UNIVERSITY OF NAIROBI

Churn prediction modelling in B2B e-commerce: A case study of Twiga Foods.

BY

Elijah Kipleting Sawe

I56/35315/2019

A Thesis Submitted to the Department of Mathematics for Examination in Partial Fulfillment of this Requirements for the Award of Degree of Master of Science in Social Statistics of the University of Nairobi

November 2022

Master Project in Social Statistics **University of Nairobi**

November 2022

Churn prediction modelling in B2B e-commerce: A case study of Twiga Foods.

Research Report in Mathematics

Elijah Kipleting Sawe

Department of Mathematics
College of Biological and Physical Science
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the Department of Mathematics in partial fulfilment for a degree in Master of Science Social Statistics

Submitted to: The Department of Mathematics, University of Nairobi, Kenya

Abstract

The informal structure of African retail, which is mostly fueled by layers of brokers and middlemen, is fundamentally changing as a result of technology. Selling locally grown fresh produce to Kenyans should be simple, but this wasn't the case at the time due to the disjointed distribution networks and ineffective supply chains. Twiga foods started in 2014 has managed to build efficiency though technology. The digital plat-form helps to on-board farmers and aggregate demand of the produce by the retailers.The farmers then supply to Twiga Foods who then deliver to retailers for free.

Evidently, acquiring a new customer is costlier than retaining a customer and therefore, one of the strategic and cost optimization initiatives by such digital platforms also known as e-commerce platforms is to retain customers. To retain the customers, such businesses must build mechanisms that allow them to predict the probability of a customer churning immediately a customer is acquired so that they are able to undertake business measures that would prevent a customer from churning.

In this research, we build a machine learning model (Logistic Regression (LR) model) that predicts if a retailer will be retained or not, test the model, put it in production and identify different use cases the model could be applied.Logistic regression is a binary classification model; the goal of the model is to predict yes or no in Twiga’s case churn or retained. It utilizes a sigmoid function to assign a probability of between 0 and 1 to each retailer.

We then evaluated the model performance by comparing two scenarios. Scenario 1; we split the data randomly. Scenario 2; we arrange the data in ascending order and split the data using delivery date.

Scenario 2 had a higher percentage accuracy of 81% compared to Scenario1 at 52% accuracy, Scenario 2 had 76% precision/specificity compared to scenario 1 at 71% and Scenario 2 had a 75% Recall (sensitivity) compared to scenario 1 at 62%. Scenario 2 is an improvement from the conventional models and has a higher performance.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: Department of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

22/11/2022

Signature Date

ELIJAH KIPLETING SAWE
Reg No. 156/35315/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

22nd November 2022

Signature Date

Dr. Idah Orowe

Department of Mathematics,
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: orowe@uonbi.ac.ke

Dedication

I dedicate this work to my dear wife Damaris and daughter Netai.

Contents

Abstract ii

Declaration and Approval iv

Dedication vii

List of Tables ix

List of Figures x

Acknowledgments xi

1 CHAPTER 1: THE INTRODUCTION 1

1.1 The Background of Study 1
1.2 Research problem 2
1.3 Research Question 2
1.4 Objectives 3
1.5 Significance of the study 3

2 CHAPTER 2: LITERATURE REVIEW 5

3 CHAPTER 3: METHODS 7

3.0.1 Logistic Regression Model 7
3.0.2 Decision Trees Model 9

4 CHAPTER 4: DATA ANALYSIS AND RESULTS 12

4.1 Data Analysis 12
4.1.1 Data Curation 12
4.1.2 Exploratory Data Analysis 12
4.1.3 Feature Engineering 13
4.1.4 Model Selection 14
4.1.5 Model evaluation 17
4.1.6 Monitoring the Model 20
4.1.7 Model Improvements 20

5 CONCLUSIONS AND RECOMMENDATIONS 21

5.1 Conclusions 21
5.2 Recommendations 21

Bibliography 22

List of Tables

Table 1. Feature Engineering 13

List of Figures

Figure 1. Logistics Regression Model 7
Figure 2. Decision Tree 10
Figure 3. Fitting 14
Figure 4. Training 15
Figure 5. Loss 15
Figure 6. Duration 16
Figure 7. Learn rate 16
Figure 8. Metrics 17
Figure 9. Precision Threshold 18
Figure 10. Precision Curve 18
Figure 11. Confusion Matrix 19
Figure 12. ROC Curve 19
Figure 13. Score Threshold 19

Acknowledgments

Above all, God has been gracious to me throughout my academic journey. I thank him for the opportunity to access the education and the gift of good health. I owe this success to my supervisor, Dr. Idah Atieno Adhiambo Orowe who guided me throughout my project. I also acknowledge the lecturers Mr. Ndiritu John, Dr. Nelson Owuor Onyango and Dr. Rachel Jelagat Sarguta who gave me a strong foundation that prepared me for this paper. Finally, I acknowledge my team, the Business Intelligence team at Twiga Foods led by Gerald Muriuki for their support through the process.

