UPLIFTING MODEL FOR PREDICTING SUBSCRIBER CHURN CONVERSION USING ENSEMBLE LEARNING: A CASE STUDY OF MOBILE TELECOMMUNCATION SECTOR IN KENYA

BY

ANTHONY C. OCHIENG 19/03586

A RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF MSC DATA ANALYTICS IN THE FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY AT KCA UNIVERSITY

NOVEMBER 2021

DEDICATION

I dedicate this work to my family, student colleagues, lectures and all those who have worked tirelessly to ensure that I achieve the best out of life.

DECLARATION

I declare that this project is my original work and has not been previously published or submitted elsewhere for award of a degree. I also declare that this contains no material written or published by other people except where due reference is made, and author duly acknowledged.

Student name: **Anthony C Ochieng** Registration number: **19/03586**

Digitally signed by Anthony C. Ochieng
Date: 20/08/2021
Masters student KCA
MSC MDA 19/03586

Sign: ____________________ Date: __________________

I do hereby confirm that I have examined the Project proposal of

Anthony C Ochieng

And have approved it for examination

Digitally signed by Simon N. Mwendia
DN: cn=Simon N. Mwendia, o=KCA University,
ou=College of Technology,
email=smwendia@kca.ac.ke, c=KE
Date: 2021.08.21 11:38:02 +03'00'

Sign: ____________________ Date: 21/8/2021

Dr. Simon Mwendia

Dissertation Supervisor

ABSTRACT

Churn is the number one topic for Telco's in Kenya and around the world. Customer churn in the telecommunication industry is still a big problem because emerging new technologies, lower costs, among other factors Churn brings with it many negative repercussions. While churn is a helpful key performance indicator for identifying areas of improvement whether in process or product, it can lead to financial disability eventually as customer acquisition cost are normally more astronomical than trying to please a disenchanted Subscriber. By analyzing churn drivers, we can safeguard the most import asset for a telecommunication company from churning. Predicting subscribers who are most likely to churn is fundamental for telecommunication companies. As a result, churn prediction is an important barometer for business success as well vastly studied and common activities that can be accomplished by machine learning applications for telecommunication industry. Telecommunication companies have since come to a realization that churn prediction only provide predictions but do not provide information for optimal decision making within a business setting. This is where uplift modelling has come to the fore. Uplift modeling is a branch of machine learning which aims at predicting the causal effect of an action such as a retention campaign or a marketing campaign on a given population by considering outcomes from the campaign treatment on that group, involving the sample populations that has been subjected to that campaign or treatment, and a control sample population. The model generated is then utilized to select the segment of population that the campaign would be profitable. This dissertation analyzes the use of ensemble methods in uplift modeling. The researcher will attempt to demonstrate higher performance compared to traditional classification and uplifting techniques. The researcher will attempt to show improved performance are a result of using ensemble classification techniques inculcating the differences in class probabilities in the treatment and control groups. The result being a Novel propensity outcome modification model. Safaricom plc was used as a case study to develop an uplifting for predicting subscriber churn conversion on various pre-existing subscriber segments. The objective of the study was to find the most profitable segment to target after using an ensemble classifier to predict probable churn customers on prepaid subscribers. Anonymized and pseudomized subscriber data was used for the study. The final results show the accuracy and precision of the ensemble predictive classifier and also the uplift scores for the various existing subscriber segments using the novel propensity outcome modification approach that identifies the probable segment to target with a retention campaign.

ACKNOWLEDGEMENT

During the past two years at KCA University, I have received enormous support from very awesome people to whom I would like to express my sincere gratitude. First and foremost, I would like to thank my supervisor, Dr. Simon Mwendia who taught me Data mining and machine learning and provided me with invaluable academic guidance, continuous advice and encouragement at every stage of my MSC Data Analytics Project Study. I cannot thank you enough. This knowledge has come in handy in step-by-step process progress in my research study

I would like to thank the reviewers of my dissertation, Dr. Kamau, Isaac Okola, Dr. Lucy Mburu, Dr. Dennis Kaburu, Dr. Rachael Kibuku and Ezekiel Kuria Mwangi Dean College of Technology from KCA university for their constructive critique and comments on my Manuscript. Without your knowledge and collaboration, this dissertation would not have been comprehensive as it currently is.

I acknowledge my gratitude to Safaricom PLC for providing anonymized and pseudomized data and materials for this research. I would like to express my gratitude to my colleagues at the faculty of computing and information management at KCA University for their support in various ways. Many thanks are due to the entire staff of the faculty for administrative support. Without you none of this would have been possible

Finally, I am thankful to the almighty God, the constant support of my family in allowing me to complete my studies. Your sacrifices were duly appreciated.

Anthony C Ochieng 2021

ACRONYMS

AI – Artificial intelligence

ARPU Average revenue per user

CLV- Customer lifetime value

CRM customer/subscriber relationship management

DEM Distinct Ensemble Model

ML Machine Learning

SMS -Short messaging services

subscriber churn prediction (SCP) modeling

subscriber churn uplift (SCU) modeling

UNDESA United Nations Department of Economic and Social affairs

Operational Definition of terms

Subscriber – Subscriber who has subscribed to services from a telecommunications service provider

Telco – telecommunications service provider

Mobile Money – Money Transfer services offered by telecommunication service providers such as MPESA

Data Analytics - The science of analyzing raw data in order to make conclusions about that information

Artificial intelligence (AI)- The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings**.**

Ensemble Learning - models that utilize a blend of trained algorithms to improve key metric performance than any single model can achieve alone.

Feature selection-

SQL - "Structured Query Language" is a language used to retrieve and manage data from relational databases.

Subscriber satisfaction is the outcome of a cognitive and emotional evaluation made after the tangible perceived experience on all levels and processes are compared to the expected standards.

Churn rate, also known as the rate of attrition, is the percentage of subscribers who discontinue their relationship to that service within a given time period.

Touch points is individual contacts between the firm and the subscriber at distinct points in the experience,

Subscriber Lifetime Value (CLV) is a prediction of the total value (mostly expressed in net profit) generated by a subscriber in the future across the entire subscriber lifecycle.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF Figures

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background of the Study

Churn prediction is a key barometer for future business performance forecasting well as one of the most business-oriented machine learning applications for telecommunication sector. In telecommunications sector, the term 'churn' refers to the loss of subscribers who switch from one provider to another during a given period. With the exponential growth of the telecommunications sector, churn estimation is emerging as one of the main activities for improving the bottom line. Churn rate is the number of subscribers who cease to relate with a firm, its products or services over a defined period in time. Gross adds is the number of new subscribers. A service provider's growth occurs when gross adds rate way surpasses churn rate on real terms reflected for example on ARPU, which results into a positive net adds. That leads to increase in the increase of absolute customer base. Today, global telecom market including Kenya are experiencing loss of income and subscribers owing to unrivalled competition. To stay afloat, mobile telephony service providers apply customer relationship management methods to acquire new subscribers and retain existing ones.

According to G. D.Olle Olle CV& Shuqin (2014) the biggest nightmare for telecommunication service providers is how the regulator has made it easy for subscribers to churn when disenchanted through a channel known as Mobile Number portability platform. This is where a subscriber is able to switch from one network to the other without changing his mobile number. Another concern is that Data generated by telecom service providers are strictly guided by data privacy laws that hinder scope of research in this area. According to an article written by Laolu Akindele of PWC(2020) personal data stored by telcos is very sensitive and must be securely protected as it forms the foundation of customers trust that has been entrenched by the Kenya Data protection ACT of 2019 that gives customer control over their data. Data Anonymization and pseudomization is key if one wants to do any form of data analytics research in this area.

This study aims to predict the type of subscribers of a Kenyan telecommunications company are likely to churn and their which inherent segment they belong to . The researcher in this case used uplift modeling techniques to get the segment of a subscriber population so as to apply a retention offer so as to retain them or encourage them to spend more. Uplift algorithms aim at getting the net difference in subscriber behavior that are a result from effecting a targeted treatment subscriber, e.g., retention offers that offer discounted products. Typically, Uplift algorithms allow estimation of the profitable subscriber segments based on a complex analysis of input variables, and decision variables that can be optimized. Uplift models thus make it possible to forecast the different subscriber segments that be used to recommend optimal decisions in application of retention measures . According to Elkan C (2001) Uplift modeling is a method of prescriptive analysis that basically applies Bayesian methods on empirical data to provide for easy decision making. Such approach involves the existence of statistics that depicts the actions of the modeled structure under different circumstances and decision-making situations.

Uplift modeling case studies for customer churn have been done before. Radcliffe and Simpson (2008) applied uplift modeling to data from two customer churn groups in telecommunications. One study group was highly effective and profitable with accuracies of 80 percent, whereas the other was observed to be counterproductive and yielding a net loss with accuracies below agreeable thresholds. Both campaigns' outcomes in terms of reducing churn improved as a result of uplift modeling with accuracies above 90 percent. Leo Guelman et al. (2015) implemented uplift modeling in customer churn for the insurance industry using random forest. He articulated that a positive impact in accuracies could be attained if the customer population was segmented and those segments would them be used in uplift modeling

In various sectors applied statistics was utilized for churn prediction using tools such as spreadsheets and commercial statistical tools. However, more recently data mining and machine learning techniques for the churn prediction and uplift modelling have been become prevalent. Applied statistics approaches are static in scope and implementation agility. Classification algorithms support data analytics to effectively predict churn in telecommunication sector. Uplift modelling becomes effective to determine Which subscriber segment that was an outcome of the churn prediction model, that didn't not churn because of a retention offer received.

According to Preeti K. Dalvi, Siddhi K. Khandge, (2012), data analytics will find the subscriber with an elevated likelihood to churn, however not essentially presenting the reason of the subscriber churning. This study aims at look at subscriber various ARPU and demographic data and utilizing data mining and machine learning method to explain the churn behavior and assist in decision making to change that customer behavior using uplift principles. The study endeavors to analyze, data extracted from telecom service provider in Kenya so as to gain insights which can be used for client churn management and taking very significant business decisions on retention of subscribers who are likely to churn. In the last ten years the number of smartphone/feature phone users have increased significantly. According to the report by UNDESA the World is estimated to be about 7.7 B in 2019, With a penetration rate of 97% we may about 7.6 billion smartphone and feature phone subscribers with an exponential growth from the 738 million Subscribers estimated in 2000. Europe and the Americas have an adoption rate way more than 100%, subscribers clearly out striping the world population average.

The Kenyan population is estimated at about 51 million while mobile penetration rate is at 97.03 percent according the Communication authority sectors statistics report (2020). The high penetration rate leads to a non-productive market which leads to higher need to retain existing subscribers rather than acquisition. Bhattacharya et al. (1998), Athanassopoulos et al. (2000), Farris et al. (2010) reiterate that new subscribers' acquisition is several times more expensive than keeping existing subscribers. Subscribers who have existed in the telecommunications network for a long period are much more likely to offer higher customer lifetime value over time than newly acquired subscribers. Long standing subscribers often recruit new subscribers by word-of-mouth casing point being colleagues' friends and family offering additional value to the telco.

The major objective for a Kenyan telecommunication company should be acquisition and retention of highly profitable subscribers through churn detection and mitigation to result in higher profit margins. Subscriber churn for Kenya telcos such as Safaricom PLC, Airtel and Telkom Kenya results in reduced sales, and causes poor and negative sentiments about the affected service providers, that eventually hurts the companies' brand image. According Rob Mattison (2005), the cost of the lost sales can be measured, the loss due to negative sentiments can at best only be vaguely estimated. Churn prevention or management and executing the mitigating activities are integral in keeping the stability of any telecommunications service provider. Churn Uplift modelling assist in decision making when implementing churn campaigns, by identifying the most probable churners segment that when targeted would be profitable to the telco when provided with incentivized retention offers, discounted calling rates and free upgrades. When using Churn machine learning and uplift modelling in identifying potential subscribers to give incentivized retention offers normally results in saving or profits for a telco. This study presents a churn model using ensemble learning to predicts subscriber segments conversion rates using ensemble learning. This model benefits telecom service providers by effectively meeting business needs for decision making during churn management lifecycles and reduce costs for the telecom service providers in comparison to other models by improving the prediction metrics significantly. Churn rate is defined as the absolute number of subscribers who cease to relate or utilize a firms, products or services with a specified period.

According to Rob Mattison (2005), churn can be categorized into three types;

Unavoidable churn: This occurs when a subscriber is incapacitated because of death or moves or is permanently removed from the network's ecosystem, travels outside the country without roaming, and possibly immigrates to places outside the outside network coverage. In general, where churn is inevitable.

Involuntary churn: This occurs when subscribers defaults on service payment and as a result the provider discontinues permanently the service. Ceasing of services due to criminal activities or fraudulent usage is also classified as involuntary churn.

Voluntary churn: Termination of service relationship by the subscriber, leaving one service provider for another because of better value for money or subscriber dissatisfaction with current service provider because of varied reasons.

According to Rob Mattison (2005), One of the possible reasons that churn is inevitable is where technology keeps changing. Newer, better, more fashionable, good to have and less expensive alternatives will always be invented, allowing subscribers have more variety. Telco technology becomes obsolete very fast. Since technology is very dynamic because of constant innovation subscribers who are always be looking for this innovation are always likely to churn. Currently in Kenya telcos are moving into 5G technology keep in perspective to where the wider telco industry is moving to.

Regulators can also contribute to churn when they make regulation to allow entry of more competitors as they endeavor to provide improved quality of service and low-cost services, Churn has many repercussions, and most of them expensive. The largest repercussion of churn is the loss of revenue. The loss of many subscribers can create a gap in terms of lost revenue especially when a telco has not diversified its revenue. Reduction in calls, SMS and mobile money billing rates to remain competitive also leads to reduction in revenues. Cost of reacquisition of lost subscribers is also very high through acquisition campaigns which may or may not be successful. Also, retention and loyalty campaigns to avoid losing subscribers also affect the bottom line. These loyalty schemes create liabilities on the balance sheets that are sometimes hard to manage.

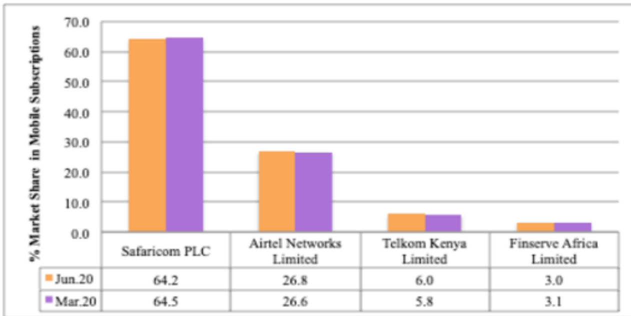
Although we have seen through innovation such as introduction of Mobile money, churn is reversible of a long period of time. Innovative products is a big factor in retention for customers. Also, when network churn is first discovered the first instinct for a telecommunication service provider is to increase its advertising budget and cut costs. The battle for media supremacy also costs money that affects the bottom line. The cost saving measures also reduce the agility of a telecommunication service provider to react to churn. Churn also affects employee's psychology, where they get apprehensive about how churn will affect their jobs. The company policies, also change to reflect cost cutting measure to counter churn as they try to cope, but often getting negative results. The finance department normally face challenges when implementing a planning, budgeting strategy and projecting of revenues. The Human resource department face challenges projecting headcount for the period. Decision making about how and where money is to be invested to prevent the churn is also a nightmare. When considering effects of churn, business executives look at costs, revenues, investments, and long-term stockholder stock which are at the core of each telecommunication service provider's financial management best practices.

According to Rob Mattison (2005), Churn sometimes leads investors to lose faith and confidence in their investment especially if there is failure to prepare or mitigate churn decisively. The unfortunate thing is that when churn prerequisites are present in the market, countering it is a difficult task. It is believed using data analytics one can convert churn to one's advantage. A telco company can be able to determine the drivers of churn by looking at why subscribers are interacting with you at all your touchpoints.

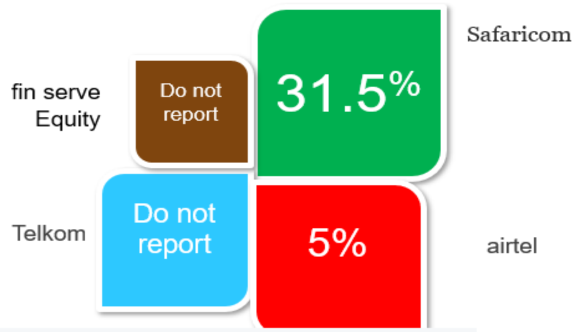
1.3 Statement of the Problem

Churn management process is a segment of subscriber relationship management. It emphasizes the monitoring of subscriber behavior and comparing the same with relevant churn drivers for a telecommunications provider so as derive action points across the business to mitigate. Currently we see Kenyan Telco's having a Churn rate of between a low of 5 percent to a high of 31.5%. This has the translated into a reduction of Market share for Safaricom plc as shown in figure 1.3 using from a report shared by the Communications authority for year 2020. This reduction in market share leads to a reduction of revenue for Safaricom PLC and is unable to meet its double-digit growth objective on all revenue touch points.

FIGURE 1.3 TELECOMUNICATIONS CHURN RATE IN KENYA



Source: CA, Operators' Returns. * Mobile Pay Ltd data not available due to non-compliance by the operator



Key performance indicators

	HY21	HY20** Restated	HY20 Reported	% Change
Churn (%)	31.45%	21.20%	21.20%	(10.3ppt)

Sources: 1. https://airtel.africa/assets/pdf/FY-20-q4-Press-release.pdf
2. https://www.safaricom.co.ke/images/Downloads/Resources_Downloads/H1_FY21_Results_Presentation.pdf

Another significant problem for any telecommunications service provider is to eliminate or lower significantly customer churn. To be able to do this they have to identify a particular segment that has been identified as probable churn subscribers and provide mitigation measures to alleviate churn within the network through targeted campaigns . Out of these segments there exists inherent segments of subscribers who when targeted with a retention campaign are likely to change their minds and stay with the telco and subscribers' segments that will not change their mind when targeted. Here, the problem emerges how to identify the later segment and remove them from any future campaigns, as targeting them will be a waste of valuable resources. In some instances, loyal customers if target by such campaigns may decide to churn as they deem such campaigns as spamming and intrusive to their privacy. For Example, a retention offers a discounted rate for prepay product, yet they are on post pay product and cause them churn as a result as they will deem the current post pay product doesn't offer value for money as with a the prepay subscribers. Customers with unresolved issues are highly likely to churn. Any attempt made to contact such a subscriber will make the subscriber to churn making any future campaigns to that said customer to have a negative impact. That type of subscriber must be removed from the list for targeted future campaigns. The existing Classical churn prediction techniques only differentiate churn from non-churn subscribers and from that analysis give while uplift modeling distinguishes the customers when targeted by a campaign will be profitable to the telecommunications service provider. Classical churn prediction only makes predictions. So, to move to optimal decision-making profit driven analytics is need hence the importance of uplift modelling Targeting subscribers identified by an uplift model will lower the impact of churn with accrued savings to the business, mitigating the first challenge associated with classical prediction algorithms. Floris Devriendt, Jeroen Berrevoets and Wouter Verbeke 2020 suggested the use of uplift modeling to optimize the subscriber churn prediction algorithm for optimal decision making.

1.3 Main objective

The primary aim of the study was to develop a churn model that inculcates uplift principles for predicting subscriber churn in Kenya telecommunication sector.

1.4 Specific Objectives

I. To identify the attributes that determine subscriber churn in the Kenya telecommunication sector.

II. To develop a predictive and uplift model for subscriber churn.

III. To evaluate the proposed models.

1.5 Research Questions/hypothesis

The research will attempt to respond to the following questions as we try to come up with a churn prediction model for the telecommunications sector in Kenya:

1.Which attributes(features) influences subscriber churn in the Kenya telecommunications sector?

2. Which is the appropriate predictive and uplift model for subscriber churn?

3.How effective is the model?

1.6 Significance of the Study

This study can be used as a foundation for further research on Telco churn conversion research using uplift modelling principles for telecommunication sector in Kenya and beyond. It can be used to determine whether Uplift modelling approaches and frameworks adequately address the Business Objectives of telecommunication sector churn in Kenya. The Telco sector stands to gain by getting an accurate model to address Churn and manage churn conversion within their network. This will eventually lead to high profitability with the Telco companies.

1.7. Justification of the Study

Telecommunication sector in Kenya is growing rapidly. Churn management will always be part of the telecommunication service provider business though a very costly experience. Churn has a lot of effects, and most of which have a major price tag. The biggest being revenue reduction. Depending on the area and economic status of the region, the average subscriber usage average between US$20 to US$80 according to Rob Mattison (2005).

The churning of many subscribers gives rise to a large dent in the corporate balance sheet. The direct loss of sales and drastic reduction of billing rates can ruin a company's bottom line. The endemic behavior of dropping prices then reassuring consumers about their competitiveness when it comes to value for money sometimes may not work depending on the business competence and power of the competitor. According Rob Mattison (2005) This rate cut contributes very naturally to a reduction in the company's annual turnover hindering other activities that may have led in churn reduction. In spite of their best attempts to avoid turnover, the company would eventually lose a chunk of its subscribers to the competition eventually and may be gain the subscribers that the competition discarded over the same period.

They could also win back lost customers by re-acquisition initiatives. These campaigns are always successful but of course, costly themselves. Indeed, more proactive telecommunication service provider are building excellent customer and product experience strategies to help mitigate people from churning. According Rob Mattison (2005) these customer experience investments, loyalty or retention offers give a chance to the company to make eye catching product offer or create out of this world subscriber experiences such as experience centers and segmented delivery of services that help in retention Closing the loop on customer complaints and acting on customer feedback help to identify problems before they become unmanageable.

Some of direct costs incurred by telecommunication service providers when trying to mitigate churn includes, increase its advertising to have more media "presence time" than the competition, this leads to media battles where the only winners are the advertising companies. Price wars are also undesirable to the service provider but beneficial to the subscriber but only for a short period of time as the market normally corrects itself after the firms realize there is not much benefit that comes from it because often the outcome leads to revenue loss. According Rob Mattison (2005) subscriber churn impacts on the working environment within the telecommunications network, take a toll on individuals, systems, and organizations as they try, sometimes unsuccessfully, to resolve issues. Another of the possible adverse impact is in planning and budgeting especially on income estimation and staff numbers estimation for the period when churn is happening. Resource spending decisions become untenable as decision on what to spend on take hours of long discussions. Subscriber churn dynamics includes management of expenses, sales, acquisitions and long-term shareholder stock at the center of any financial management policies. ideally, all these other actionable items have an impact on investors trust.

1.8. Motivation of the Study

The major Motivation for this study is to find the drivers that contribute to churn so that the telecom service providers can address them and to use uplift modelling to guide campaigns offers so as to mitigate churn.

- The other motivation is to understand subscriber lifetime value across different subscriber segments and sales regions
- Another motivation is to get insights on subscriber ARPU across different subscriber segments and sales region
- To develop an appropriate model that can be used to predict churn and uplift conversion for telecommunication sector in Kenya

1.9 Scope of the Study

As such, the research will assess Telecommunication churn in Safaricom. We will explore the various dimensions in churn using Data mining and machine learning technologies required to predict churn effectively. The study will be restricted to prepaid subscribers in the telecommunication sector and will be guided by enabling technology through data mining and machine learning, data governance, regulatory policies and best practices in churn management in the Telecommunications Sector specifically in Kenya. The data will primarily be for prepaid subscribers for the Kenyan Telco.

1.10 Assumption of the Study

- Due to machine power constraints the sample data will not exceed 1.2 million subscribers
- Python will be used for visualization and machine learning algorithms
- Data will always be anonymized to protect the identities of the subscribers

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 Introduction

This Literature review chapter compares the different arguments, hypotheses, methodologies and conclusions in telecom churn research and try to relate the same with the study under research. We will discuss the telecom churn concepts in the global, African and Kenyan context. For our case the literature review will contrast and critique the various prediction techniques, arguments, themes and methodologies and approaches expressed in the literatures. The focus here is on Telecommunication sector in Kenya with the case study on Safaricom PLC. This chapter provides the connection between various literatures, predictive techniques and the one under study using our ensemble learning uplift modelling approach to predict churn.

2.2 Theoretical Framework Review

The theoretical context is the mechanism that can hold or endorse the hypothesis of a research study that it proposes and outlines the theory that determines whether there is a research issue under study. (Bolton, 1998) .The research stream followed here emphasizes the link between telecom subscriber attributes , subscriber churn and uplift modelling.

2.2.1 Subscriber churn in telecommunication Sector

a. **Definition of churn** Loudon and Laudon, (2012) defined subscriber churn as the number of subscribers who cease buying or consuming a products or services from a service provider. The famous scholars demonstrated churn rate as a very important sign of growth or decline in any company's consumer base. They reiterated that if these indicators were not addressed sufficiently, they would lead to a loss of revenue Reicheld (1996) reiterated that an increase in the churn rate results in a reduction in business cashflows, even if the telco is able to replace lost subscribers through new acquisition. This is because the cost of acquisition is normally very high compared to the cost of retention. McKinsey estimated that reducing churn could increase earnings of a typical telecommunication technology company by as much as 10%. This is because the same subscriber can be induced to acquire new products from the same company through upselling and cross-selling.

Lemmens and Croux, (2006) consider the definition of churn in the marketing domain by using it to segment subscribers and offering them the best service before they decide to take their business to some other service provider. The mantra being all subscribers are important but they are never the same. Matrics, (2013) noted that subscriber churn also known as subscriber defection or subscriber attrition and stated churn as the rate at which a business is losing subscribers or revenue through leaving the service provider. He goes on and looks what the possible drivers of churn could be.

b. **Subscriber churn as global phenomenon**. Subscriber churn is a continuous occurrence affecting both sides of the balance sheet namely the assets and liabilities. (Lemmens and Gupta 2013) in their scholarly study called Managing Churn to Maximize Profits in the USA proposed a predictive model that took in consideration the profit item when predicting the subscribers who are likely to churn. They used a profit loss optimization function that can assist a telecommunication service providers target subscriber who were likely to churn by ranking according profit margins if retention was successful. Some insights on the drivers of churn were provided during the Teradata Center Hackathon at Duke University, where 44 participants were interviewed. The subscribers interviewed noted that cost of services and subscriber service dissatisfaction were the main drivers of churn. (Kumar, Thomas, and Reinartz 2005), used probability models and classifiers to predict churn and their drivers by observing data generated from a subscriber management process. For example, find the approximate probability that a subscriber would likely churn if a process was not amended used heuristics to predict churn.

Ascarza (2016) created a profit loss functions using random forest method for change in likelihood to churn prediction attributed to a retention proposition. These retention propositions if applied correctly would mitigate churn (Neslin et al. 2008, pp. 256-259). They used logistic regression to model their churn prediction. This logistic regression method was only successful in some types of data sets only that displayed nonlinear tendencies to generates the coefficients of a formula to predict a logit transformation of the likelihood of occurrence of the feature of interest Logit (p) = b0 + b1X1 + b2X2 + …… + bnXn. Where p is the likelihood of occurrence of the feature of interest and Xi are features they defined their logit transformation as the logged odds Logit (p) = ln (p / 1-p). Thamsaranasakul (2008) carried out a research in Thailand with the Objective of showing the significant drivers that cause subscriber churn of telecommunications Sector in Bangkok which revealed a link among demographic characteristics and some of the marketing process or company response factors. These demographic factors include ability of subscriber to afford services based on employment status, level of education, poverty levels, location etc. (Buckley et al 2012) used neural networks and Artificial intelligence methods to predict subscribers likely to churn (Huang et al. 2012), used random forests classifiers to predict churn using various subscriber management processes (Van den Poel and Lariviere 2005; Verbeke et al. 2012), used bagging and SGB regressors with a loss function to predict churn (Lemmens and Croux 2006), used hazard methods to predict churners that used structured and systematic technique for churners examination and churn management.

(Bhattacharya 1998; Braun and Schweidel 2012; Bolton 1998) all used hidden Markov functions to predict churn which were mathematical models that tried to predict churn. (Hardie & Ascarza 2013; Schweidel, Bradlow, & Fader 2011; Knox & Schweidel 2013) used multiple regression methods to predict churn which was an extension of linear regression models that allow predictions of churn with multiple independent subscriber variables; (Knox and Oest 2014; Verhoef, Risselada, and Bijmolt 2010), did classification using the probit methods to be able to predict the likelihood of a subscriber churning. This method explored the relationship between a stimulus (dose) and the quantal (all or nothing) response using Quantitative responses (Ascarza et al. 2016a; , Heerde, Foubert and Datta 2015), used decision trees models or CART to predict whether a subscriber likelihood to churn may be affected by subscriber factors and churn drivers. These CART methods using a loss function to predict the subscribers likely to churn.

c. **Subscriber churn in Africa.** In Africa Abaidioo (2011) carried out a study “Predicting subscriber churn in the mobile Telecommunication industry a case study of MTN Ghana to find out the impact of subscriber churn in MTN Ghana, the causal reasons of subscriber churn and to build a predictive model for churn in the telecommunication sector in Ghana. The outcomes were that the propensity for a subscriber to churn was found to be 1.03 times which is extremely high, it was also noted that the cost of local and international calls and poor subscriber service were the causes for churn in MTN Ghana.

In Zambia`s Banda (2016) proposed a system dynamics approach to subscriber churn management in the mobile Telecommunication industry. The study proposes a model that uses system dynamics to give significant insights into managing subscriber churn in the Telecommunication sector in Zambia. The data was qualitative and was collected through surveys and interviews from more than seven hundred mobile subscribers in six districts of Zambia, that included the three telecommunications service providers and Zambia Information Communication & Telecommunications Agency (ZICTA). The results of the analysis indicate that, on aggregate, telecommunications service providers in the Zambian telecommunications sector encounter annual rates around 3.73 and 9.14 %. The model demonstrated that if the telecommunication service providers implemented the recommendations suggested, annual churn rate can reduce to about 1.02 percent. The scholar also discovered that the pattern of churn changes over time and are affected by several factors that may be known or unknown making its management complex.

In South Africa (Hoffman, 2013) did research on ‘Reducing churn from price increase; an experimental intervention’, the study took an in depth look into successful interventions that assist a business in retaining subscribers while increasing prices. The findings were that by simultaneously offering Subscriber’s products that were customized to their needs that would subscriber retention efforts within the telco industry.

d. **Churn review in Kenya.** Literature on subscriber churn is very scarce in Kenya except for a few. One notable study was done by Kirui (2013), In the study the scholar was Predicting Subscriber Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining, noted that subscriber churn in the mobile telephony industry is a Perpetual problem due to tough competition, emerging technology, low switching costs, deregulation by governments, among other factors. The findings were that subscriber churn prediction is playing a central role in churn management in telecommunications sector and to detect and manage the various costs associated with subscriber churn, it is imperative that mobile service providers deploy churn predictive models to detect subscribers who are likely to churn, then implement intervention strategies.

Another notable study was done by Patricia Kemunto (2017) predicted churn using Decision trees for local telco called Orange now known as Telkom Kenya. In her study she used a survey to ask customers about their probable reasons for churning, she looked at various variables such usage frequency, airtime top-ups length of service and others variables in her study . She concluded from the data analyzed that churn was between 2 and 3 percent in

Orange. The causes of customer churn were poor network quality and mobile money customer related issues.

Finally, a study by James Macharia (2012). He looked at customer predictive modelling using the Cox model and Decision trees and based his research on survival theory analysis. His conclusions stated that the decision trees model showed improved churn prediction results when analyzing churn in the telco industry in Kenya and gave recommendation that the churn models should be run monthly so as to be able manage churn with the telco business.

2.2.2 Attributes/factors that determine subscriber churning in telecommunication sector

Churn is a complex concept that entails analyzing each aorta of subscriber behavior, analyzing scientific advances in the smartphone industry (5G) and realizing the role of a competitor in innovations of products and services (Mobile Money). Wong and Sohal (2003) looked at the impact of customer experience by offering quality services through subscriber loyalty/stickiness. According to Wong that aids in subscriber retention, eventually lowering subscriber churn. Antreas (2000), studied service standards that influences quality of customer experience, price points and innovation and their significance on subscriber satisfaction. He concluded providing convenience through such measure helped in customer retention.

Kim and Yoon (2004) explored the factors contributing to subscriber churn and subscriber satisfaction within the telecommunications sector in Korea, utilizing a binomial logic framework. Researchers observed that the amount of good customer experience, inter-call quality; tariff points, devices, brand presence, revenue and age on the network/tenure were the salient factors causing the subscriber churn. Inter-call reliability with no call drops and brand presence have made subscribers aware of the brand. The research found out that subscribers did not churn mainly because of porting out costs. Misra (2014) researched on the subscriber satisfaction on service quality for post-pay and prepay subscribers and its significance on them churning. The author found that prepaid subscribers' quality of customers experience was superior when compared to postpaid subscriber's customer experience. She also concluded that prepay subscribers are more price sensitive and are not churn averse, if they get higher value for money or when they listen to their family members and friends on which service provider to stick with.

Ahn et al. (2006) did research, where he utilized subscriber's billing records and subscription records to discover the attributes causing subscriber churn in telecommunication sector. They gathered the data from a telecommunications company for subscribers who were continuously utilizing their mobile lines for a period of three months and monitored gathered records for thirty-six weeks. They discovered that subscriber churn was affected by porting out cost. If the cost of porting out is high likelihood of subscriber switching is reduced. The researchers discovered, with the fulfillment of mobile number portability (MNP), which gave the subscribers the ability to exchange service providers, while remaining with their current registered mobile line has churn. Ahn also concluded that subscriber churn was directly dependent on loyalty points a subscriber has or is getting for every interaction made while using a service provider's products and services, as the objective of these transactional points accumulation schemes is to reduce churn by providing free products and services for theses subscribers. The authors discovered subscribers churn decisions were not arrived at abruptly, but behavioral change or an event such as non-use or suspension of services led to such decisions. Even so, the researcher did not include the age of the subscriber and the age on the network variables for that study.

Chen and Ching (2007) investigated effects surrounding subscriber experience sustainability & telecommunications company on points gathering schemes and impact of brand presence on a structured model of interactions. To capture the whole ecosystem of telecommunications services experience they included billing records and network coverage key performance indicators in the research. A clear positive relationship between subscriber service experience, brand presence and subscriber stickiness existed after analysis. In the course of their study, they discovered brand presence had a significant job for creating and maintaining subscriber stickiness/loyalty and impacted only by billing complaints. While network coverage had no correlation when compared to brand presence. Sustaining and maintaining subscriber loyalty schemes seemed to be the most critical aspect for customer retention, where service providers rely on a long-standing symbiotic relationship between it and its subscribers to retain them.

Rousan et al., 2010 discovered experience metrics performance was highly affected with subscriber stickiness to a service provider, impacting positively to subscriber retention while complementing market size growth. Similarly, Kotler and Keller (2006) emphasized that customer experience was a prerequisite for subscriber retention and acquisition activities. Poku et al. (2013) researched on how excellent customer experience provision through service quality impacts subscriber loyalty. The authors selected 50 subscribers randomly from a city in Ghana to be included in their study. They collected their views and surveys responses. They used discernibility, positive declaration, dependability, responsibility and emotional IQ as their research parameters to determine the correlation between subscriber loyalty and service experience quality. The researchers came to an understanding that dependability is an indicator of customer experience and service experience quality. They noted there is a positive association between customer experience, service efficiency and subscriber stickiness. The Authors discovered subscribers would be in the lookout for value-enhanced products and services which impacted their likelihood to churn.

A research on correlation between subscriber experience and churn reduction in telecommunications sector by Omotayo and Joachim (2008) discovered a close correlation with subscriber satisfaction and churn reduction. These researchers utilized subscriber satisfaction, product worth and behavioral intuition variables in their study. A questionnaire was prepared to gather relevant information from subscribers in one telecommunication service provider in Nigeria. The revelation was when subscribers' satisfaction determines whether a subscriber will be more or not likely to churn and hence excellent subscriber satisfaction decreases the likelihood of churn.

Product price is a significant factor in developing markets such as the telecommunications industry in India According to Srivastava et al (2006). Provision of incentivized value propositions, innovative products, excellent service experience at fair rates, is a prerequisite to mitigate subscriber churn. Service Experience Quality, together with price, also plays a very crucial part for subscriber's churn in India (Mittal and Sirohi, 2007). Their study revealed that although most telecommunication service providers offer excellent subscriber experience, the callings rates set by telecommunication providers was the most critical variable when analyzing subscriber churn. Customer experience and service quality, including fair pricing are a prerequisite to mitigating churn. Sathish et al. (2011) studied the changing behavior of telecommunications subscribers in Chennai and demonstrated subscriber service experiences, customer complaints on services, cost of services etc. as critical determinants of subscriber churn. Researchers proposed that call rates play a significant role for deciding whether a subscriber will churn together with network coverage; incentivized innovative value services and subscriber experience. The outcome of the study also concluded that marketing advertisements contributed the least. The scholars discovered that family influence towards churn was at 41%, while marketing & media presence was at 2% of the users.

Revathy and Padmavathy (2005) discovered that a person's age and family income had an influence on the choice of a service provider, while the academic qualifications; gender and professional alignment did not affect their choice. The scholars also studied product and service charging complaints, internet connection issues, lack of network coverage and found them to influence user's churn decisions. The scholars noted, lowering of tariff rates and revving up their online and mass media presence greatly reduced churn. Benjamin et al. (2012) studied reason why subscribers churn in a west African city Lagos in Nigeria. A survey was done to get response from 800 subscribers through random selection. They utilized wages, age, level of education, whether employed or not as their features and churn likelihood as the dependent variable for their study. Univariate and multivariate analytics were utilized for their study. Network coverage issues across the country was found to be the biggest contributor to churn. Other salient factors contributing to churn included high call tariffs, customer experience, mass media presence, multi sim attribution and incentivized tariffs. Age and gender were found to have no effect on the subscriber's likelihood to churn.

Omar et al. (2014) utilized multilayer perceptron (MLP) neural networks techniques to analyze the subscriber churn. The scholar created a model using randomly selected 5,000 subscribers from a Jordanian service provider. His features included, subscribers' monthly rates, calling rate, SMS sending rates, calling and SMS behavioral tendencies and network type (3G). They discovered that subscription transactions, total minutes usage and availability of fast 3G internet service contributed to a subscriber likelihood to churn. They define service experience quality as the cumulative total of core features such as network reach, voice quality, transmission power, subscriber service experience, equal business opportunity. According to Aydin and Ozer, 2005; Ashraf et al, (2013) mass media and brand presence of a service provider is also a very critical tool to get subscriber satisfaction approval and to reduce churn. According to Malik et al. (2012) greater service experience quality and positive brand presence results in increased subscriber satisfaction and good Net promotor scores. If the subscriber is content, telco provider loyalty is increased as well. Subscriber loyalty is a result of excellent customer experience. Subscriber loyalty results in the consistent utilization of a service providers products and services reducing subscriber churn, hence the inverse correlation between subscriber loyalty and churn. Unlike post-paid subscribers, prepaid subscribers are not bound by bundled contracts and are susceptible to churn. In conclusion Subscriber retention we have seen to be correlated to Customer experience and product experience. Customer complaints like network coverage and inept customer service increase probability of churn. Other salient factors include ineffective complaints handling and Charge disputes. Factors such as price points, lack of innovative products, and ancient technology may also cause subscribers to churn.

2.2.3 Techniques (algorithms) used for prediction

a. Classification Algorithms

According to Har-Peled, S., Roth, D., Zimak, D. (2003), Classification is a technique requiring the use of machine learning algorithms which learn how to implement a class label to a specific item in a set or problem case. An interesting subject case is to identify emails as "spam" or "ham." There are many different types of classification activities that one can undertake in machine learning and data mining. In machine learning, classification denotes a predictive modeling task where a category label is predicted for a specific type of data. Types of classification activities can be applied to the below scenarios:

- When you have a database for emails that is received by the business and there is a need to classify if one is spam or not.
- Text represent more than 90% of a business data pool whether it is customer sentiments or social media text one can classify each character of text as one of the known product category.
- When we have customer social economic and demographic data one can classify them as churned or active.
- Viewed with machine learning perspective, classification utilizes a training subset of the population with several features and dependent variables for learning purposes

The algorithm would use the subset of the data identified for training where intuitively it would try to match independent variables to class labels known as dependent variable. Meaning, the subset of data utilized for training must define the problem set to include several instances of every category label or Class labels as they are commonly referred to. These category or class labels are mostly named variables, e.g., "spam," "ham," and must delineated to an integer until the modeling algorithm is given. In the data science realm, it is known as label encoding, where a special integer is delineated to every category label, for example. "spam" = 1, "ham"= 0.

A number of classification algorithms exist can be used for prediction of classification-based problems. Classification model's performance are judged based on certain metrics. These metrics include accuracy, recall precision, F1 score etc. Accuracy may not give one desired result for metrics measurement, though it could be guide towards further analysis together with the other fore mentioned metrics. Different problem sets may require likelihood probabilities predictions for every category membership for each example instead of class labels. Precision and recall can be used for these instances. This induces additional ambiguity in the estimation that an application or individual will be able to understand. The ROC Curve is a diagnostic tool for determining expected probabilities. There are four major categories of classification activities which you may consider;

- Paired Classifications commonly referred to as Binary Classifications
- Multiple classes classification commonly referred to as Multi-Class Classification
- Multiple label classification commonly referred to as Multi-Label Classification
- Imbalance labels classifications commonly referred to as Imbalanced Classification

a. Binary Classification

According to Har-Peled, S., Roth, D., Zimak, D. (2003), binary classification applies to those classification activities which have a pair of class names.

- Spam Email categorization (spam or ham).
- Predicting churn customers (churn or no churn).
- Predicting conversation that could to lead to an outcome (buying or not buying).

Differential labeling activities include one category that is an acceptable and unacceptable states. E.g., 'ham' is an acceptable state and 'spam' is an unacceptable state. Further examples include "cancer not detected" which is that acceptable state which requires a diagnostic examination & that "cancer detected" which is the unacceptable state. This category of the acceptable state is given the category label 0 & the category of the unacceptable state is given a category label 1. Typical examples of algorithms that can be utilized to classify binaries include:

- Logistic Regression

- k-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Naive Bayes

There are model frameworks that are made for binary classification and do not primarily act with more than a pair of categories; e.g., Logistic Regression and Support Vector

b. Multi-Class Classification

According to Har-Peled, S., Roth, D., Zimak, D. (2003), multi-class classification refers activities for classifying multiple category labels. Types of such include:

- Face detection.
- Plant species identification.
- Optical character recognition (OCR).

With exception of binary classification, multi-class classification has no concept of acceptable/normal or unacceptable/abnormal results. Instead, items are listed as contained in several known groups. On certain problem cases, the number of category labels can be numerous. E.g., a model could predict a picture of man or woman from a group containing millions of pictures in a face-detection system. Activities involving prediction of a series of items, such as language translating models, are referred to as a unique type of multi-class classification. Every word in the set to be predicted need a multi-class classification of which the scale of the vocabulary determines a range of additional word categories which will be predicted that can number in millions. It's also popular to model a multi-class classification activity using a framework that predicts the spectrum of Multinoulli probabilities for every instance. The Multinoulli distribution is a discreet probability distribution that covers an instance which an occurrence would have a categorical outcome, e.g., K at {1, 2, 3, ..., K}. For classification, this implies that the algorithm estimates a likelihood of an item belonging to every category label. Numerous model frameworks utilized for binary classification can be utilized for multi-class classification. Example of sample model frameworks which are utilized for a multi-class classification entail:

- Naive Bayes.
- Random Forest.
- Gradient Boosting.

- k-Nearest Neighbors.
- Decision Trees.
- **Naïve Bayes**

  Definition: According to Hastie, Trevor. (2001), Naive Bayes algorithm is structured on Bayes' theorem with the assumption of independence between every pair of features/variables. Naive Bayes classifiers predict well in real-world situations such as document/Topic classification and spam classification.

  Advantages: This algorithm requires a small proportion of training data to estimate the necessary parameters/variables. Naive Bayes classifiers are fast visa vis other complex techniques.

  Disadvantages: Naive Bayes is known to be a bad estimator/predictor.

- **Random Forest**

  Definition: According to Hastie et al (2008). Random Forest classifier is a meta-estimator/predictor that fits several decision trees on various subsets of data and uses average to improve the predictive/estimation accuracy of the model and manages over-fitting very well. The subset size is always the same as the original independent variable sample size, but the samples are drawn with replacement.

  Advantages: Over-fitting is minimized, and random forest classifier has better accuracy than decision trees.

  Disadvantages: Slow prediction times, hard to implement, and complex algorithm.

- **Stochastic Gradient Descent**

  Definition: According to Bottou, Léon; Bousquet, Olivier (2012), stochastic gradient descent is a simple & efficient approach to fit linear models using gradient boosting techniques. It is particularly useful big sample size. It supports several loss/cost functions and penalties for classification.

  Advantages: Efficient and easy to implement.

Disadvantages: Requires hyper-parameters tuning and it is sensitive to large number of features/columns.

- **K-Nearest Neighbors**

  Definition: According to Piryonesi S. Madeh; El-Diraby Tamer E. (2020) K-Neighbors classifier is a category of lazy/malaise learning that doesn't try to construct a general internal model, but stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors' distances of each point.

  Advantages: Easy to implement, robust to noisy/outlier prone training data, and effective large datasets.

  Disadvantages: Need to get the value of K and the computation cost is punitive as it needs to compute the distance of each instance to all the training subsets.

- **Decision Tree**

  Definition: According to Cormen, Thomas H. (2009), given a subset of features together with its classes/labels, a decision tree produces a sequence of rules for classification of datasets.

  Advantages: Simple to understand and visualize, requires little data preparation, and is functional for both numerical and categorical data.

  Disadvantages: Decision tree can create complex/deep trees that do not generalize well as they are prone to overfitting, and decision trees can be unstable because small variations in the data might result in a completely different tree generation.

Binary classification could be modified for use in multi-category tasks, especially where the use of a technique to match several binary classification algorithms with every category vs most other categories (named one-vs-rest) or one algorithm with every class couple (called one-vs-one).

- **One-vs-Rest**: Match one binary classification model per category vs. all other categories.
- **One-vs-One:** Match a binary class algorithm for every category pair.

Binary classification algorithms that may utilize these techniques for a multi-class classification are:

- Logistic Regression.

- Support Vector Machines.

c. Multi-Label Classification

According to Heider, D; Senge, R; Cheng, W; Hüllermeier, E (2013), multi-label classification are those classification activities which have one or more class labels, in which one or more classes can be predicted for each case. E.g., photo/picture classification, of which one given picture could have many attributable items, a model may predict the availability of multiple identifiable attributes in the picture, such as “boat,” “orange,” “boy,” etc. This is distinct from binary classification and multi-class classification, where one category is expected for every label. The popularity of modeling multi-label classification activities with a model that forecasts multiple dependent variables is on the rise, with each dependent variable taking predicted as a Bernoulli probability distribution. These are algorithms that makes multiple binary classification forecast for every type. Classification models utilized for binary or multi-class classifications can’t be utilized for multi-label classifications. Unique versions of normal classification algorithms/models can be used, so-called multi-label versions of the models, e.g.:

- Multi-label Decision Trees
- Multi-label Random Forests
- Multi-label Gradient Boosting

One can also use different classification models to forecast the labels per category.

d. Imbalanced Classification

According to Max Kuhn, Kjell Johnson (2018), Imbalanced classification refers to classification/model activities where the number of samples in each class is not equally distributed. Typically, imbalanced classification activities are binary classification activities where most of samples in the training dataset belong to the acceptable/normal class and a minority of samples belong to the unacceptable/abnormal class. Types include:

- Fraud detection problem case.
- Outlier detection problem case.
- Hospital patient diagnostic tests problem case.

Many problem cases are modeled as binary classification activities but could need advanced techniques. Advanced frameworks could be utilized to alter content of training data by under sampling the dominant category or oversampling the non-dominant category. types are:

- Random Under sampling.

- SMOTE Oversampling.

Advanced modeling models could be used, that concentrates on the non-dominant category when fitting the model on training data, such as cost-sensitive machine learning models. Types include:

- Cost-sensitive Logistic Regression type.
- Cost-sensitive Decision Trees type.
- Cost-sensitive Support Vector Machines type.

To conclude, other key metrics that can be used in conjunction with accuracy. Types are:

- Precision.
- Recall.
- F1 score.

b. Regression Algorithms

According to David A. Freedman (2009). regression analysis is used to estimate a continuous variable. Predicting prices of stocks given the features of stock, price etc. is one of the types of Regression modelling cases. It is a supervised technique. Regression predictive modeling is the activity of approximating a mapping function (f) from independent variables (X) to a continuous dependent variable (y). A continuous dependent variable is a real-value, such as an integer or floating-point value. These are also numbers, such as quantities and proportions. A regression problem case requires the prediction/estimation of a quantity.

- A regression can have real values or discrete independent variables.
- A problem case with multiple independent variables is often called a multivariate regression problem case.
- A regression problem case where independent variables are ordered by time is called a time series forecasting problem case. Because a regression predictive model/algorithm predicts/forecast a quantity/proportion, the metrics of the model must be reported as an error in those predictions. There are many ways to estimate the metrics of a regression predictive model, but perhaps the most common is to calculate the root mean squared error, (RMSE).

a. Types of Regression

- Simple Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression

- Random Forest Regression

- **Simple Linear Regression.** The most popular and interesting type of Regression technique is called the simple linear regression that is used to predict a feature variable Y based on the explained variable X. A linear correlation should exist between dependent variable and explanatory variable hence name Linear Regression. While training and generating a regression algorithm, coefficients are learned and fitted to training subsets of a population. The objective of the training is to look for the best fit line that minimizes cost function. The cost function aids in measuring the error. During the training process, the aim is to minimize the error between actual and fitted values and thus minimizing the cost function. The objective is to get coefficients which minimize the cost function. The most popular cost function is Mean Squared Error (MSE) which is defined to be equal to the mean squared difference between actual and fitted values. The coefficient values are computed using the Gradient Descent technique. The Gradient descent starts with random values of coefficients, calculate the gradient of cost function on these elements, update the coefficients and calculate the cost function again. This process is iterated upon until a minimum value of cost function is gotten.

- **Polynomial Regression.** Here, the original attributes are transformed/changed into polynomial attributes of a given degree and then Linear Regression is done on them. If the degree is scaled to a very high value, leads to an overfitted curve.

- **Support Vector Regression.** Here, hyperplane identification is done with maximum margin so that the maximum proportion of data points are in that area proportion required.

- **Decision Tree Regression.** Decision trees can either be used for classification or regression. At every level, splitting attribute is identified. With regression, the ID3 algorithm is used to identify the splitting node by lowering the standard deviation (in classification information gain is utilized). A decision tree is built by categorizing the data into subsets containing examples with homogenous values. Standard deviation is used to compute the homogeneity of a numerical subset. If the numerical subset is totally homogeneous, its standard deviation is zero.

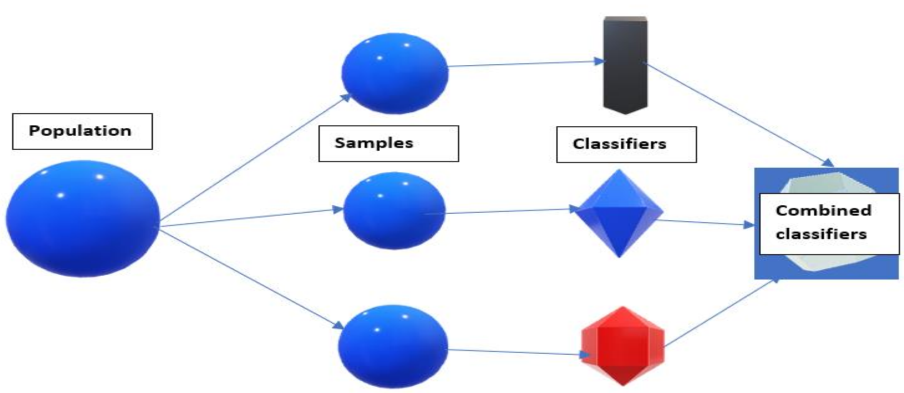
- **Random Forest Regression** is an ensemble technique which considers the estimations of a combination of decision regression trees. Random Forest prevents overfitting by creating random subsets of the attributes and building smaller trees using these subsets.

c. **Ensemble Algorithms** Ensemble learning is a common method for building machine learning models. According Alok k and Mayank j (2020), ensemble techniques combine outcomes of machine learning models using three major ways. Ensemble methods show agility and are versatile in how they perform their machine learning functions. They are powerful and have many advantages compared to other machine leaning models. Ensemble learning are very accurate. Ensemble techniques achieve their high accuracy by putting together the outcomes of machine learning algorithms in many facets; Ensemble learning methods are categorized in three categories: mixing training data, mixing combinations, and mixing models.

I. Mixing Training Data.

Here we separate the training data into multiple chunks, and train separate classifiers on each training subset data. According Alok k and Mayank j (2020) **,** the output of these classifiers is put together (see Figure 2.1). The process is called bagging This technique makes that classifier get enough diversity as they are trained (evolve) on a subset of the total population. By putting together, the output of these diverse learner elements, we get greater accuracy in comparison to where a single learner is trained on the total population (training data).

FIGURE 2.1 MIXING TRAINING DATA USING BAGGING

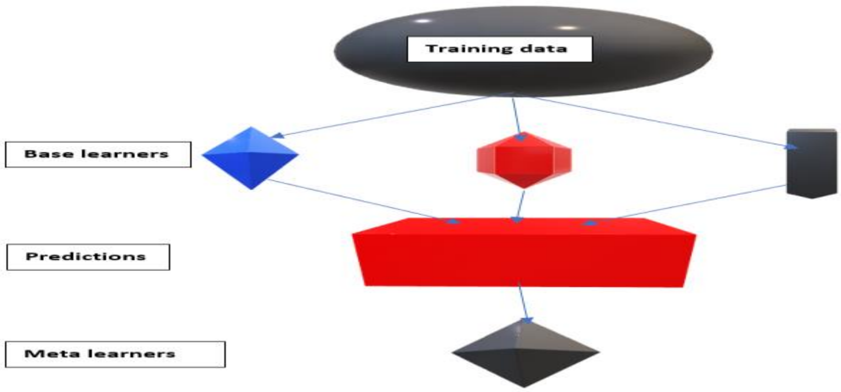


II. Mixing Combinations

- **Boosting.** Here we use a number of algorithms/learners, where every ML learner is trained on a subset of elements. According Alok k and Mayank j (2020), if the algorithm learner has a weak performance, more emphasis is given to that algorithm/learner to improve performance. This joining of algorithms is known as boosting.
- **Stacking.** Here we stack algorithms/learners one on top of the output of another to form stacks of ML models. According Alok k and Mayank j (2020) **,** we train a number of algorithms together to get a predictions/estimations learner. When we merge these predictions, there could

be errors. In stacking, we treat the result of individual predictions as the next training input data. (The first levels of ML models are called base learners.) We stack another layer of machine learning algorithm on top of the base layers; the secondary level is called a meta learner (see Figure 2.2). Both techniques—boosting and stacking—are derived by combining ML algorithm models in a number of ways,

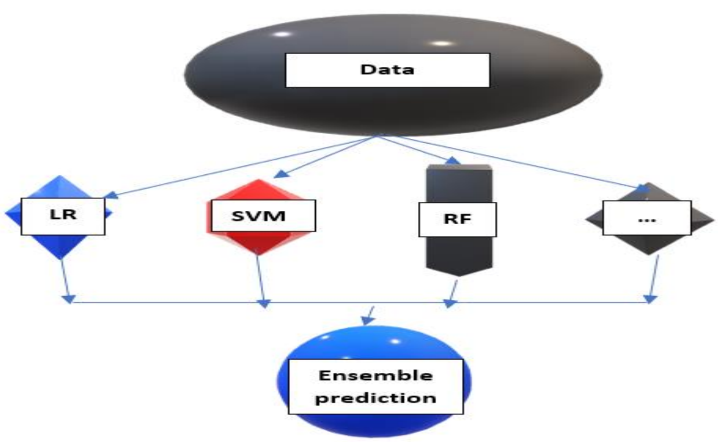
FIGURE 2.2 STACKING ENSEMBLE MODEL



III. Mixing Models.

Several models can be utilized (see Figure 2.3) or within a single machine learning model, by changing settings/ hyperparameters which will be differentiated within training runs. According Alok k and Mayank j (2020), this makes thee model to have better performance metrics with lower bias in comparison to using a single model or singular settings

FIGURE 2.3 ENSEMBLE MIXING MODEL



2.2.4 Types of Ensembles learning techniques

Ensemble techniques combine outcomes of machine learning models. Ensemble learning is a combination of multiple machine learning techniques performed together. Ensemble methods do this through combining the output of machine learning models.

a) Max Voting Techniques

Voting ensembles train different machine learning models. where the same data is trained on different machine learning models for example logistic regression, support vector machine (SVM), and random forest. According Alok k and Mayank j (2020)**, t**he output of these three models is combined to get an ensemble prediction. An election using different machine learning models is done. If it is a classification problem, each ML model votes for a category. In majority voting, the class that earns the most votes is the preferred class. It is widely observed that the resulting class often has higher accuracy than any single model. Their output is then combined using a Voting classifier. When accuracy is measured for each of the individual models, as well as the combined model on the test dataset, one gets a boost in accuracy.

b) Averaging Techniques/soft voting

Averaging is another way to combine the output of different classifiers. According Alok k and Mayank j (2020) **, t**he major difference between voting and averaging is that in averaging, we take the prediction probability of each class separately from the model and then combine the resulting probabilities by taking the average of these predictions. This combination method is called soft voting. The initial steps to train different models are the same, but instead of using the Voting classifier, we do inference of our model on our test dataset, take out the prediction probabilities of each class, and then take out the average of all the probabilities as a resultant class probability on the test dataset.

c) Weighted Averaging Techniques

This approach is like Averaging/ soft voting techniques the only difference is that for soft voting we assign equal weight to all the models when calculating the average output. According Alok k and Mayank j (2020), in the case of Weighted averaging techniques, we depict a specific model as more significant based on runtime performance or any other metric that is deemed superior in comparison to other algorithms. To reflect its importance, we increase the weight of that particular model and decrease the weight of all the other models when calculating the average output.

d) Horizontal voting techniques

Mixing model (voting, averaging, and hyperparameter tuning) work very effectively in classical machine learning. But in situations especially in deep learning where our training data size, training data time, and model size are very large. There might arise cases where training takes too much computation and time. According Alok k and Mayank j (2020), in situations like this, it is often impractical or cost-prohibitive to train multiple models and multiple instances of the same model with different hyperparameters. This is where horizontal voting is considered. During long running machine learning job, one can observe after certain number of epochs of training the model accuracy has stopped improving making it hard to select an accurate epoch time for a model. With horizontal voting ensemble, you save models after a minimum number of epochs. The resulting models are recombined using voting techniques to get an accuracy boost

e) Snapshot ensembles

Snapshot ensembles are an extension of a horizontal voting ensemble. Instead of saving models after the minimum threshold, you modify the learning rate of the model itself. When training a machine deep learning model, it is often desirable to start the initial higher learning and then slowly decrease the learning rate. According Alok k and Mayank j (2020) **, t**his approach leaves a lot of the optimization on the table, as training could be struck at any local minimum. One of the leading solutions for this local minimum problem is a cyclic learning rate, in which we increase and decrease learning rates in cycles. Like horizontal voting ensembles, all the models at each of the local minima states are combined. This approach results in a very good model when compared to using an individual model alone.

2.2.5 Types of Uplifting Modeling Techniques

According to R. Michel, I. Schnakenburg, T. von Martens (2019), uplift modeling is also known as incremental modeling, treatment effects modeling, true lift modeling, or net modeling. Uplift is the increase in likelihood of the outcome *with* the treatment as compared to the outcome *without* the treatment. Uplift modeling can supplement experimental data from A/B testing by identifying the incremental impact on particular individuals of a specific treatment, as opposed to the overall lift or decrease caused by a treatment. This technique may help you assess whether other attributes of those individuals (e.g., demographic characteristics) could help explain their response. This nuanced analysis allows for future targeting only of those most likely to respond positively to a treatment. Uplift modeling is a technique that helps to determine probability gain that the customer by getting the marketing materials will buy a product. The field is relatively new. The two most common approaches are (Lee 2018):

a) Two Model

In this method two classifiers are built. The one is trained on observations that received treatment and the second is trained on observations that didn't receive treatment. Thereafter, the uplift for particular observations is calculated. If the observation experienced treatment, then it is an input to the *model one* and the probability that the customer will buy a product is predicted. Next, it is investigated what could happen if the customer didn't receive treatment. In that case, the treatment indicator in observation's feature is changed to 'zero'. This kind of modified record is an input to the *model zero* that predicts the probability that specific subscriber will accept an offer. The uplift is calculated as the difference between the outputs of the *model one* and *model zero*. The higher the difference, the more profitable it is to address retention with distinct offer to a specific subscriber. Analogically, uplift is computed for the people that didn't experienced treatment.

b) One Model

This approach is similar conceptually to the *Two Model* method with such a difference that instead of building two classifiers only one is used. Therefore, every observation is an input to the model that generates prediction. Later, the indicator in the treatment column is changed into the negation and such a vector is used as input to the model that once again outputs probability that the customer buys a product. The uplift is the difference between the two predicted probabilities.

FIGURE 2.4 TWO MODEL APPROACH (OWN ELABORATION) WHERE P1=P(CHURN |T=1) & P0=P(CHURN |T=0)

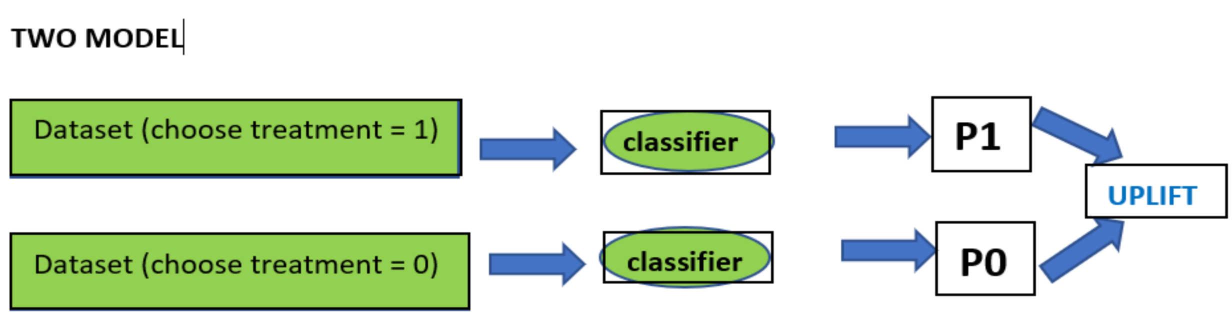
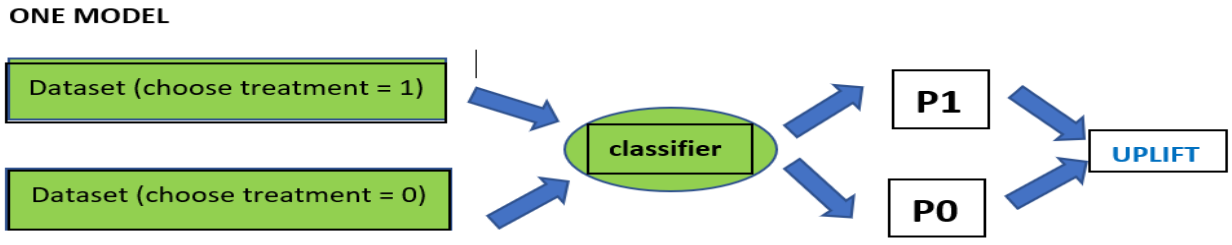


FIGURE 2.5 ONE MODEL APPROACH (OWN ELABORATION) WHERE P1=P(CHURN |T=1) & P0=P(CHURN |T=0)



<u>How do we model Uplift?</u>

a) Traditional Propensity Model

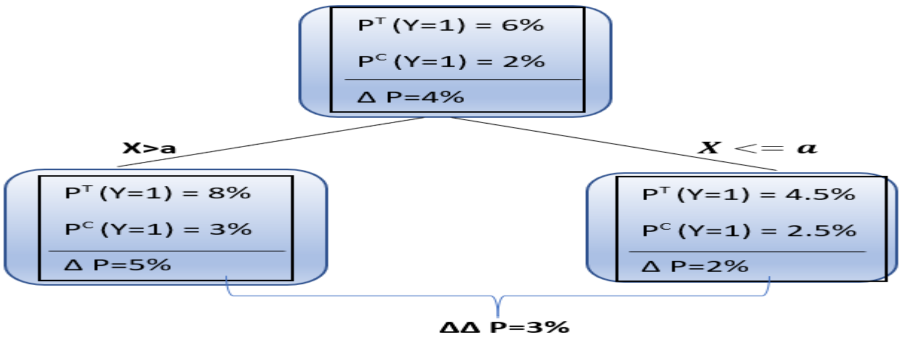
The application of mathematical models to data to try to predict whether someone will take a particular action. In other words, it's a way of identifying who among your audience is most likely to make a purchase, accept an offer, sign up for a service, or to leave. Whereas propensity modelling is concerned with modelling the response to our customer retention campaign, uplift modelling attempts to quantify the incremental response as a direct result of the customer retention offer.

c) Direct Modelling

There are two main approaches here – direct modelling and indirect modelling. The main difference between the two approaches stems from how we measure and evaluate uplift models.

In direct modelling, we are "directly" modelling the difference in probabilities between two distinct groups. There are many approaches to do so, with almost all of them relying on tree-based algorithms, slightly altered to accommodate uplift modelling. Tree-based models are ideal, as they naturally model at group level by iteratively splitting a group in a further two groups with every splitting decision. Where traditional tree-based models are designed to split the data into smaller and smaller homogeneous groups, uplift models instead are designed to split our customers into heterogeneous groups each time they split (by maximizing a measure of uplift). They are using certain splitting criterions, such as Kullback-Leibler divergence, Euclidean Distance, p-value or Chi-squared Distance. Like traditional tree-based methods, hundreds of trees would be fitted in an ensemble fashion. The image below shows a simplified Uplift tree grown to depth 1.

FIGURE 2.6 UPLIFT TREE (OWN ELABORATION)



d) Indirect Modelling/ Meta-Learner Models

According to R. Michel, I. Schnakenburg, T. von Martens (2019), indirect Uplift modelling techniques (meta-learners) are regular response models that are then re-purposed to infer uplift and can be based on any base algorithm. We're not attempting to optimize some measure of uplift directly and are instead modelling the expected value of the response, for different treatments. For our customer retention campaign, we would compute the probability that the customer is not going to churn if we send the Offer through SMS and the probability of churning if we do not send an offer through SMS. The difference between the two estimated probabilities is the estimated uplift. This can be implemented in a two-model approach (separate model fitted to all control / treatment groups) or a unified model (single model with the allocated treatment part of the feature space).

e) Evaluation of Uplift Models

We cannot view uplift directly as it's impossible to both get the Offer through and not get it. This not only determines how we go about building models, but also how we can evaluate how well the model performs. This means that common evaluation metrics for classification problems such as Precision, Recall or Accuracy are not useful to us out-of-the-box and we need to use non-standard metrics to evaluate model performance. We use two intuitive graphs to explain and evaluate these models. In the figure 2.8 the Qini bar-plot, you can see the actuals of net lift for each decile of predicted uplift; the deciles providing the grouping. This model is performing very well – we see a monotonically decreasing lift as we go from the best customers to the worst. Note that the bottom 5 deciles have a negative lift, which means that it would be better to not send these customers the SMS offer; it makes them less likely not to churn! The average net lift is plotted horizontally, this shows that deciles 1-5 are above the average (4 borderline). In the figure 2.7 the Area Under the Uplift Curve, demonstrates the cumulative version of the bar plot. The black line represents the 'random' model and the green line is the cumulative lift. Essentially, we're trying to maximize the area under this curve; if there is one metric to optimize for, then this is most likely the best metric to quote, in the absence of a better understanding of the business application. We can see that we reach a global maximum at the 3rd decile – so in real life, this might be a sensible number to treat (depending on budget and other constraints of course).

FIGURE 2.7 AREA UNDER THE UPLIFT CURVE

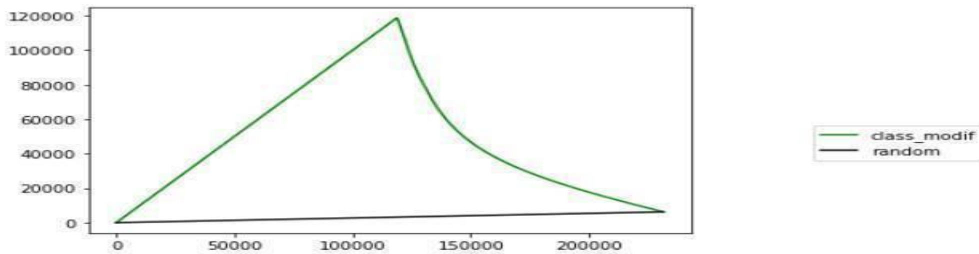
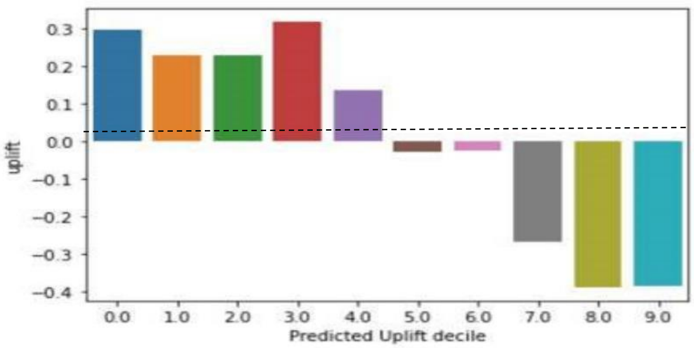


FIGURE 2.8 QINI UPLIFT BAR PLOT



For our study we will use the independent variables described in the table2.1 below

TABLE 2.1 INDEPENDENT VARIABLES TABLE

Age	**Continuous variables between 6 -10 Years**
Gender	Male Female
Sales region	Coast
	Mount Kenya
	Nairobi East
	Nairobi west
	Greater Western
	unspecified
	Rift

ARPU	Voice
	SMS
	Mobile money
	Subscription
	Data
	Other subscription
Age on the network	0-20 years
Calls by subscriber to call center	one to thirty calls
Subscriber interactions	interaction category
	interaction sub-category
Segments	Youth
	Hustler
	Mass
	Discerning professional
	unspecified
subscriber lifetime value	Sum of all ARPUs multiplied by age on the network

2.3 Empirical Review

The review of empirical literature allows evidence-based and factual based analysis of related works done locally and internationally in the same area of study or related studies. Here I will review most of the current ensemble learning and uplift modelling models that were built for subscriber churn prediction. As we will see they are not many that have employed ensemble learning for uplift modelling method for churn conversion prediction.

2.3.1 Application predictive algorithms for churn in the telecommunication Sector

Churn management is a pertinent issue in any industry more so the telecommunications sector. Several data mining and machine learning techniques have been utilized across the board to try to predict probable churn population. More recently the emphasis has been to predict churners that are profitable that would make the service provider some money. Other methods that have been explored include churn for cost reduction or increase of savings. In this study we will review three previous studies that have applied the review uplift algorithms in various sectors sector.

2.3.1.1 Enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling Ammar A. Q. Ahmed and D. Maheswari 2020

Regular classification technique would not conclusively predict churn because of lack of congruence between algorithm metrics and business goals. The proposed model/framework by Ammar A. Q. Ahmed and D. Maheswari 2020 classifies the churn population into three broad categories, namely, active, passive and rotational. This study presented an ensemble stacking model framework together with cost savings-based strategies for telecommunications churn prediction framework. Analysis have been performed based on model metrics key performance indicators (True negative rate (TNR) and False positive rate (FPR) and a savings/loss function measurement criterion with a great emphasis on the cost savings measurement/loss function. This process mode shows a high relationship between key performance metrics and business aims, thus making the said algorithm suitable for most cost or savings sensitive projects. A heterogenous ensemble model applicable to stacking or bagging strategies to accurate key performance metrics. Any incorrect predictions are handled at the tertiary level using a cost savings measurement-based blender to provide the final key performance metrics estimations. Blender measurements are adjusted according to the cost savings targets to better predict key performance metrics concentrating on business aims. eventually, Subscriber cost savings uplifting is implemented on final predictions, 50% more cost savings are achieved through this technique compared to the other ensemble models. The cost savings matrix that is used was proposed by Elkan et al. [14] and was created and metrics analyzed for accurate churn prediction by Bahnsen et al. [15].

TABLE 2.2 COST CALCULATION TABLE

Predicted	Actual	
	churn	Not churn
churn	$y(Co + Ca)$+(1-y) (CLV + Ca)	$Co + Ca$
Not churn	CLV	0

Where c is the number of subscribers kept within the network, CLV is related to Customer/subscriber lifetime value [16], Co relates to cost the service provider will forgo to give of the offer/campaign and Ca relates to the cost a service provider will forgo by contacting via

Call or SMS the subscriber. This is normally calculated heuristic based on call handling sustenance model applied for customer experience. The researchers recommended a cost calculation associated with churn campaigns reflected using Table 2.2. By integrating data from confusion matrix into the cost matrix the researchers were able to perform a cost calculation. The cost calculation they came up with is shown in Eq. (1).

$$Cost_i = y_i\,(c_i\, c_{TPi} + (1 - c_i)\,(c_{FNi}) + (1 - y_i)\,(c_i\, c_{FPi} + (1 - c_i)\, c_{TNi}) \quad (1)$$

where c_i is the predicted label and y_i is the actual class label, $y \in \{0, 1\}, c \in \{0, 1\}$ referring to non-churn and churn classes, C_{TP}, C_{FP}, C_{TN} and C_{FN} are derived from the cost matrix. From his results he used Random Forest and decision trees as his base learners and then uses gradient boosting as his blender. The methodology he has utilized is summarized as below

1. Segregate independent variables to training and test data

2. Apply the selected models on the training subset of data

3. Compute performance measure from the resultant estimations

4. Select the best model for estimating churners a. select true Selection model that satisfy the true Threshold criteria (True Positive Rate>true Threshold) b. If no model meets the true Threshold criterion, 30% of the best performing models are taken into account.

5. Select the best models for estimating non-churn population a. Identify false Selection models that satisfy the false Threshold criteria (TNR> false Threshold) b. If no algorithms/models meet the false Threshold criterion, 30% of the best performing algorithms/models are used.

6. implement true Selection models on testing subset data to get the true Data

7. Get true estimations using true datasets & remove duplicates

8. implement false Selection models on testing subset data to get false data

9. Get false estimations using false data & remove duplicates

10. Implement intersection to discover data found within true data & false Data

11. Discover best true estimation models judged on True Positives & cost savings

12. implement best true estimation models with common data

13. Estimations by maximum proportions of models vindicating the claim as taken as the final estimation parameters.

In this study derived calculations have formidable cost savings of almost 50% to a telecommunication's service provider due to the inclusion of subscriber uplift performance metric methods. With accurate churn estimation levels, it came with a downside of exhibiting high levels false negative rates. This model/algorithm benefits service providers by efficiently achieving business aims and by increasing savings from the low costing efforts in comparison to other algorithms. The proposed stacking ensemble method hopefully will enhance the prediction process and provided an accuracy metric of above 90%.

2.3.1.2 Managing Churn to Maximize Profits Aurélie Lemmens and Sunil Gupta (2017)

In this study, Aurélie Lemmens and Sunil Gupta (2017) demonstrated how their profit-based loss function greatly improved the overall profitability of retention activities by ensuring accurate prediction of high-profit subscribers in comparison to low-profit subscribers. They discovered that using this new loss function led to, on average, to a 62% increase in profit lift with no additional costs. In this study, they applied the profit-based loss function to stochastic gradient boosting (SGB) as it showed superiority in predictive performance (Hastie, Tibshirani, and Friedman 2009). SGB is a greedy numerical optimization algorithm, created at Stanford University by Friedman and co (Friedman 2002; Friedman, Hastie, and Tibshirani 2000). It works by sequentially combining the predictions made by regression trees (Breiman et al. 1983).

Intuitively, SGB works by making an initial guess of each subscriber's propensity to churn and then, it tries to estimate the residual errors by fitting a tree. At each repetition, a new tree is estimated to fit the residuals of the previous repetition. The estimation runs until no improvement can be realized. Below is an explanation of regression trees and how they are used for SGB. The idea behind SGB can be summarized as follows. Before estimation, a loss function Ψ is identified and utilized at every iteration to calculate "residual error" between the fitted and actual values, denoted as $F_m(x_i)$, and actual values y_i. Once the loss function is derived, the estimation starts by setting each observation to an initial number, denoted by $F_0(x_i)$, which is typically set as follows:

$$F_0(x_i) = 1\ 2\ log\ (\ p0\ 1{-}p0\ )\ , \qquad (2)$$

where $p0$ is the percentage of churners in the calibration sample. Thus, $\hat{F}_0 (x_i)$ can take any value in]−∞, ∞[. From this initial guess, we can calculate the error (i.e., difference between the fitted values $\hat{F}_0 (x_i)$ and actual values y_i). The next activity consists of fitting a tree model $T (x_i, \Theta_0)$ of the errors against the predictors x and calculating the fitted values of these errors. The proportion of terminal nodes is kept small (max 8 nodes) to avoid overfitting. These fitted errors are then put together with the predicted values $\hat{F}_0 (x_i)$. The combination of both values is called the "boosted" fitted values (i.e. the original guess is boosted with the fitted errors) and can be denoted $\hat{F}_1 (x_i)$. This process is repeated to calculate the error from the boosted fitted values (i.e. difference between the fitted values $\hat{F}_1 (x\ )$ and actual values of y_i), fit a tree model of the new errors, and combine the fitted values of these new errors to $\hat{F}_1 (x_i)$. We repeat these steps M times until the model converges.

They calculate the optimal target size of the retention campaigns by taking into consideration the trade-off between reducing the loss from defection and cost of retention activities. The outcome showed the profit-based loss function led to a profitable retention campaign in comparison with techniques that ignored profits during estimation as well as heuristics based on managerial judgment. They argued, that these profits did not incur additional cost to service providers since the actioning of retention campaign remain relatively constant, only the constituents and proportion of the target group changed.

To derive the profit of targeted retention actions they based it on the conceptual framework proposed by Neslin et al. (2006), they postulated that profitability of a retention campaign was given by

$\Pi = \sum \pi i\ N\ i \in$target, (3)

where N is the total number of subscribers and πi is the lift in profit generated when targeting subscriber i with a retention action, which we refer to as profit lift. It is captured by

$\pi i = \gamma i\ (\Delta ECLVi - \delta)$, (4)

where $\Delta ECLVi = ECLVi\ ' - ECLVi$, is the difference between the expected residual CLV (Fader and Hardie 2010) of subscriber i if she is targeted ($ECLVi\ '$) and the expected residual CLV of subscriber i if she is not targeted ($ECLVi$), γi is the probability that the targeted subscriber would accept the retention offer and stay with the firm, and δ is the action cost. Note an assumption is that the cost of the retention activity is constant and is not additional cost is incurred when the subscriber rejects the offer and leaves (γi = 0).2

Their profit definition is similar to the one suggested by Neslin et al. (2006) and Provost and Fawcett (2013, p. 286) and relates to the work of Ascarza (2016), who proposes to target subscribers with the highest lift, defined as the change in churn probability due to targeting. This notion is captured by γi the probability that a targeted subscriber accepts a retention offer and stays with the firm. their profit formulation extended her definition of lift by considering two new elements: the net value that the subscriber brings to the firm, and whether the subscriber would churn or not in absence of a retention action (Provost and Fawcett 2013).

They rewrote equation (2) depending on whether the targeted subscriber would have churned or not in absence of a retention action. Let $y_i = +1$ if subscriber i would churn in the next time period if no action is taken, and $y_i = -1$ if subscriber i would not churn if no action is taken. For a would-be churner i, $ECLV_i = 0$ in the absence of a retention action. Therefore, her expected profit lift simplifies to $\pi y_i{=}{+}1 = \gamma i\ (ECLV_i{}' - \delta)$.

They developed a profit los matrix shown in Table 2.3.1.2, they start from the (dichotomous) case where subscribers with $F\,(x_i) \geq 0$ are targeted and those with $F(x_i) < 0$ are not. The optimal decision would be to target subscribers that would generate a positive profit lift ($\pi_i >$ 0) and discard subscribers with a negative or zero profit lift ($\pi_i \leq 0$). We categorize subscribers based on whether $\pi_i > 0$ or $\pi_i \leq 0$ and $F\,(x_i) \geq 0$ or $F\,(x_i) < 0$. The new margin sign $(\pi_i)F\,(x_i)$ is positive in the diagonal and negative in the off-diagonal elements of Table 2.3

TABLE 2.3: PROFT BASED LOSS MATRIX

Loss Matrix	**Targeted Subscriber** $F(xi) \geq 0$	**Non-Targeted subscriber** $F(xi) < 0$
positive expected profit lift	**Positive margin**	**Negative margin**
$\boldsymbol{\pi i > 0}$	Targeted profitable churner $\psi = 0$	Non-Targeted profitable churner $\psi = \pi i$

Negative expected profit lift	**Negative Margin**	**Positive margin**
$\boldsymbol{\pi i \leq 0}$	Targeted non-churner or targeted unprofitable subscribers $\psi = \lvert\pi i\rvert$	Non-Targeted non-churner or non-targeted unprofitable subscribers $\psi = 0$

In the upper-left hand quadrant of this table, they targeted profitable churners and in the lower-right hand quadrant, they did not target unprofitable subscribers, which include non-churners and unprofitable churners. By targeting the right subscribers in the diagonal quadrants, they incurred no loss, i.e., Ψ = 0. In the off diagonals, the margin is negative. The loss associated with not targeting a profitable churner (upper-right quadrant) is $\pi_i = \gamma_i\,(ECLV_i\,' - \delta)$. Similarly, the loss associated with targeting a non-churner is $|\pi_i| = |-\delta| = \delta$, and the loss for targeting an unprofitable churner is $|\pi_i| = \gamma_i\,(\delta - ECLV_i\,')$ when the cost of 11 retention offer is higher than subscriber's expected residual CLV. The profit-based loss matrix can directly translate into a so-called 0-1 loss function, i.e., $\Psi 0{-}1(\pi_i, F\,(x_i)) = |\pi_i|\, I$ (sign $(\pi_i)F_i < 0$). Given that this loss is non-differentiable, they extended their example to the continuous case and wrote the profit-based loss function by including πi into the binomial function. Using our new margin sign

$(\pi_i)F\,(x_i)$, $\Psi\,(\pi_i, F\,(x_i)) = |\pi_i\,| log\,(1 + e\,{-2}\text{sign}\,(\pi_i)F_i)$. (5)

This study also contained several limitations that were fruitful research opportunities. They did not focus on the actual response rate of retention campaigns offered to the subscribers, which would allow us to estimate the response probability of a given action, e.g., by using repeated observations of subscribers' response to past promotions (Neslin et al. 2009). A pertinent challenge would be to study variations in subscriber response depending on the type and breadth of the retention activities, and to derive the retention cost at which the response metric would be maximal (Ventakesan et al. 2007). Secondly, their approach does not model the time dynamics in a subscriber's propensity to churn. Using longitudinal data would allow us to ignore this assumption by jointly modeling future ARPU of subscribers and their defection behavior (e.g. Ascarza and Hardie 2013, and Schweidel et al. 2011). Third, they did not model the possibility that a subscriber's decision to churn might depend on her expectation of the firm's retention offers in the future (Lewis 2005). The recent rise of this phenomenon makes it an interesting area for future research.

2.3.1.3 A Subscriber Churn Prediction Model in Telecom Industry Using Boosting

Ning Lu, Hua Lin and Jie Lu (2012) did a real-world investigation on subscriber churn estimation and proposed the utilization of boosting algorithm to boost a subscriber churn prediction algorithm. In comparison to researches/studies that used boosting to improve the base learner accuracy, this study divides subscribers into 2 weight-based clusters so as to identify higher risk churners. Logistic regression was utilized as a base learner throughout this analysis and a churn prediction algorithm was developed on either cluster. The outcome was corelated with a single logistic regression algorithm. Analysis showed that boosting gave a good categorization of churners and non-churners; thus, boosting was suggested as to be viable to do prediction analysis for churn.

According P. Datta, B. Masand, D. R. Mani, and B. Li (2000), Boosting refers to the effective technique that tries to 'boost' the accuracy of a machine learning algorithm. Most boosting models involve continuous learning and weak learners' addition to come up with a significant meta learner. Each weak learner that is added is weighted accordingly using accuracy as a metric and trained with reweighted training data. According to these researchers, there exists two types of churn related behavior: voluntary, where a subscriber decides to cease services; and involuntary, in which the service provider decides to cease a subscriber's services (typically because of unpaid bills or fraudulent activities). The study considered voluntary churners, as involuntary churners are easily identifiable and are of less significance from a churn management point of view. They analyzed the metrics of their churn estimation algorithm using a training subset of subscriber sample data gotten over a 6-month period. Each subscriber was categorized in two pre-defined categories and his/her churn propensity was monitored and updated according to his/her latest 3-month sample data so as to emulate the churn prediction realities.

This study conducted an investigation of subscriber churn prediction/estimation based on real-data sets. Their model/algorithm allowed for an "Implementation Zone" where subscribers with the greatest churn likelihood being focused on for retention purposes. Instead of boosting a base learner using an algorithm, this study categorizes the training data based on the difficulty of fitting a base learner and built an estimation model for each defined group. The results were tested on a live data-streams and corelated with a logistic regression model fitted to an entire training set.

Assessment shows that churners are highly skewed, the weights given by the AdaBoost algorithm suggested quality categorization that defined accurately high-risk subscriber churners. To improve performance and limitations of the estimator, they proposed trial of the other classification methods to deal with class imbalance. Another limitation was it provided accurate churn prediction which provided a basis for producing subjects for targeted offers but did not identifying the reason for a subscriber's churn behavior or whether retention was because of the model or other marketing campaigns.

2.3.1.4 Why you should stop predicting customer churn and start using uplift models Floris Devriendt, Jeroen Berrevoets and Wouter Verbeke 2019

A real world research on the subscriber churn uplift model was conducted in 2019 and suggested the use of uplift modeling to optimize the subscriber churn prediction algorithm. He Created a novel appraisal criterion called the Maximum Benefit Uplift (MBU) measure that enables maximum possible profit-based success assessment that can be accomplished by implementing an uplifting model. He concluded that there was a lack of case studies that explicitly contrasted classical churn and uplift modeling methods, and that the success of predictive models and uplifting models could not be linked adequately because the performance assessment approaches were distinct. He represented the lift-up curve and the measure used to evaluate the uplift algorithms as being taken side by side with the lift curve and the measure used to assess the predictive models. In brief, the contributions of researchers are outlined below:

1. They apply the maximum profit factor to evaluate the uplift models by adding the maximum uplift profit measure.

2. In addition, lift-up curves and lift-up measures were adopted to evaluate uplift models corresponding to the lift curve and lift metric used to assess predictive models.

3. They also provided empirical proof of the benefits of uplift algorithms over predictive modeling by conducting a case study using a financial institution data .

They expanded a campaign profit formula by Neslin [23]. The campaign formula stated as below

$$\pi = N\alpha\, [\beta\gamma(b - C_{contact} - C_{incentive})\ \beta(1-\gamma)(-C_{contact})(1-\beta)(-C_{contact} - C_{incentive})] - A \qquad (6)$$

The interpretation of the formula for benefit is shown by dividing P into five parts.:

(a) $N\alpha$ reflects the number of subscribers targeted by the offer; with the exception of the fixed admin costs represented by A, only the targeted subscribers will accrue costs and benefits as a result of the offer.

(b) $\beta\gamma(b - C_{contact} - C_{incentive})$ construes the net profit provided by the offer that equals the savings by the business due to less churn costs of the offer, $b - C_{contact} - C_{incentive}$; multiplied by the proportion γ of potential churners within the correctly identified potential churners β targeted by the offer.

(c) $\beta(1-\gamma)(-C_{contact})$ construes the cost related to giving offers to subscribers that were predicted not to have churned but churned anyway.

(d) $(1-\beta)(-C_{contact} - C_{incentive})$ denotes the cost resulting from targeting subscribers that were not to churn with an offer and subscribers took up the offer.

(e) A reflects the fixed admin cost such cost of sending an offer through SMS that impacts the bottom-line profitability of a retention offers.

In his experiments he uses logistic regression and random forest to come up with the profit curves. The MP measure is utilized to assess the performance of logistic regression and random forests, three different sets of variables are used to estimate MP on the basis literature reported figures and is comprised of low, medium and high returns resulting from Subscriber retention. The results were shown to be consistent whichever parameter values were used. The profit curves show the profits earned per subscriber of the population of the total customer base of targeted subscribers with the retention offers. The researchers ranked the values based on estimated probability of churn and based on the estimated uplift score. They plotted profit earned per subscriber as it is independent of the size of the subscriber base and commensurate to the total profit. In their conclusion the profit curves show the adequate proportion of subscribers to be targeted by the offers, resulting in the maximum profit.

2.3.2 Research Gap Analysis

From the literature I have reviewed less than 20 Highlights the ineffectiveness of using traditional churn forecasting in two field trials where uplift algorithms are shown to be more accurate in consumer retention. None of the literature I have looked at have inculcated stacking ensemble model in their uplift models. Therefore, in this dissertation we corelate subscriber churn prediction (SCP) and subscriber churn uplift (SCU) modeling for Subscriber retention by comparing their performance when they are applied to a case study in the Telecommunication sector in Kenya. SCP and SCU models are evaluated using separate evaluation methods as their outputs are different. For evaluating SCP models, the receiver operating characteristic (ROC) curve and the lift curves forms the basis of performance evaluation, enumerated using the area under the ROC curve (AUC) or top-decile lift. According to Mauricio AP et al (2016), in performance evaluation for uplift models, Qini curves or uplift-per-decile plots are utilized, & performance is enumerated using the Qini coefficient or top-decile uplift. In this dissertation, we develop an evaluation procedure for churn prediction by inculcating stacking ensemble model into uplift modelling to generate better performance. In Uplift Modeling the main point lies in making good and reliable predictions about the subscribers that are most likely to churn *because* of a treatment, such as an offer through SMS. Predicting the right customer type and measure if they are e.g., customers that might churn or accept an offer because of the treatment is a clear focus of Uplift Modeling (Radcliffe, 2007). The goal of Uplift Modeling is the estimation of difference in an outcome (e.g., Churn) probability of an individual if he/she receives a treatment or not (Radcliffe and Surry, 2011):

Uplift Modeling deals with customer targeting, which focuses on targeting the customers who are most likely to respond *because* of a treatment. This means, that these customers wouldn't respond if they wouldn't have been treated (Devriendt et al., 2018). They represent the lift-up curve and lift-up measure that characterizes the performance of any uplift model. These serve as counterparts of the lift curve and lift measure sed in assessing predictive models. The Uplift model ensemble algorithm discussed in this dissertation will be used to compare the performance of SCP and SCU models developed using ensemble stacking and extra trees classifier methods in an experimental telecommunication company case study. In summary, my main contributions are summarized below.

1. the researcher expands uplift modeling by introducing the ensemble stacking using the m-lens python library and extra trees ensemble methods.

2. The researcher provides proof of the upside of uplift modeling versus predictive modeling by doing an experimental case study using our novel technique.

This dissertation is arranged as follows; the researcher first talks about subscriber churn prediction modeling then discussing the combination of uplift modeling as an alternative approach with predictive modeling. Next, we discuss the ensemble stacking and extra-trees classifier that is inculcated in uplift modelling using the m-lens library in python in the novel approach of building the SCU model. nest, we present the design of the telecommunication service provider in Kenya case study and thereafter discuss the results of our experiments. At the end, we present our conclusions.

2.4 Conceptual Framework

Mugenda and Mugenda (2003), define a conceptual framework as a hypothesized model showing the concepts under research and their correlations. According to Robson (2011) conceptual frame work is a collection of concepts, assumptions, needs, beliefs, and theories that supports and contributes to research, it is a key part of design. This section presents a new approach for churn prediction that we have named the Telco Churn Analysis Framework (TCA framework) This Framework tries elaborating the practical steps that any Telco firm can take while doing churn analysis in their company. It covers four main stages as listed below

1. **Data literacy.** This step involves identifying subscriber churn data and analytics opportunities throughout the business. This involves identifying business aspects that contribute directly or indirectly to churn as these variables will be included as part of the model.
2. **Data Acquisition and Governance:** This is an important step as there are legal frameworks how data can be used and shared based on data privacy laws. In this step you ensure any data that will be used is either anonymized on its standards are within the set regulations to allow for preprocessing of the same. This ensures identity protection of the subscribers who might be used for the study
3. **Knowledge mining and machine learning:** In this step we apply knowledge mining and machine learning algorithms to capture insightful stories with the data to make them actionable. Machine learning helps in identifying subscriber population that is highly likely to churn.
4. **Business Implementation:** in this step we use the output of the knowledge mining and machine learning process to design and structure campaigns that will retain the identified population is likely to churn. This process involves budgeting for the campaigns and review of the success rate of such campaigns based on the output of the machine learning algorithms.

This steps also informs the people requirements, process transformation and decision-making leg for churn management in the firm.

FIGURE 2.9 TELCO CHURN ANALYSIS FRAMEWORK (TCA)

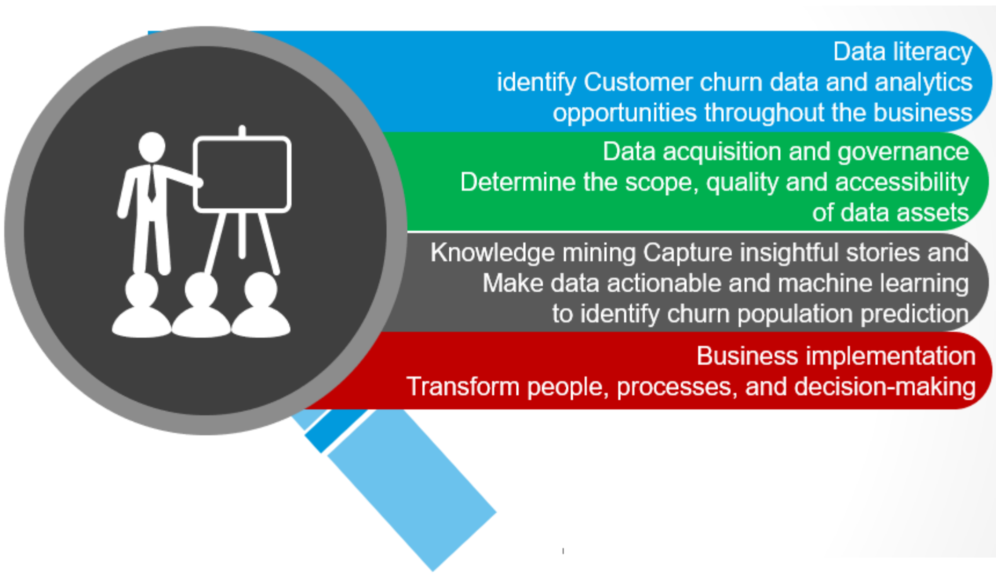
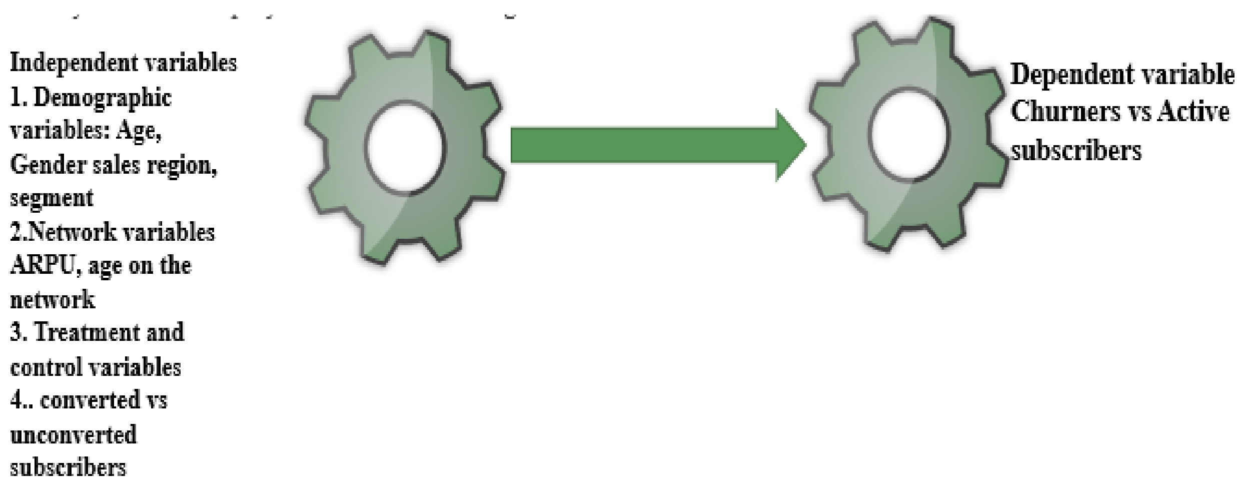


FIGURE 2.10 CONCEPTUAL FRAMEWORK FOR TELCO CHURN MODEL



2.5 Operationalization of Variables

TABLE 2.4 THE VARIABLE OPERATIONALIZATION TABLE

Variable	sub-variable	Indicators	Values
independent variable	Demographic variable	Age	continuous variables between 6 -10 Years
		Gender	Male Female
		Sales region	Coast
			Mount Kenya
			Nairobi East
			Nairobi west
			Greater Western
			unspecified
			Rift
	Subscriber network variables	ARPU	Voice
			SMS
			Mobile money
			Subscription
			Data
			Other subscription
		Age on the network	0-20 years
		Calls by subscriber to call center	one to thirty calls
		Subscriber interactions	interaction category
			interaction sub-category
		Segments	Youth
			Hustler
			Mass
			Discerning professional
			unspecified
		subscriber lifetime value	Sum of all ARPUs multiplied by age on the network
dependent variables		Churn	Active subscribers
			Churned Subscribers

2.6 Summary

In this chapter we have looked at the empirical, conceptual and theoretical frameworks for this study. We have look at various literature with an emphasis from Aurélie Lemmens and Sunil Gupta (2017) Managing churn to maximize profits, A Subscriber Churn Prediction Model in Telecom Industry Using Boosting by Ning Lu, Hua Lin and Jie Lu and an enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling by A. Q. Ahmed and D. Maheswari to try to form the basis of this study. This has laid the groundwork for this very interesting research work that will be conducted.

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1 Introduction

The research methodology chapter introduces the overall methodology that will be adopted in building the descriptive and predictive model for Telco subscriber population. This chapter also indicates how the methodology approach fits the overall research design. It also explains the sampling technique employed, the target population, data collection method used, the experiments that will be undertaken and how the results are analyzed and interpreted

3.2 Research design

The primary data (both quantitative and qualitative) will be collected for the study. According to (Yin R. k, 2014) a case study allows deep understanding of the research area the researcher in focusing on. The case study was research design was carried out in Safaricom PLC. Data was collected through opensource software called NIFI from the Data warehouses of all subscribers in telecommunications service provider. The historical data was collected from the source data base and put in a staging area ready for analysis. Thereafter it was ingested using SPARK and stored in a PostgreSQL Database. The required sample of data will be stored as a flat file ready for preprocessing, analysis and prediction on Jupyter notebook. There after

FIGURE 3.1 RESEARCH METHODOLOGY FOR TELCO CHURN UPLIFT MODELLING

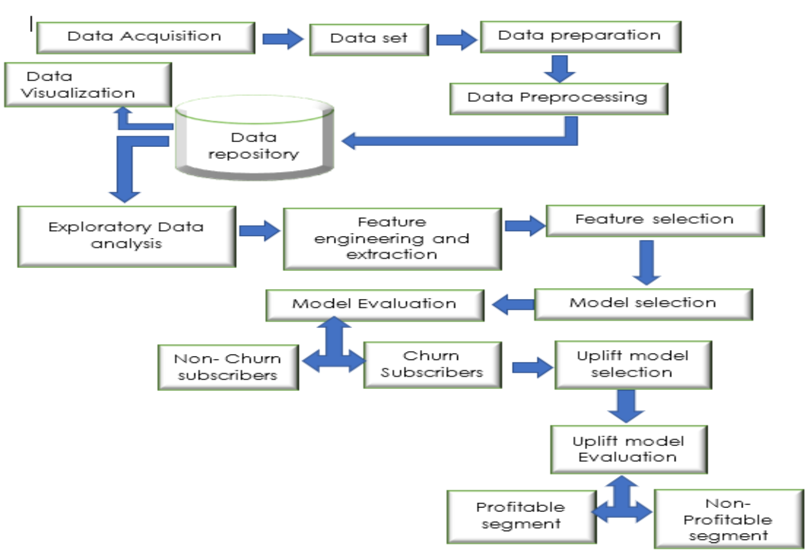
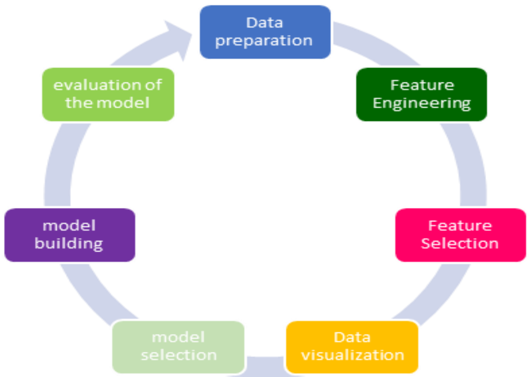


FIGURE 3.2 RESEARCH DESIGN FOR TELCO CHURN UPLIFT MODELLING



Using the Jupyter notebook the data was analyzed by preprocessing the ingested data, performing some feature engineering and selection. The data was visualized using the appropriate visualization libraries. Appropriate model selection process was undertaken before building the model. The results of the model were evaluated using the appropriate evaluation metrics such as ROC Curves, Confusion matrix, AUC curves, Qini curves, ROC curves and uplift bins, AUUC Curves and results tabulated.

3.3 Target Population

The target subscribers were registered prepaid subscribers with Safaricom PLC in the 6 sales regions of Nairobi East, Nairobi West, Rift, Coast, Greater Western and MT Kenya between August 2020 to October 2020

3.4 Sampling and Sampling Procedure

In this study, the probability sampling technique will be used as this technique is a variation of the Kothari sampling which permits specification of the probability of each subscriber segment being included in the sample. To determine the sample size of the subscriber population, a variation of standard statistical formula as mentioned by Kothari (2004) will be used. Sample size will be calculated at 95% confidence level and 5% margin of error. We have added y constant based on data of the total subscriber base.

The method for determining the size of the sample is as follows:

$$n = (z^2 .p.q.N)/(e^2.(N-1).y^2 + z^2.p.q) \tag{6}$$

Where,

n = Sample size; z = The value of the standard variate at a given confidence level and to be calculated from table matrix representing area under Normal Curve. For this investigation standard deviation will be taken at 95% confidence level =1.96;

p = Sample size, based on experience will be used. In the current study value of p will be estimated as 0.50; q = 1-p (In the present study q=1 - 0.50 =0.50); e = Acceptable margin of error (the precision), usually considered 0.05; and N = Size of population.

y = is the value of sample population derived from domain knowledge of the telecom industry

Whenever a sample study is made, there arises some sampling error which can be controlled by selecting a sample of adequate size. Any person undertaking research will have to specify the precision that he wants in respect of his estimates concerning the population parameters. For example, when estimating the mean of the universe within ± 3 of the true mean with 95 per cent confidence. In this case we will say that the desired precision is ±3, meaning that the acceptable error, e, is equal to 3. Keeping this in view, by adding the y component and maintaining its desired precision in my view will not impact the overall outcome desired and we can use the above in the determination of sample size of a telco sample size or any quantitative research if the precision is ensured. The y value is an existing technique in telecommunications sector normally applied in generating sample for network analysis. This has now been applied in this study for churn metrics.

$$1.96 * 1.96 * 0.5 * 0.5 * 35000000/(((0.05 * 0.05) * (35000000 - 1) * 0.0222 * 0.0222)) + ((1.96 * 1.96) * 0.5 * 0.5) = 762501 \text{ subscribers} \quad (7)$$

3.5 Research Instrument

The research instrument here will be Jupyter Notebook and SQL

3.6 Validity and Reliability of the instrument

Validity is defined as the amount to which a concept is accurately measured in a quantitative analysis. Reliability relates to the consistency of measure. In order to establish the content validity of measuring instrument I identified overall content to be represented by making sure variables were randomly chosen with a six-month period. To represent the information in all areas. By using this method, the researcher that the selected variables were representative of the content of trait and properly measured. Experts in the field of study were also used in identifying a content area and also advised where applicable. This facilitated a number of necessary revisions and modifications of the research instrument

3.7 Data collection procedure

Information for the analysis from primary and secondary sources. Primary data will be collected from the Data warehouse of subscriber data at the telecommunication company. Secondary information will be collected from books, journals, newspaper articles and published articles. Both quantitative and qualitative data will be used to achieve the objectives of this study

3.8 Data Processing and analysis

Python will be used for analyzing the data, data mining and machine learning. Collected data will be presented in graphs and charts. Statistical datamining and machine learning tools, such as, percent, correlation, classification and regression will be used to interpret findings of the study. Emphasis will also be given on quantitative analysis with qualitative data.

3.9 Experiments

In this study we will compare the classical churn modelling approach evaluating its performance base on accuracy precision and recall. We also evaluate the performance of area under the curve using Receiver operator characteristic curves. We will then extend the experiment using uplift modelling. We will undertake two common approaches for uplift modeling: two models' approach and class transformation approach [36]. The objective is to be able to predict uplift, that is the difference of probability in outcome that is generated by the treatment for each subscriber. For the class model approach this we will undertake the below

- stack the target and control data
- flip the target for the control dataset
- train a model on this target

how we calculate uplift will be 2 times predicted probabilities minus 1. For the two-model approach we model uplift by the difference between probability of outcome in target dataset minus the probability of outcome in control dataset. We will then evaluate the performance using uplift bins, uplift curves and Qini curves [36]. Our evaluation will be based on time taken to compute, accuracy precision and recall for the ensemble models

4. Evaluation metrics Definitions

The performance of our churn prediction models was evaluated based on test data. The model predicted two classes "positive" or "churn/yes" and "negative" or "not churn/no". Some of the widely used class output evaluation measures [41] for binary predictive models are:

1) **Accuracy:** Accuracy is the percentage of correctly predicted instances to the total instances. Accuracy is a useful metric only when all the classes are equally represented.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Where, True Positive (TP): total sample population that are positive and predicted correctly as positive.

True Negative (TN): total sample population that are negative and predicted correctly as negative.

False Positive (FP): total sample population that are negative and predicted wrongly as Positive.

False Negative (FN): total sample population that are positive and predicted wrongly as negative.

2) **Precision:** Precision measures how many of sample population classified as positive are positive. Also called Positive predictive value. The ratio of correct positive predictions to the total predicted positives

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

3) **Recall:** Recall measures how many of the total positive sample population were classified correctly as positive. Also called Sensitivity, Probability of Detection, True Positive Rate The ratio of correct positive predictions to the total positive's population samples.

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

4) **ROC curve (receiver operating characteristic curve):** represents the performance of a classification algorithm at all thresholds. (by exercising thresholds: when computing TPR and FPR for the threshold at 0.7, you apply the model to each example, get the score, and, if the score if higher than or equal to 0.7, you predict the positive class; else, you predict the negative class). Lowering the threshold classifies more elements as positive, increasing both False Positives and True Positives. It plots 2 parameters:

1.True positive rate (Recall) discussed above

2. False Positive rate: Tells what % of people who were classified negatively but were positives.

$$FPR = \frac{FP}{FP+FN} \quad (10)$$

5) **AUC (Area under the ROC Curve):** It gives an aggregate measure of performance across all possible classification thresholds. The higher the area under the ROC curve (AUC), the better the classifier. A perfect classifier would have an AUC nearing1 while FPR nears 0.

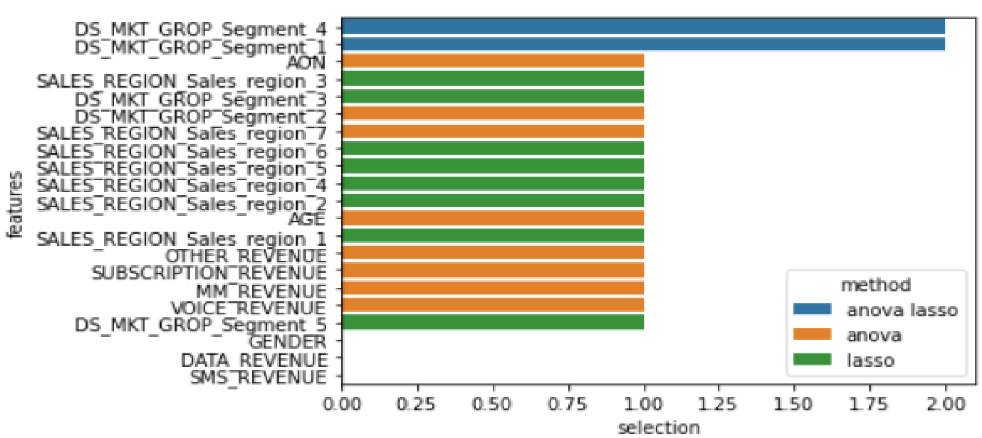
4.5. Discussions on Experiments don during the research

When we were doing the research study, we did various experiments for Objective 1, 2 and 3

4.5.1.1 Features selection Objective 1 experiments 1

We also looked at lasso regularization which is a regression analysis method that performs both variable selection and regularization in order to enhance accuracy and to make it more understandable and well interpreted. The results are shown in fig 3.3 that shows feature selection metrics. the blue feature are the ones selected by both ANOVA and lasso, the other are selected by just one of the two methods.

FIGURE 3.3 FEATURE SELECTION METRICS FOR LASSO AND ANOVA

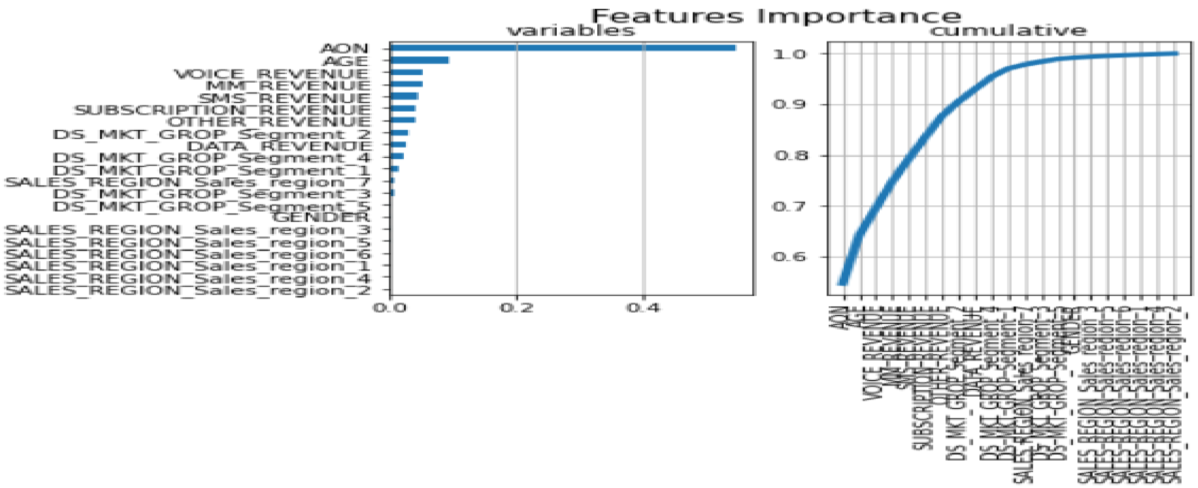


4.5.1.2 Features selection Objective 1 experiments 2

We used the feature importance metric for the ensemble extra trees. Extra trees are an ensemble method that consists a high number of decision trees in which every node is a condition on a single variable that split the dataset in two so that like values and up in the same subset. Feature importance is calculated using the entropy reduction in each tree. The results are shown in fig 3.4 that shows feature importance for extra trees. The important features here include AON AGE

Voice revenue, MM revenue, subscription revenue, other revenue market segment 4 market segment 1 sales regions seven, customer segment 3, customer segment 5 and gender. The rest that follows can be discarded. This is informed by the elbow curve where it flattens on the curve.

FIGURE 3.4 FEATURE SELECTION METRICS FOR ENSEMBLE TREES.



Finally, for experimentation we use all the variables first and compare on the impact when we discard the variables as advised by the feature selection methods above. From the results we selected the one-way ANOVA and CRAMER V chis square analysis for this study which will be discussed under results section

4.5.2 Model Building Objective 2 experiments 1

In this section we created a ensemble stacking model using the MLENS library. In this section we will describe the methodology and results of the experiment.

4.5.2.1 Ensemble Modelling m-lens classifier experiment 2.

The ensemble learning using the m-lens library followed a three-step process.

1. Initializing the ensemble, to become the Super-Learner.
2. Adding the intermediate estimators. Here we added two classifiers: extratrees ensemble and. Adaboost classifier. The execution was done in parallel.
3. Finally added the meta-learner, which is GaussianNB.

The fit data is displayed on figure 3.5. The first column, score-m, contains the score. The suffix -m denotes mean values, and -s denotes standard deviation across folds for brevity. ft and pt. stand for fit time and prediction time, respectively.

FIGURE 3.5 M-LENS META-LEARNER MODEL FIT DATA METRICS

```
Fit data:
                                   score-m  score-s    ft-m  ft-s   pt-m  pt-s
layer-1  adaboostclassifier           0.90     0.00   84.17  2.21   7.86  0.44
layer-1  extratreesclassifier         0.94     0.00  198.61  0.54  34.12  0.17
layer-2  adaboostclassifier           0.94     0.00    9.95  0.01   4.54  0.01
layer-2  gaussiannb                   0.94     0.00    0.23  0.01   0.12  0.01
```

From figure 4.5.2 M-lens learner Model fit metrics we can deduce that the accuracy of the model was at 94% with an execution time of around sixty seconds. When we compare with the Extra-trees ensemble method it came in second. The extra-trees ensemble model will be discussed in the results section

4.5.2.3 The Two-Model Approach Experiment 2

In this method two classifiers were built using extra trees ensemble classifier. One was trained on observations that received treatment (called *model_T1*) and the second is trained on observations that didn't receive treatment (called *model_T0*). The uplift for those observations were calculated. If the observation experienced treatment, it was an input to the *model_T1* and the probability that the subscriber will accept the retention offer that was predicted. Thereafter we investigated what could happen if the subscriber didn't receive treatment. The treatment indicator in the observation's feature was changed to 'zero' which served as an input to the *model_T0* that predicted the probability that Subscriber will accept a retention offer and not churn. The uplift was calculated as the difference between the outputs of the *model_T1* and *model_T0*. The higher the difference, the more profitable it was to give a retention campaign to a Subscriber. Uplift was calculated for the subscribers that were not given treatment. The results comparison between the propensity model and the two-model approach will be discussed in the results section.

4.5.2.3 The Novel Propensity Outcome Modification Approach

In this method a classifier was built using extra trees ensemble classifier. Here we stack target and control data. This is achieved by first multiplying the treatment and control that represent outcome one. Thereafter we divide the outcome one by the product of one minus the treatment data representing outcome two and one minus control data representing outcome three. The result being the modified outcome. We then train extra tree ensemble classifier on this modified outcome.

This is shown using the equation below

$$xtrain['modified\_outcome'] = xtrain['conversion'] * xtrain['Tratment'] \,/\, (1 - xtrain['conversion']) * (1 - xtrain['treatment']) \quad [11]$$

4.5.3 Models Evaluation Objective 3 experiments 1

For the results of Experiment 1 for Objective 2 evaluated and analyzed the results as shown in this section

4.5.3.1 Evaluation metrics m-lens ensemble Experiment 2 results

FIGURE 3.6 M-LENS ENSEMBLE CLASSIFIER CONFUSION MATRIX

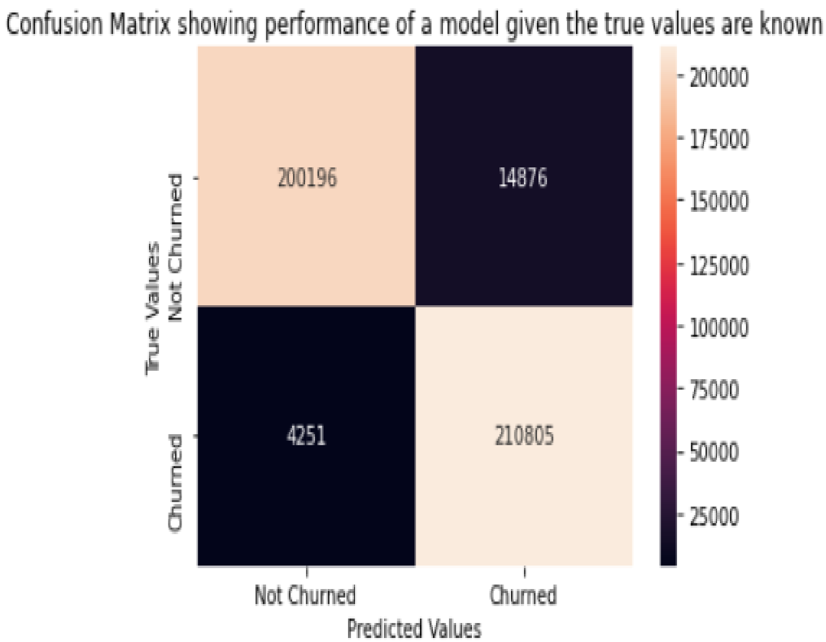
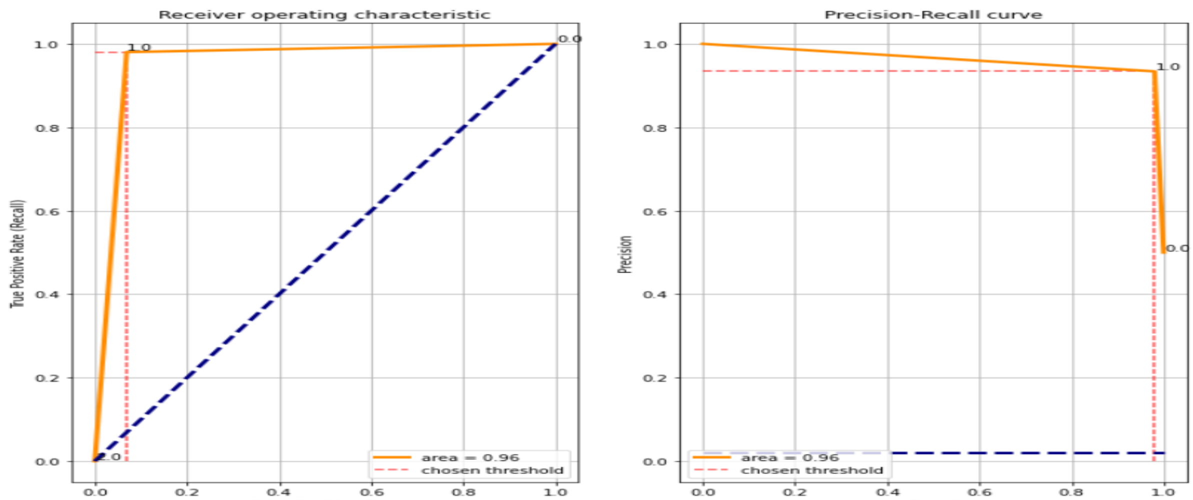


Figure 3.6 shows the M-lens Ensemble Model Confusion Matrix. After some tuning, we can see the model predicted 215,072 (14,876+200,196) of which 200,196 are true positives and 14,876 are false positives so it has a **Precision** is 200,196/215,072 93.1% the model predicted 215,056 4251 being false negatives (4251+210805) of which gives **Recall** of 98.%. We have chosen a threshold of 1 to decide whether a prediction is churn or not churn leading to the results in figure 4.6.2 that represent the receiver operating characteristics curve and the precision recall curve for the m-lens ensemble meta learner model. Every point of these curves represents a confusion matrix obtain with different thresholds (the numbers printed on the curves in figure 3.7 Receiver Operating Characteristics curve. For a threshold of 0.5 and get a recall of 98% meaning the model would predict 98% of churn subscribers correctly but the precision would drop to 93.1% meaning the model will predict a little of false positives

FIGURE 3.7 RECIEVER OPERATING CHARACTERISTICS AND PRECISION RECALL CURVE



The lime package aided to build an explainer. Lime is algorithm agnostics meaning it can be applied to any machine learning model. The method tries to explain how input variables predict change. It provides local algorithm explanations by modifying a single random data sample by tweaking the feature values and showing the resulting outcomes of a model specifically how the prediction was made, and which variables caused the prediction. An explanation is done by analyzing the model locally in comparison with an interpretable one e.g., decision tree. The interpretable models are train on a data subset that will give good local predictions. The dataset is created by adding noise to continuous features. In our current study models are used to approximate local behavior by looking at a small subset of the data. To give an illustration I took a random observation from the test set just to see what the model predicted and the results are shown in figure 3.8

print('True:', y_test[134738], "-- > pred:", predicted[134738],"|prob:", np.max(predicted_prob [134738]))

output

```
True: 0 -- > pred: 1 |prob: 0.63
```

FIGURE 3.8 EXPLANATION FOR CHURN VARIABLE USING LIME PACKAGE

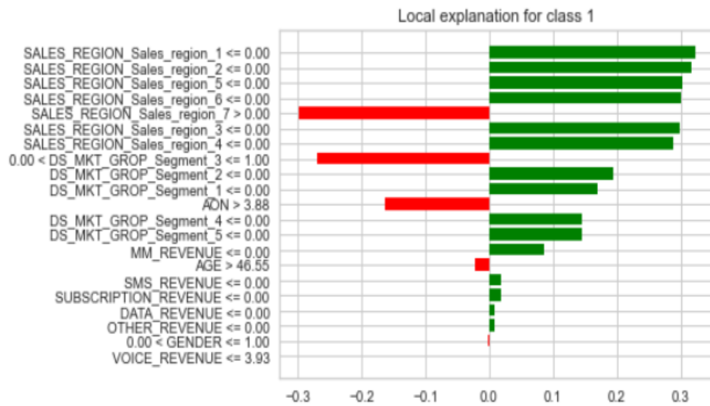
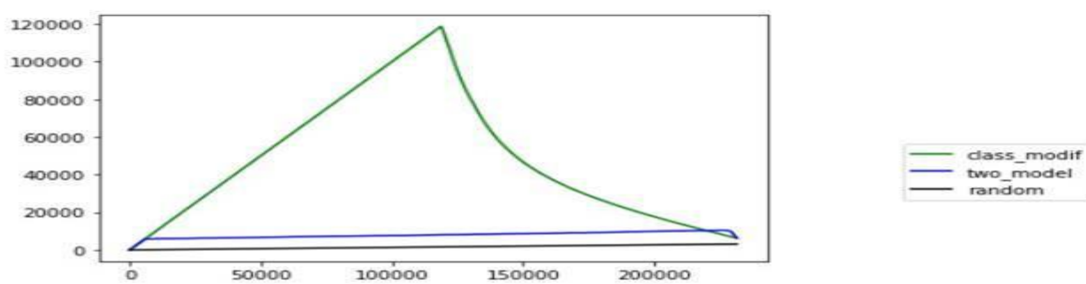


Figure 3.8 shows how Lime Package in python can be used to explain factors affecting the churn variable on a random subscriber in the population. The model thinks that this observation is a churn case with a probability of 0.63 and in fact this customer did churn. The main factors for this particular prediction are sales regions, customer segments and revenues streams except sales region 7, customer segment 3 , ag on the network and customer's age.

4.5.3.2 Evaluation metrics Uplift modeling Experiment 2 results

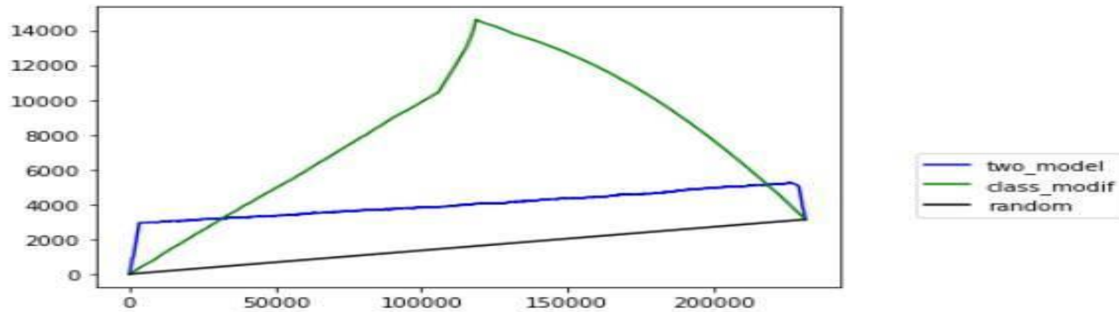
For uplift modeling experiments the researcher evaluated the two-model approach and the novel propensity class modification approach. The evaluation metrics that was studies was the Area under the curve and the uplift curve.

FIGURE 3.9 AREA UNDER THE UPLIFT CURVE FOR THE BEST ENVISAGED MODEL, TWO MODEL APPROACH & PROPENSITY CLASS MODIFCATION APPROACH



In the Area Under the Uplift Curve, we have the cumulative version of the bar plot. The black line represents the 'random' model, the blue line is the cumulative lift for the two-model approach, the green line is the cumulative lift for the propensity class modification model. When we map the outcomes of two model approach and the novel propensity class modification approach the area under the uplift curves (AUUC) we get figure 3.9. Her we find that the best performing model for this data set is the propensity class modification model followed by the two-model approach model. The better the uplift model the higher the area under the uplift curve.

FIGURE 3.10 THE QINI CURVE FOR THE BEST ENVISAGED MODEL, TWO MODEL APPROACH & PROPENSITY CLASS MODIFCATION APPROACH



In the Qini Curve, we represent Total Subscriber Lifetime Value. The black line represents the 'random' model, the blue line is the cumulative lift for the two-model approach, the green line is the cumulative lift for the propensity class modification model. When we map the outcomes of two model approach and the novel propensity class modification approach the area under the uplift curves (AUUC) we get figure 3.9. Her we find that the best performing model for this data set is the propensity class modification model followed by the two-model approach model. The better the uplift model the higher the maximum of the Qini curve

TABLE 3.1 UPLIFT COMPARISON PER SEGMENT BETWEEN EXPERIMENT 1 AND 2

Customer segment	two model approach uplift		Propensity Outcome modification approach uplift	
	No retention offer	Retention offer	No Retention offer	Retention offer
segment 1	0.000034	0.000495	-0.000167	0.002535
segment 2	0.000102	-0.000306	0.00025	-0.0022301
segment 3	0.000204	-0.000033	0.00052	-0.000348
segment 4	0.000045	0.000262	-0.00015	0.001759
segment 5	0.000003	0.000367	-0.000159	0.000822

For segment 1 the classification approach proved more effective with greater uplift for the subscribers that got the retention offer at 0.002535 compared to the two-model approach at 0.000495. The uplift for the subscribers who did not get the retention offer was insignificant at 0.000034 for the two-model approach and negative at -0.000167 for the propensity outcome classification approach. For segment 2 both the models experienced negative uplift. So, this segment should not be targeted for any retention campaigns. For the subscribers who received the retention offer for this segment was recorded at -0.000306 for the two-model approach and - 0022301. Meaning that these churned customers were lost causes even if targeted will be irredeemable. For the subscribers who did not receive the retention offer uplift was at 0.000102 for the two-model approach and 0.000250 for the classification approach. For segment 3 the two-model approach proved more effective then then the classification approach. The uplift experienced was very low. So, this segment should not be targeted. For the subscribers who received the retention offer the uplift was at -0.000033 for the two-model approach and at -0.000348 for the classification approach meaning the subscribers were irredeemable if targeted. For the subscribers who did not receive the retention offer, the uplift was 0.000204 for the two-model approach and 0.000520 for the classification approach. For segment 4 the best approach was the classification approach. Where better uplift was experienced. For the subscribers who received the retention offer the uplift was at 0.000262 using the two-model approach and at 0.001759 when using the classification approach. This was the best segment to be targeted showing the highest gain. For the subscribers who did not receive the offer, the uplift was at 0.000045 for the two-model approach and -0.00015 for the classification approach. For segment 5 the classification approach gave a better uplift than the two-model approach. The uplift for the subscribers who receive the retention offer was at 0.000367 using the two-model approach and at 0.000822 using the classification approach while for the subscribers who did not receive the retention offer was at 0.000003 for the two-model approach and -0.000159 for the classification approach

CHAPTER FOUR

DATA ANALYSIS, RESULTS AND DISCUSSIONS

4.0 INTRODUCTION

This section of the dissertation is devoted to application of models and approaches described in chapter 3 on a real data set of approx. 760,000 anonymized subscribers from a Kenyan telecommunication company. The dataset was in two sets. One contained active customers and the other churned customers. The variables, which could help to reveal leaving customers, were thoroughly selected based on advice of domain experts that contained two major categories of data – demographic data, such as age, gender, sales region, customer segment, age on the network, customer ARPU on various products and customer lifetime value The input variables divided into demographic group and ARPU variables are described and descriptive statistics are showed at the beginning of this chapter. The variables are analyzed using histograms, boxplots, density plot, stacked bar graphs and contingency tables to uncover interesting patterns. The model has two parts. Prediction modelling and uplift modelling. Two approaches to predictive modelling are compared. The first one is based on the ensemble learning using extra trees and stacking ensemble using the m-lens library. This step is used to predict the probable churners in a subscriber population. The second step we perform uplift modelling where 3 approaches are compared mainly propensity modelling, direct modelling and meta learners/ indirect modelling. The estimated models are further described and explained so as to prescribe which subscribers should be targeted with an offer so as to retain the customers or increase their usage of products on the network. The testing data set is then used to calculate performance statistics to check the behavior of these models on unseen data. Evaluation metrics are used to describe the performance for both ensembles learning and uplift modelling approaches.

4.1 Description of the data

Data for the thesis were obtained from a Kenyan telecommunications service provider. Raw input data for this task are stored in various production systems. The process of preparing data into the form suitable for modelling was done through ETL. Extract data from source systems, transform it and load it into database. Data engineers provided the data in CSV format. Sample of the churn datasets are shown below in figure 4.1 .

FIGURE 4.1 SAMPLE DATA FROM THE CHURN DATA SET

	Unnamed: 0	UNIQUE_IDENTIFIER	ID_MNTH	AON	GENDER	DOB	AGE	STATUS	CALLS	INTERACTION_TYPE	...	SMS_REVENUE	DATA_REVENUE	DATA_REVENUE	MM_REVENUE	SUBSCRIPTION_REVENUE	OTHER_REVENUE	SALES_REGION	DS_MKT_GROP	Churn
0	0	15264047	NaN	521.0	F	31-Dec-75	16387.0	Expired	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_5	YES
1	1	43992874	NaN	1097.0	F	1-Jan-71	18212.0	Active	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_3	YES
2	2	6408224	NaN	521.0	F	8-Sep-78	15405.0	Active	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_3	YES
3	3	28605538	NaN	1813.0	M	12-Aug-91	10684.0	Expired	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_1	YES
4	4	19412253	NaN	259.0	M	20-Aug-78	15424.0	Expired	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_3	YES

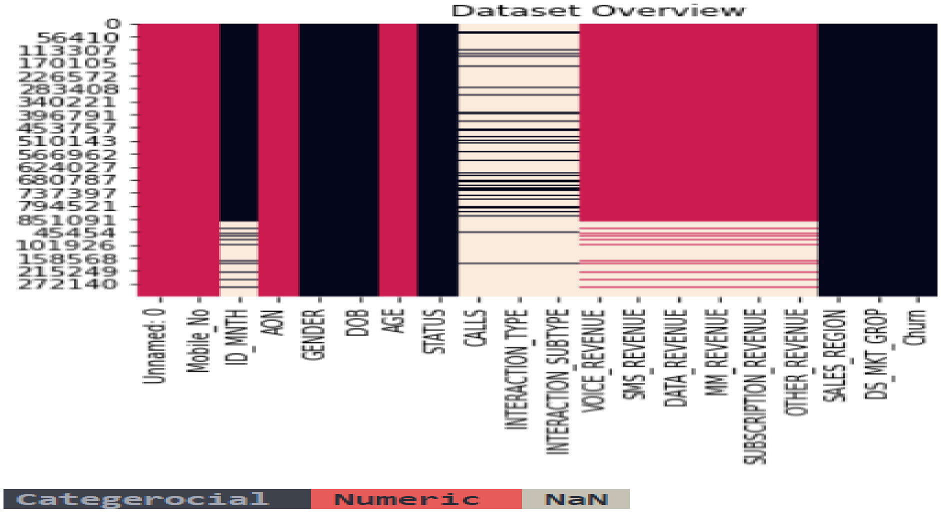
You can see descriptive statistics (minimum, the first quartile, median, mean, the third quartile and maximum) of numeric variables in the figure 4.2

FIGURE 4.2 DESCRIPTIVE STATISTICS

	ID_MNTH	AON	GENDER	DOB	AGE	STATUS	CALLS	DATA_REVENUE	MM_REVENUE	SUBSCRIPTION_REVENUE	OTHER_REVENUE	SALES_REGION	DS_MKT_GROP	Churn
count	925679.000000	1.185368e+06	1192039	1183323	1.183323e+06	1185368	193647.000000	925679.000000	925679.000000	925679.000000	925679.000000	1192039	1192039	1192039
unique	NaN	NaN	2	20263	NaN	6	NaN	NaN	NaN	NaN	NaN	7	5	2
top	NaN	NaN	M	1-Jan-70	NaN	Active	NaN	NaN	NaN	NaN	NaN	Sales_region_7	Segment_3	NO
freq	NaN	NaN	687902	40271	NaN	930868	NaN	NaN	NaN	NaN	NaN	343393	724306	862833
mean	202009.015913	1.608811e+03	NaN	NaN	1.446122e+04	NaN	1.262493	60.985967	194.604470	189.287237	46.406612	NaN	NaN	NaN
std	0.817313	1.883946e+03	NaN	NaN	5.775289e+03	NaN	7.464713	589.024500	686.166227	1629.098264	241.485657	NaN	NaN	NaN
min	202008.000000	4.000000e+00	NaN	NaN	1.000000e+01	NaN	1.000000	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN
25%	202008.000000	2.360000e+02	NaN	NaN	9.955000e+03	NaN	1.000000	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN
50%	202009.000000	8.340000e+02	NaN	NaN	1.317100e+04	NaN	1.000000	0.000000	0.000000	15.000000	4.000000	NaN	NaN	NaN
75%	202010.000000	2.508000e+03	NaN	NaN	1.821200e+04	NaN	1.000000	5.000000	123.540000	93.500000	27.000000	NaN	NaN	NaN
max	202010.000000	4.414400e+04	NaN	NaN	7.125300e+05	NaN	1151.000000	203598.560000	32057.400000	458890.306400	41174.000000	NaN	NaN	NaN

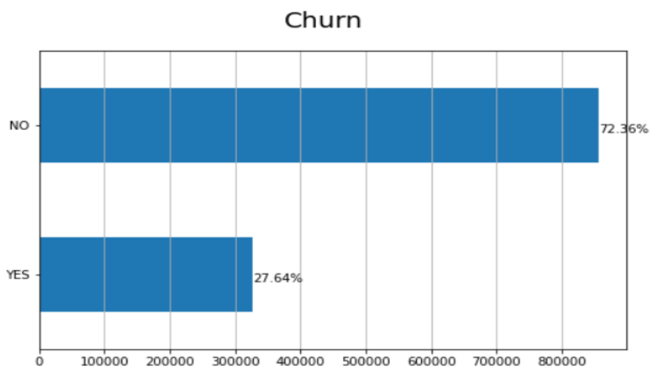
there were clearly some outliers on the AON (age on the network) where some were more than years that the telco was in existence. This was corrected by creating a maximum age on the network for all the subscribers that had gone beyond the threshold. The distribution of categorical and numeric variables was shown as below.

FIGURE 4.3 CATEGORICAL AND NUMERICAL DATA DISTRIBUTION



The was a lot of missing data that required feature imputation so as make the data ready for machine learning. The feature that had imputation performed include gender, sales regions and customer segments represented by DS_MKT_GROP variable. Python package simple Imputer was used to impute these missing values. This package uses multiple imputation and sophisticated algorithms which combines expectation minimization algorithm and bootstrapping

FIGURE 4.4 CHURN VARIBLE DISTRIBUTION



The dataset was imbalanced as 72% were active customers and only 27% were churn customers. Data also included customers with multiple numbers but the unique customers total 760,000. The issue of multiple sim by subscribers was resolved using the group-by python package that aggregated data for the individual subscribers

For subscribers in the telecommunications sector, it can be interesting to see, how the typical subscriber that churned or did not churn looks like. This information is obtained when only subscribers with churn variable equal to yes are selected or subscribers with churn variable equal to No are selected and mean and median for numeric variables and mode for categorical and binary variables are computed. We can see these statistics in the Table. 4.1 and 4.2

TABLE 4.1 PROFILE OF TYPICAL CHURNERS AND NON-CHURNERS – NUMERIC VARIABLES

	Churners		**Non-churners**	
Numeric variable	**Mean**	**Median**	**Mean**	**Median**
AON	1.9	0.7	5.3	3.5
AGE	40.5	34.9	39.3	36.5
CLV	18113	5888.7	4653.2	1030
VOICE_REVENUE	138.6	35.9	338.6	86.8
SMS_REVENUE	31	1.1	35.7	3.3
DATA_REVENUE	34.9	0	63.4	0
MM_REVENUE	4.4	0	210.1	0
SUBSCRIPTION_REVENUE	60.8	0	150.4	17.5
OTHER_REVENUE	18.1	1	48.9	4

The Mean and Median were calculated for Numerical variables. From the Table 4.1. A typical churner the CLV mean and median is higher compared to a non-churner with 18113 and 5888.7 respectively. This most likely they spend a lot of money on product and services for short periods of time. A typical non- churner tends to stay on the network a longer period of time with its mean and mode at 5.3 years and median at 3.5 years much higher than the churned customers who have a mean of 1,9 years and a median of 0.7 years. Typical non-churner revenue metrics tend to be much higher over time compared to churned customers across the board as shown in figure 4.2.2

TABLE 4.2 PROFILE OF TYPICAL CHURNERS AND NON-CHURNERS – CATEGORICAL VARIABLES

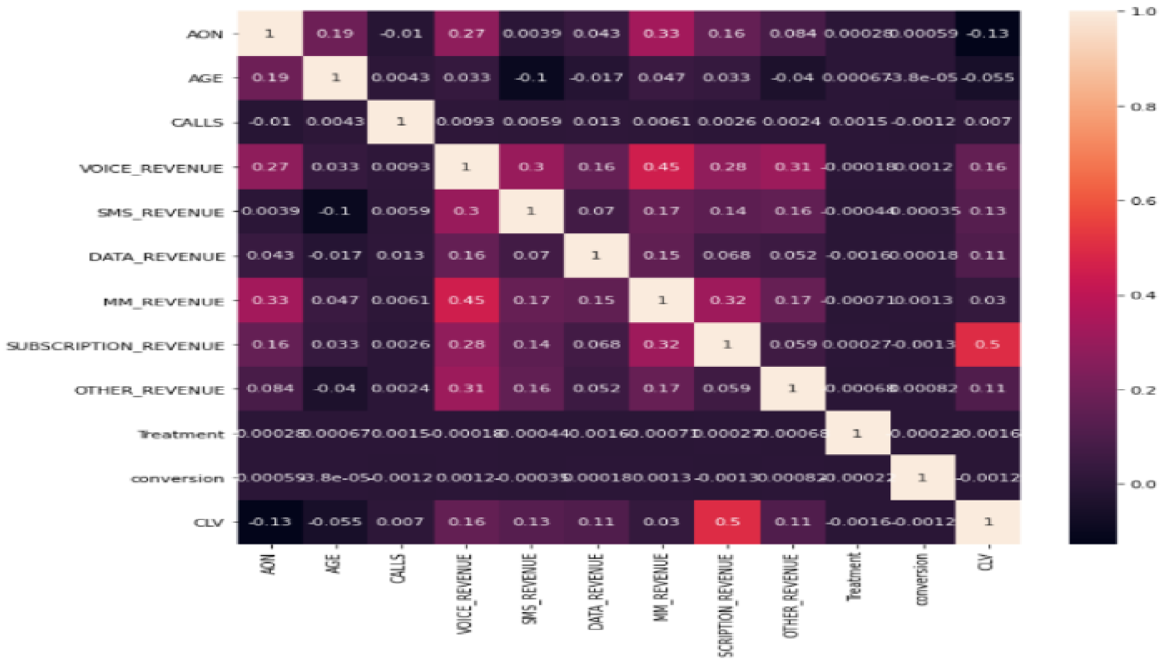
	churners	**non-churners**
GENDER	M	M
SALES REGION	Sales region 6	Sales region 7
Segment	segment 3	segment 3

Table 4.2 shows a profile of a typical churner and non-churner using Categorical variables. The most frequent values – modes – were calculated for categorical variables. From the Table 4.2 A typical churner and non-churners are male, both of which originate from customer segment 3, where the typical churner comers from sales region 6 while typical non- churner comes from sales region 7.

4.2 Exploratory data analysis

Exploratory data analysis is done before modeling to give more data on the development of predictive algorithms. We did a heatmap correlation analysis.

FIGURE 4.5 CORRELATION HEAT MAP. (Method =PEARSON)



A nice visualization of correlation coefficients among variables is shown on fig 4.5 None of the Variables are highly correlated. The size of correlation coefficients is represented by shade of the color. The lighter the color the strongly the variables are correlated. The plot expresses the values of correlation coefficients also in absolute values.

4.3.0 Features selection Objective 1 To identify the attributes that determine subscriber churn in the Kenya telecommunication sector.

Feature selection is a process of obtaining relevant subset of features to build a machine learning models and is significant when trying to improve the performance of predictive algorithms when it comes to prediction values such as high TP and low FP for churn analysis. Obtaining good features is hard. Most of the feature sets that have been proposed so far in telecommunication sector can still be made better. In order to improve the prediction rates for churn recognition, based on the dataset obtained for this study, we did several feature selections processes. This was done to answer objective 1 for this study. The first step was to perform correlation analysis shown in figure 4.5 None of the Variables are highly correlated. One-way ANOVA test was used thereafter to test whether the means of two independent samples were significantly different in the churn dataset. This was ascertained by checking the p value was small enough (<0.05) to reject the null hypothesis of sample means equality. The results are shown in table 4.3 for numerical variables. The table also shows the degree of freedom, sun of square f score and mean sum of squares for every variable. From the results we can see all the numerical variables in the dataset had a p value less than 0.05 hence were dependent except the calls to the call center variable which was independent and hence was not included for model building.

TABLE 4.3 ONE-WAY ANOVA ANALYSIS– NUMERICAL ANALYSIS

	One Way ANOVA						
	x variable	**y-variable**	**p value significance**	**df**	**sum_sq**	**mean_sq**	**f**
1	Age	Churn	< 0.05	1	357190.8	357190.76	1428.325
		residual	N/A	1183074	295859100	250.076593	N/A
2	AON	Churn	< 0.05	1	2777610	2.78E+06	132815.4
		residual	N/A	1183074	24741990	20.91331	N/A
3	voice revenue	Churn	< 0.05	1	2331642000	2331642000	4004.029
		residual	N/A	918583	5.34913E+11	582324	N/A
4	MM revenue	Churn	< 0.05	1	2467618000	2467618000	5234.003
		residual	N/A	918583	4.33074E+11	471458.9	N/A
5	DATA revenue	Churn	< 0.05	1	47314000	47314000	135.3857
		residual	N/A	918583	3.21022E+11	349475.7	N/A
6	SMS revenue	Churn	< 0.05	1	1312062	1312062	196.0996
		residual	N/A	918583	6146051000	6690.795	N/A
7	Subscription revenue	Churn	< 0.05	1	467629700	467629700	1529.236
		residual	N/A	918583	2.80896E+11	305793	N/A

8	Other revenue	Churn	< 0.05	1	54945230	54945230	936.2063
		residual	N/A	918583	5391093000 0	58689.23	N/A
9	calls	Churn	0.369	1	0.284655	0.284655	0.807688
		residual	N/A	189419	66757.27273	0.352432	N/A

To test correlation between 2 categorical variables we will a chi square test assuming that the two variables are independent (null hypothesis) it tests whether their values are uniformly distributed. If the p value is small enough (< 0.05), the null hypothesis can be rejected making the two variables dependent. Cramer V which is a measure of correlation that follows the chi square approach that is symmetrical (like Pearson's) and ranges between zero and one and where there are no negative values. Table 4.4 shows the results of Cramer V on the churn dataset under study for the categorical variables. From the results we can see all the categorical variables in the dataset had a p value less than 0.05 hence were dependent

TABLE 4.4 CRAMER V ANALYSIS– CATEGORICAL VARIABLES

	Cramer V		
	categorical	**y-variable**	**p value**
1	Gender	Churn	< 0.05
2	LINE status	Churn	< 0.05
3	identified Month	Churn	< 0.05
4	Sales Region	Churn	< 0.05

4.4.0 Predictive modelling Objective 2 to develop a predictive and uplift model for subscriber churn.

This part of the dissertation is focused on detailed explanation of the ensemble classification models used. We will discuss the ensemble extra trees classification and stacking ensemble using the m-lens library this to answer objective tow which is to develop a predictive and uplift model for subscriber churn. We will also discuss the feature importance element to answer objective one which to identify the attributes that determine subscriber churn in the Kenya telecommunication sector and finally evaluate the models to answer objective three.

4.4.1 Ensemble Modelling extra trees classifier experiment 1.

To select the ensemble classifier, we used the data set on several models based on the metrics of accuracy, precision recall, AUC, F1 score and run time of the model as shown in figure 4.6. Ensemble extra trees was selected as it had the highest recall score of 98.11%. Also, the accuracy was higher than the ones we came across under literature review of 94%. Finally, its run time was much faster than the random forest.

FIGURE 4.6 MODEL SELECTION METRICS

	Name	Accuracy	Precision	Recall	AUC	F1 Score	Run Time
8	RandomForestClassifier	0.956776	0.947258	0.967572	0.956757	0.957307	311.036947
7	ExtraTreesClassifier	0.956039	0.934364	0.981150	0.955996	0.957186	254.813144
0	DecisionTreeClassifier	0.949659	0.948112	0.951567	0.949656	0.949837	13.188610
6	GradientBoostingClassifier	0.926024	0.902849	0.955072	0.925975	0.928226	269.834659
3	MLPClassifier	0.913126	0.877373	0.960841	0.913044	0.917212	786.414708
5	AdaBoostClassifier	0.903036	0.898699	0.908848	0.903026	0.903745	83.024237
9	LogisticRegression	0.839813	0.775871	0.956474	0.839613	0.856758	7.718093
1	LinearSVC	0.824296	0.755162	0.960656	0.824063	0.845604	291.096855
4	RidgeClassifier	0.784980	0.706366	0.976712	0.784652	0.819826	0.550460
2	GaussianNB	0.729927	0.659771	0.951381	0.729547	0.779187	0.547660

The feature importance of the model showed that age and AON variables had the greatest contribution to the model as shown on figure 4.7. Table 4.5 show the actual contribution of the top 9 variables to the model. Age contribution was at 21.3 % and AON contribution was at 13.6 %

FIGURE 4.7 EXTRA TREES CLASSIFIER MODEL

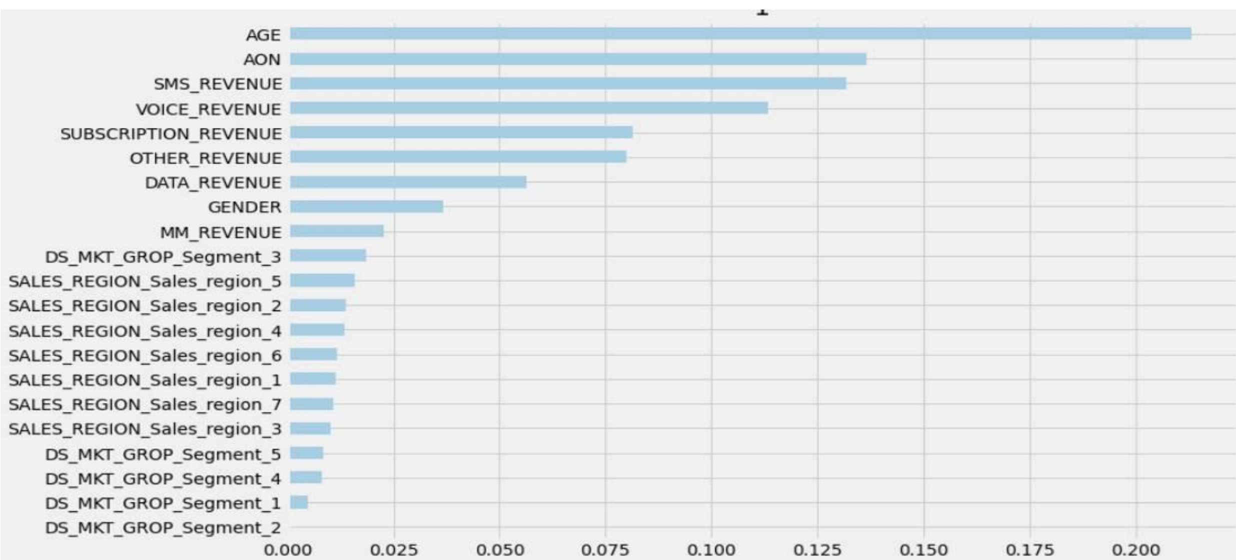


TABLE 4.5 IMPORTANT VARIABLES PICKED BY THE EXTRA TREES ENSEMBLE MODEL FOR PREDICTING CHURN

Feature importance	
AGE	0.213207
AON	0.136573
SMS_REVENUE	0.131648
VOICE_REVENUE	0.113266
SUBSCRIPTION_REVENUE	0.081572
OTHER_REVENUE	0.079852
DATA_REVENUE	0.056417
GENDER	0.036643
MM_REVENUE	0.022634

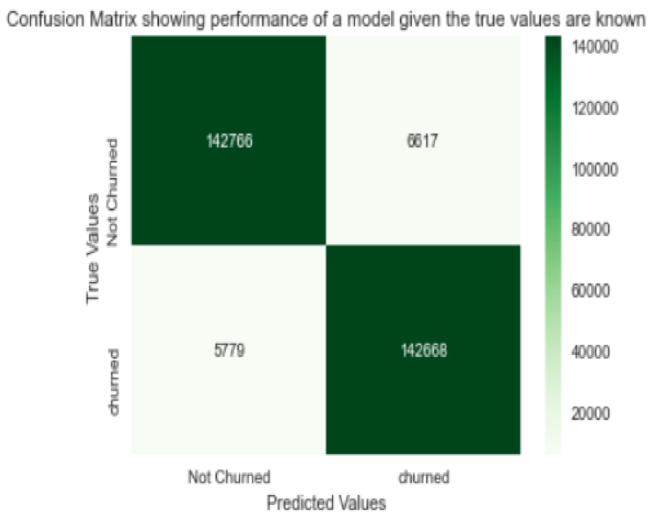
4.5.0 Evaluation metrics Objective 3 to evaluate the proposed models.

4.5.1 Evaluation of extra-tees ensemble method experiment 1

Fig 4.8 shows the model confusion matrix. After some tuning, we can see the model predicted 149,383 (6,617+142,766) of which 142,766 are true positives and 6,617 are false positives so it has a **precision** is 142,766/149,383 **95.6%** the model predicted 142,668 true negatives 5,779 being false negatives (6,054 +142,668) of which gives **recall** of 142,668/148722= **95.9%** .

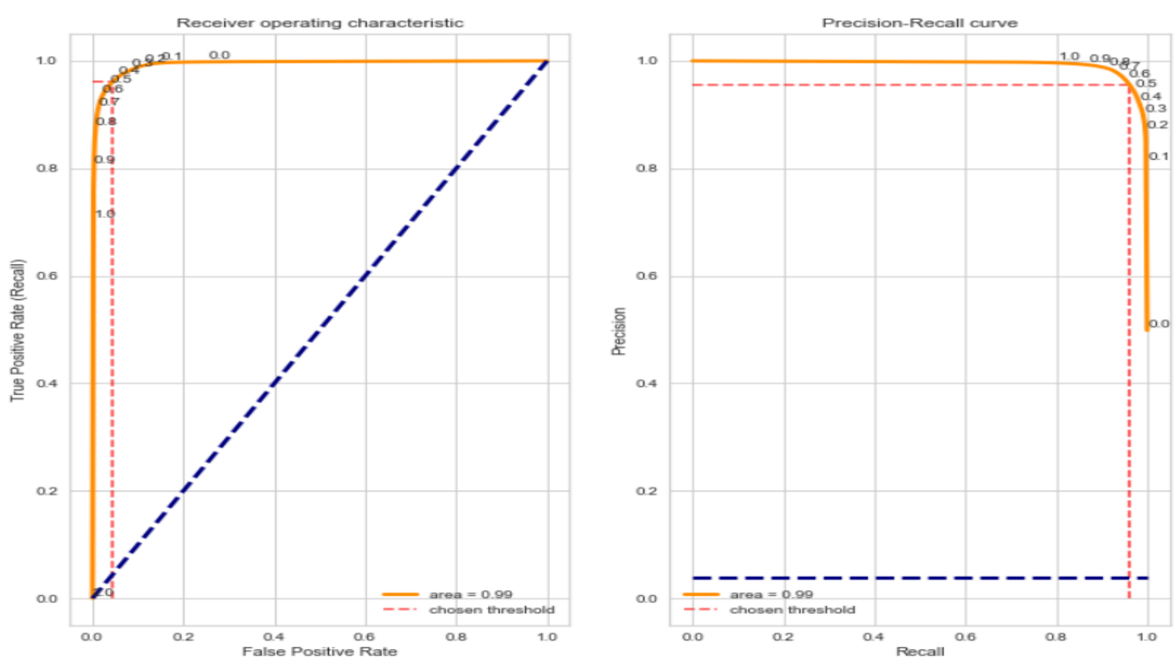
We have chosen a threshold of 0.5 to decide whether a prediction is churn or not churn leading to the results in figure 4.9 that represent the Receiver Operating Characteristics Curve and the precision recall curve.

FIGURE 4.8 ENSEMBLE EXTRA TREES CLASSIFIER CONFUSUION MATRIX



Every point of these curves represents a confusion matrix obtain with different thresholds (the numbers printed on the curves in figure 4.9 Receiver Operating Characteristics Curve. I could use a threshold of 0.2 and get a recall of 95.9% meaning the model would predict 95.9% of churn subscribers correctly but the precision would drop to 95.6% meaning the model will predict a little of false positives

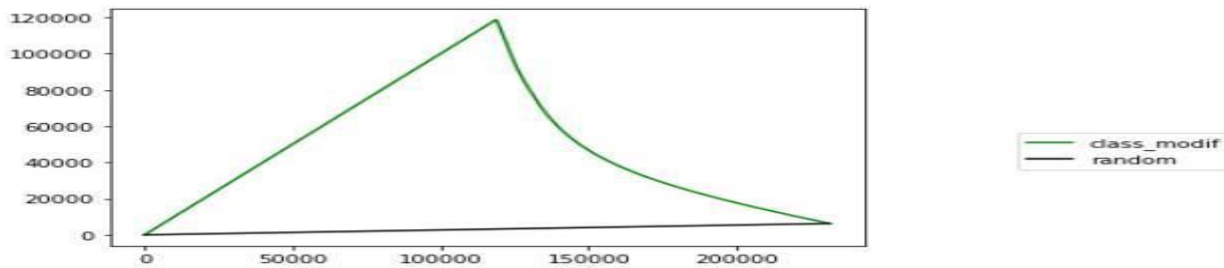
FIGURE 4.9 RECIEVER OPERATING CHARACTERISTICS AND PRECISION RECALL CURVE



4.7.1 Evaluation metrics uplift propensity

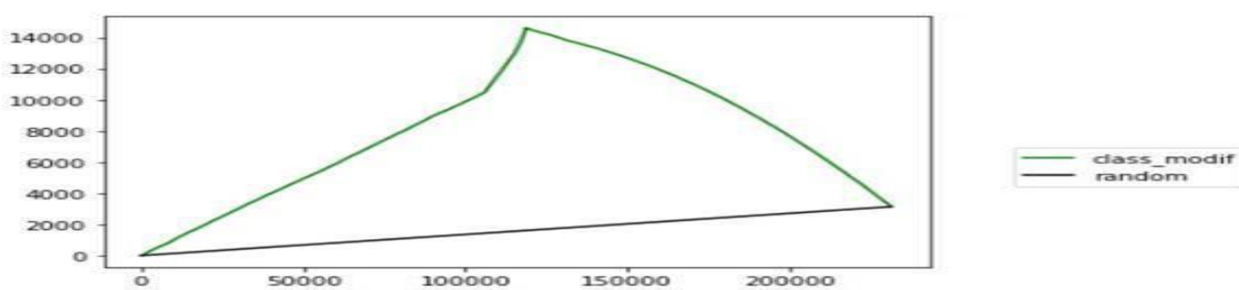
Uplift modeling is also known as incremental modeling, treatment effects modeling, true lift modeling, or net modeling. Uplift is the increase in likelihood of the outcome *with* the treatment as compared to the outcome *without* the treatment. We can't observe this difference, or causal effect, directly, but must infer it from an experiment. Uplift curve represents potential additional revenue generated by targeting a specific subscriber segment. Sorting Subscribers according to uplifts provides the prioritization list for a retention campaign based on subscriber segments. Selecting subscribers with revenue uplifts higher than retention unit cost is the strategy for maximizing ROI. In figure 4.10 the segment with the highest uplift is segment 4. represented by the curve.

FIGURE 4.10 AREA UNDER THE UPLIFT CURVE COMPARISON FOR PROPENSITY CLASS OUTCOME MODIFCATION APPROACH



Qini Curve represents effectively total subscriber lifetime value. The better the uplift model the higher the maximum of Qini curve(and area under the uplift curve AUUC). Value creating model should perform better with higher revenue than no model(or random model). In figure 4.11 the propensity class outcome modification model has performed very well with the AUUC well above the parameters that uses no model(or random model)

FIGURE 4.11 AREA UNDER THE QINI CURVE FOR PROPENSITY CLASS OUTCOME MODIFICATION APPROACH



4.7.4 Propensity Outcome modification classification approach Experiment 3

Propensity Modelling is concerned with modelling the response to our retention campaign. It is an approach that accounts for all the independent variables that affect churn in the telco industry. In our study propensity was the probability that the subscriber was treated. We divided our data in to test and training set and we used the extra-trees classification model. We trained our model using the telco dataset Then we calculated the propensity scores thereafter representing the same inform of a graph. The telco data set has 5 inherent segments that the subscribers have been categorized. To find the best segments to target with either the two-model approach or the outcome modification/classification. The overall model does not seem quite efficient when you look at it initially. But when you look at the customer base of telecommunication industry it is more than 45 million subscribers. When you multiply the population per segment with the uplift bearing in mind the ARPU for every customer per year the gains can be quite significant depending on the telco the model is applied to

TABLE 4.6 UPLIFT FOR PROPENSITY OUTCOME MODIFICATION APPROACH

Customer segment	**Propensity Outcome modification approach uplift**	
	No Retention offer	**Retention offer**
segment 1	-0.000167	0.002535
segment 2	0.00025	-0.0022301
segment 3	0.00052	-0.000348
segment 4	-0.00015	0.001759
segment 5	-0.000159	0.000822

As shown in table 4.6 for Segment 1 the propensity outcome classification approach proved more effective with greater uplift for the subscribers that got the retention offer at 0.002535. The uplift for the subscribers who did not get the retention offer was negative at -0.000167 for the propensity outcome classification approach. For Segment 2 the propensity outcome classification approach experienced negative uplift. So, this segment should not be targeted for any retention campaigns. For the subscribers who received the retention offer for this segment was recorded at -0022301. Meaning that these churned customers were lost causes even if targeted will be irredeemable. For the subscribers who did not receive the retention offer uplift was at 0.000250. For Segment 3 the uplift experienced was very low. So, this segment should not be targeted. For the subscribers who received the retention offer the uplift was at -0.000348 meaning the subscribers were irredeemable if targeted. For the subscribers who did not receive the retention offer, the uplift was 0.000520.

For Segment 4 better uplift was experienced. For the subscribers who received the retention offer the uplift was at 0.001759. This was the best segment to be targeted showing the highest gain. For the subscribers who did not receive the offer, the uplift was at -0.00015. For Segment 5 the uplift for the subscribers who receive the retention offer was 0.000822 while for the subscribers who did not receive the retention offer was -0.000159

4.8 Discussion of Results

4.8.1 Comparing the impact of ensemble models

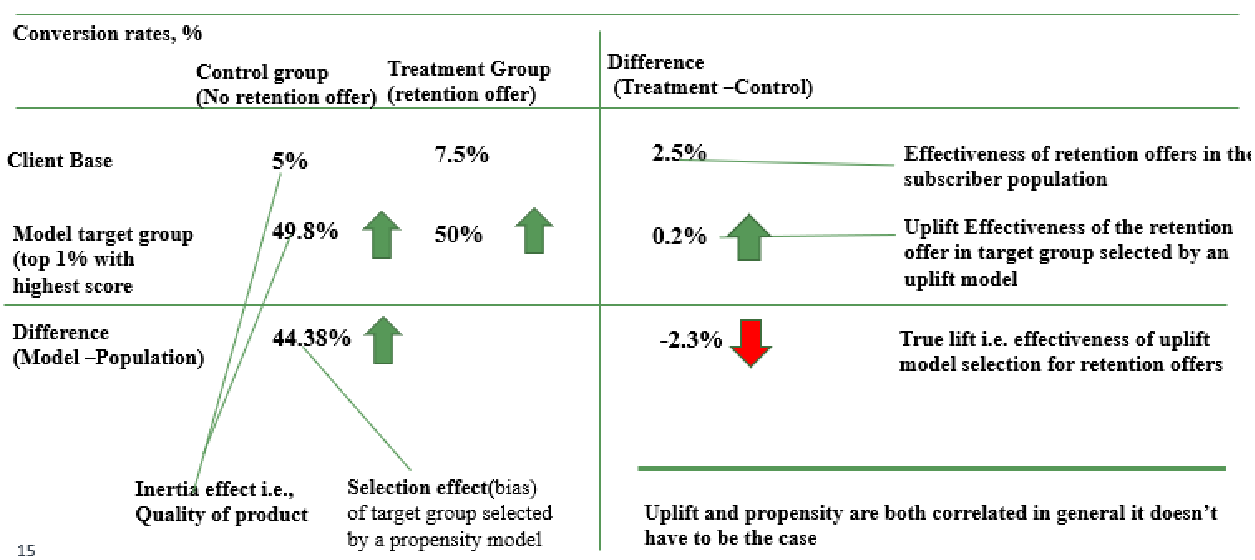
In this study, we observed that using uplift models together with ensemble models can make a lot of difference in retention strategies for a business. The campaign results were observed to be quite successful for top segments which contribute the most towards the business revenue. The combined ensemble and uplift modeling approach observed in this study can be used for other industries as well. The model selected for deployment was extra-trees ensemble combining this approach with uplift modelling techniques that gave excellent results. When we compare the studies by A. Q. Ahmed and D. Maheswari (2019), Aurélie Lemmens, Floris Devriendt, Jeroen Berrevoets and Wouter Verbeke (2020) and Sunil (Gupta 2017) ,Ning Lu, Hua Lin and Jie Lu(2012) our extra trees model gives an accuracy precision and recall of above 95% better than the accuracy achieved by the other studies whose accuracy, recall and precision were between 90-94 per cent.

Uplift analysis done was able to show the effect of treatment, rather than the outcome directly based on retention activities. The propensity class outcome modification uplift model showed better metrics to guide our retention campaign based on the retention offers rendered to our sample study population. The propensity class outcome modification (PCOM) model quantified the incremental response as a direct result of the retention incentive to our sample successfully. The PCOM model has higher precision of uplift estimation than the 2-model approach adopted by Chen aet al (2020). The PCOM model can easily be presented and interpreted. It is also less intuitive and easy to explain. When we compare to the studies done by Floris Devriendt, Jeroen Berrevoets and Wouter Verbeke (2020)on their Maximum benefit uplift model which provided uplifts based on predetermined 4 segments to generate the uplift metrics our approach gets uplift metrics based on inherent segments in the telecommunication sector which provides a seamless benefit realization on which are the most profitable segments to target with retention incentives . We can there after calculate cost savings to the business there after in future studies.

The variables used to in the extra-tree ensemble learning and propensity class outcome modification uplift model were age on the network, demographic variables such as age, gender, sales region and subscriber ARPU for calls, SMS, DATA subscriptions and mobile money usage. When we compare the with the variable use by other recent scholars they varied between calls, SMS and data usage together with demographic data and age on the network data. This model can be therefore applied to telco service provider across regions to help mitigate churn.

4.12 Measuring and comparing the impact of subscriber retention campaigns using propensity modelling.

FIGURE 4.8.1 MEASURING THE IMPACT OF SUBSCRIBER RETENTION CAMPAIGNS



From our telco data a conversion rate 5 % of the sample was from the control group, the group that did not receive a retention offer and a conversion rate of 7.5 % from the sample subscribers that received the retention offer. The effectiveness of the retention of was 2.5 per cent. The conversion rate for the top 1 percent for the control group was 49.8% while for the treatment group was 50% resulting to the uplift effectiveness of 0.2% in the target group selected by the uplift model. The selection effect (bias) of the target group selected by the propensity model was 44.38%. the true uplift i.e., the effectiveness of uplift model selection for a given retention incentive was -2.3%. the results are summarized in figure 4.8.1 measuring the impact of churn retention campaign.

From the results above we can conclude that the Propensity is negatively correlated with uplift which results in negative value added from retention measures based on a propensity model. Leveraging on uplift model brings about a positive value added from a retention campaign. Propensity is only weakly positively related correlated with uplift which results in positive value add from a retention campaign based on a propensity model. its effectiveness can improve by leveraging on an uplift model.

Traditional propensity modelling enables one to foresee the future as it is strictly decreasing response rate prove that propensity models do work at what they are designed for . Customers with predicted high likelihood to accept an offer has a high probability of being retained compared with subscribers with low predicted likelihood to accept a retention offer. Traditional propensity modelling is not helpful at affecting the future in our favor. Propensity models do not predict incremental churn trends, nor does it predict churn management failures whether the retention campaign was not able to create any incrementality when compared to the mean treatment effect nor does it detect analytical failures as it does reflect decreasing uplift rates. Uplift modelling is good at identifying subscribers that are most sensitive to retention offers Uplifts are decreasing with each decile. Uplift modelling is not good at foreseeing the future. Propensity is not decreasing with deciles

CHAPTER 5

5.0.1 Introduction

This chapter provided the summary of the findings from Chapter Four, Conclusions and Recommendations of the study based on the objectives of the study. It was a very crucial chapter as it would make summary of the whole research, conclude and make recommendations. The objectives of this study were: To identify the attributes that determine subscriber churn in the Kenya telecommunication sector, to develop a predictive and uplift model for subscriber churn and to evaluate the proposed models.

5.0.2 Conclusions

For Safaricom plc the major problem was identifying drivers of churn and mitigating churn across the network. The greatest drivers for churn based on the data collected was AGE and Age on the network which is the length of time a number has stayed on the network. For age this implies that marketing initiatives need to be implemented for the younger generation and affordable products need to innovated for that segment as they have the highest likelihood of churning. For age on the network, it was noted the highest churn rate was experienced for lines that were less than 90 days on the network. This pointed out Gaps on the onboarding process and also indicated high likelihood of fraudsters who buy lines temporarily and discard them after executing their fraudulent initiatives.

From the Objective one results it can be concluded that among the various methods to perform feature selection. In this study we explored the one-way ANOVA for numeric variables, Cramer V for categorical variables, Extra-trees feature importance method and Lasso with ANOVA techniques. We adopted the feature selection using one-way ANOVA for numeric variables, Cramer V for categorical variables. It was adopted because of its ease of implementation and accuracy in results. The features that were selected included Age, AON, voice revenue, MM revenue, DATA revenue, SMS revenue, Subscription revenue, other revenue, Gender and Sales Region.

From Objective two results it can be concluded that among the various techniques we experimented on such as extra-trees ensemble method and the MLENS ensemble method. Better results were achieved using the extra trees ensemble method in terms on accuracy it was 96% and for MLENS the accuracy was at 94 %. The model proved very effective at predicting churn when compared to the MLENS model. For uplift modelling. For the uplift modelling we experimented with uplift tree, two model approach and propensity classification approach. Classification uplift modeling was the best achieving an uplift of close to 5.24% in segments compared to the other uplift modeling techniques such as uplift tree and two model approach that was less than 2%. Hence the propensity modeling approach was best suited for predicting uplift for the telecommunication industry in Kenya.

From Objective three results it can be concluded that among the various techniques we experimented on such as extra-trees ensemble method and the MLENS ensemble method better results were achieved using the extra trees ensemble method in terms on recall, AUC and precision. Though MLENS ensemble technique performed well when it came to recall at 98% its AUC was 96 % and precision was at 93.1% runtime was much longer than the extra tree ensemble. The recall achieved for extra trees was 95.9 % , AUC metric was 99% and precision was 95.6%. . The Extra trees model hence was very effective when it came it its evaluation for its performance metrics. The AUUC metrics for the uplift tree, propensity and two model approaches were27, 62 and 56 respectively.

Limitations of the study was due to the highly sensitive nature of the data it had to be anonymized and pseudomized a process that takes a long period of time. Also computing power was a limitation based on the large amount of data required to produce a viable model

5.0.3 Contributions to knowledge.

The researcher expands uplift modeling by introducing the ensemble learning using the m-lens python library and extra trees ensemble methods. This technique emphasizes the use of ensemble learning to predict churn on the network and to create the uplift model to identify profitable segments that needs to be targeted with retention measures. The researcher provides proof of the upside of uplift modeling versus predictive modeling by doing an experimental case study using our novel propensity class outcome modification approach. This approach when compared to other existing models by Chen et al (2020) provides better uplift metrics and is able to identify profitable segments that can be targeted with retention measures thus saving costs for telcos in Kenya.

The study can easily contribute in the development of Churn management systems to help management to select and manage the portfolio of churn in telecommunication industry. Such systems would ideally be incorporated in an effective retention management Systems. The key to this is the development of a combination of ensemble and uplift models in prediction of churn and prescribing what measure in terms of retention incentives that can be taken to reduce churn. In addition, more thought needs to be given to the Data privacy issues and other impediments that may hinder wide implementations of such models across industries.

5.0.4 Recommendations of further research

This study focused on try to find churn drivers that contribute to churn from our findings its recommended that more effort has be put onboarding processes for new customers within the network and tracking of registrations of new customers to eliminate fraudulent subscribers from joining the network. The marketing team has to put in more effort to create affordable products to certain age groups to alleviate those segments from churning. The other drivers for churn include loan status for lines from mobile loans, billing issues multi sim adoption by customers and lack of proper handling of customers. To mitigate these drivers its recommended for Safaricom plc to have efficient onboarding processes through automation, efficient recycling of subscriber lines to ensure new subscribers have working lines. Its also recommended that Safaricom plc to adopt automated fraud detection on the network and to perform continuous customer education on fraud. Finally, its recommended for Safaricom PLC to have excellent product and service offerings to remain competitive.

This study also focused on one type of retention incentive to achieve uplift on our telco dataset. Therein lies and opportunity to research where multiple retention incentives are given to improve conversion rates based on specific segments. This telco dataset had five segments inherently identified by their attributes. This will go a long way in improving churn management practices in the telco industry. More research can also be done to improve the uplift tree modelling to get even better results on saving incurred by business. And finally, more research can go into calculating the cost savings from the uplift calculations given by the retention measures.

6.0 Appendix

6.0.1 APPENDIX A: TIMELINES

Project Timelines		
Activity Description	**start**	**Finish**
supervisor allocation	13-Nov-20	18-Nov-20
Literature review	18-Nov-20	27-Dec-20
Research design	27-Dec-20	7-Jan-21
Proposal Drafting	7-Jan-21	13-Feb-21
Proposal Presentation	13-Feb-21	20-Mar-21
Data Analysis	20-Mar-21	10-july-21
Project Defense - Final Dissertation	29-May-21	31-Jul-21

6.0.2 APPENDIX B: PROJECT BUDGET

ITEM	
1. Stationary	4000
2. Printing and photocopying	4000
3. Airtime	2000
4. Internet	20000
5. Binding Documents	2000
Total	32000

REFERENCES

R. Jadhav and T. Usharani (2011), “Churn Prediction in Telecommunication Using Data Mining Technology,” International Journal of Advanced Computer Science and Applications, vol. 2, no. 2.

M. Sewell (2008), "Ensemble learning," RN, vol. 11, no. 02.

E. Chen,(2011, October) “Winning the Netflix Prize: A Summary,” Edwin Chens Blog Atom. [Online]. Available: http://blog.echen.me/2011/10/24/winning-the-netflix-prize-asummary/.

Georges D.Olle Olle and Shuqin Cai (2014, February), “A Hybrid Churn Prediction Model in Mobile Telecommunication Industry”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1.

Preeti K. Dalvi, Siddhi K. Khandge (2016), “Analysis of Subscriber Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression”, PVG’s COET, Pune, India

DeSA UN (2020) World population prospects: the 2012 revision. Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, New York

Sector-Statistics-Report-Q1-(2020-2021).pdf (ca.go.ke) https://ca.go.ke/wp-content/uploads/2020/12/Sector-Statistics-Report-Q1-2020-2021.pdf CAK webpage

Bhattacharya C (1998) When subscribers are members: subscriber retention in paid membership contexts. J Acad Mark Sci 26(1):31–44

Athanassopoulos A (2000) Subscriber satisfaction cues to support market segmentation and explain switching behavior. J Bus Res 47(3):191–207

Farris PW, Bendle NT, Pfeifer PE, Reibstein DJ (2010) Marketing metrics: the definitive guide to measuring marketing performance. Wharton School Publishing

Yin, R. K. (2014). Case study research: Design and methods. Los Angeles, CA: Sage. Los Angeles

Bolton R N KN lemon et al (2006) The Effect of Service Experiences over Time on a Supplier's Retention of Business Subscribers

Mugenda, O.M. &Mugenda A.G. (1999, 2003). Research Methods, Quantitative and Qualitative.

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on artificial intelligence, 4–10 August 2001, Seattle, WA, pp 973–978

Bahnsen AC, Aouada D, Ottersten B (2015) A novel cost-sensitive framework for subscriber churn predictive modeling. Decis Anal 2(1):5

Mauricio AP, Payawal JMM, Dela Cueva MA, Quevedo VC (2016) Predicting subscriber lifetime value through data mining technique in a direct selling company. In: 2016 international conference on industrial engineering, management science and application (ICIMSA), Jeju, pp 1–5. https://doi.org/10.1109/ ICIMSA.2016.7504027

Buckinx W, Van den Poel D (2005) Subscriber base analysis: partial defection of behaviorally loyal clients in a non-contractual FMCG retail setting. Eur J Oper Res 164(1):252–268

Glady N, Baesens B, Croux C (2009) Modeling churn using subscriber lifetime value. Eur J Oper Res 197(1):402–411

Verbraken T (2012) Toward profit-driven churn modeling with predictive marketing analytics. In: Cloud computing and analytics: innovations in e-business services. Workshop on e-business (WEB2012), Orlando

Verbraken T, Verbeke W, Baesens B (2013) A novel profit maximizing metric for measuring classification performance of subscriber churn prediction models. IEEE Trans Knowl Data Eng 25(5):961–973

R. Michel, I. Schnakenburg, T. von Martens (2019). „Targeting Uplift". Springer, ISBN 978-3-030-22625-1

Neslin, Scott A., Sunil Gupta, Wagner Kamakura, Junxiang Lu, and Charlotte H. Mason(2006), "Defection Detection: Measuring and Understanding the Predictive Accuracy of Subscriber Churn Models," Journal of Marketing Research, 43 (2), 204-211.

Neslin, Scott A., Thomas P. Novak, Kenneth R. Baker, and Donna L. Hoffman (2009), "An Optimal Contact Model for Maximizing Online Panel Response Rates," Management Science,55 (5), 727–737.

Provost, Foster, and Tom Fawcett (2013), Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly Media, Inc.

Reichheld, Frederick (1996). The Loyalty Effect. Harvard Business School Press, Boston, MA. Rei

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2000), "Additive Logistic Regression: a Statistical View of Boosting," The Annals of Statistics, 28 (2), 337-374.

Friedman, Jerome H. (2002), "Stochastic Gradient Boosting," Computational Statistics and Data Analysis, 38, 367-378.

Ascarza, Eva, Peter Ebbes, Oded Netzer, and Matt Danielson (2016a), "Beyond the Target Subscriber: Social Effects of CRM Campaigns," Journal of Marketing Research (forthcoming).

Ascarza, Eva, Raghuram Iyengar, and Martin Schleicher (2016b), "The Perils of Proactive Churn Prevention using Plan Recommendations: Evidence from a Field Experiment." Journal of Marketing Research, 53(1), 46-60

Ascarza, Eva (2016), "Retention Futility: Targeting High Risk Subscribers Might be Ineffective," working paper.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009), The Elements of Statistical Learning (second edition), Springer: New York

Breiman, Leo (1996), "Heuristics of Instability and Stabilization in Model Selection," Annals of Statistics, 24, 2350-2383.

Venkatesan, Rajkumar, V. Kumar, and T. Bohling (2007), 'Optimal CRM Using Bayesian Decision Theory; An Application to Subscriber Selection," Journal of Marketing Research, 44 (4), 579-594.

Schweidel, David A. and George Knox (2013), “Incorporating Direct Marketing Activity into Latent Attrition Models,” Marketing Science, 32 (3), 471-87.

Lewis, Michael (2005b), “Incorporating Strategic Consumer Behavior into Subscriber Valuation,” Journal of Marketing, 29 (October), 230-238.

Y. Freund and R. E. Schapire,(1999, Sep) "A Short Introduction to Boosting," Journal of Japanese Society for Artificial Intelligence, Vol. 14, No. 5, pp. 771-780.

P. Datta, B. Masand, D. R. Mani, and B. Li(2000, Dec), "Automated Cellular Modeling and Prediction on a Large Scale," Artificial Intelligence Review, Vol. 14, No. 6, pp. 485-502.

Chen, H, Harinen, T. lee, J, Y, Yung, M. & Zhao (2020). CausalML: python package for Causal Machine Learning arXIV preprint arXIV:2002.11631

Diemert, E, Betlei, A., Renaudin, C., & Amin, M., R. (2018, August). A large-scale Benchmark for uplift Modelling. in proceedings of AdKDD and Target Ad Workshop, KDD, London, United Kingdom

W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens (2012), New insights into churn prediction in the telecommunication sector: a profit driven data mining approach, Eur. J. Oper. Res. 218 (1) 211–229, https://doi.org/10.1016/j.ejor.2011.09.031.

N.J. Radcliffe, R. Simpson (2008)., identifying who can be saved and who will be driven away by retention activity., J. Telecommunication. Manager. 1 (2)

Perez-Marin (2012), Random forests for uplift modeling: an insurance customer retention case, in: K.J. Engemann, A.M. Gil Lafuente, J. Merigo (Eds.), Modeling and Simulation in Engineering, Economics and Management, Lecture Notes in Business Information Processing, 115, Springer Berlin Heidelberg, pp. 123–133, https://doi.org/10.1007/978-3-642-30433-0_13.

M. Sokolova and G. Lapalme (2009), “A systematic analysis of performance measures for classification tasks,” Information Processing & Management, vol. 45, no. 4, pp. 427–437, Har-Peled, S., Roth, D., Zimak, D. (2003) "Constraint Classification for Multiclass Classification and Ranking." In: Becker, B., Thrun, S., Obermayer, K. (Eds) Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference, MIT Press. ISBN 0-262-02550-7

McCallum, Andrew (2019, October). "Graphical Models, Lecture2: Bayesian Network Representation" (PDF). Retrieved 22.

Hastie, Trevor. (2001). The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations. Tibshirani, Robert., Friedman, J. H. (Jerome H.). New York: Springer. ISBN 0-387-95284-5. OCLC 46809224.

Heider, D; Senge, R; Cheng, W; Hüllermeier, E (2013). "Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction". Bioinformatics. 29 (16): 1946–52.

Max Kuhn, Kjell Johnson (2018). "Applied Predictive Modeling 2nd Edition. Springer ISBN-13: 978-1461468486

David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. ISBN 978-1-139-47731-4.

Bottou, Léon; Bousquet, Olivier (2012). "The Tradeoffs of Large-Scale Learning". In Sra, Suvrit; Nowozin, Sebastian; Wright, Stephen J. (eds.). Optimization for Machine Learning. Cambridge: MIT Press. pp. 351–368. ISBN 978-0-262-01646-9.

Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175.

Cormen, Thomas H. (2009) Introduction to algorithms. (Third ed.). Cambridge, Mass.: MIT Press. ISBN 978-0-262-27083-0. OCLC 676697295.