Elijah Kipleting Sawe

Nairobi, 2022.

1 CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Churn rate is a measurement of how many people or things leave an organization over a given time period. It is one of the two main variables that affect how many consumers a company will serve on a consistent basis. Churn is derived from *'butter churn' a device used to convert cream into butter*. A butter churn can separate butter from cream by squeezing out the butter from the cream. Customer churn works in the same manner; a business acquires X number of customers over period Y but gets to carry with them a portion of X customers to period $Y + 1$. That processes of losing a proportion of X customers is referred to Customer Churn.

Retailer switch/churn is a metric of subscriber attrition from a mobile network operator service provider, according to Karianga James Macharia (2012). It is calculated by dividing the number of subscribers who stop using a specific service over the course of a given time period by the typical number of customers or people over the course of that same time period.

Twiga foods limited (*Twiga*) was cofounded by Peter Njonjo and Grant Brooke in 2014 with an aim of exporting bananas. In 2015, the cofounders quickly realized that there was enormous potential to not only support farmers to access market but also to contributeto Africa's access to quality and affordable food by creating efficiency around retailer economy.

Their main goal was to create a formal, organized, efficient, fair, and transparent local market in order to close gaps in food and market security. As of November 2022, the online platform has since connected over 9,149 farmers to marketand have supplied food to over 140,000 retailers who then sell to customers across 12Cities. Twiga, first operated in Nairobi before expanding to Kisumu, Eldoret, Kakamega, Nakuru and The Coast. The company has expanded to Kampala Uganda and has plans to expand to other African cities by end of 2022.

Sokoyetu (*twiga.shop*) is Twiga's digital platform developed by its internal Technology team. It is the marketplace where the retailers subscribe, place orders, and track the delivery of their orders. The use digital platform allows Twiga to be one of the most successful data driven entities due to massive data collected, analyzed, and used in improving both farmers and customer experience and food quality.

Twiga foods spends 20USD on average to acquire a new customer and about 2USD to retain an existing customer. As a business built through Private Equity funds, cost optimization is key and such initiatives as churn prediction are overly critical in the journey to ensuring that the business is profitable.

1.2 Research problem

The digital wholesale industry that Twiga operates in, over the past 5 years have seen lots of new entrants and although it has huge opportunity, the competition to retain high value and consistent customers has increased. Other than other digital/online platforms, there are legacy local and unstructured wholesalers who stage extremely high competition and have over the years earned customer loyalty. Switching from one digital wholesale to another or back to a local whole seller is free! If you consider the cost of acquiring a new customer (20USD) and take an average of 1 year to return the investment, customer churn is a priority.

To manage churn, Twiga foods has embraced sales force automation (SFA) which uses Machine Learning (ML) prediction model to classify a customer as churn or non-churn. The sales agents then use the SFA application, which receives the customer profile, to take the appropriate actions to guarantee that clients who are profiled as likely to leave are given extra attention and engagement in order to keep them. To achievesuch level of automation, there is need create a dependable Machine Learning Model that will be able to profile a customer as soon as the customer is onboarded.

1.3 Research Question

The research question states that:
Are we able to improve prediction of probability of a new customer churning?

To address this research question above, the following two sub questions are formulated:

1. Does the prediction model perform better when data is arranged before splitting or when data is randomly split?
2. What percentage of retailers that end up churning does the algorithm successfully find?
3. Of all the Retailers that the algorithm predicts will churn, how many of them do churn?
4. Does the model correctly identify the true positives?

1.4 Objectives

Primary objective:

The primary objective of the research paper is to identify the most dependable e-commerce customer churn prediction model using machine learning by studying three approaches, logistic regression model one where data is ordered and another when its randomly split and the Decision Trees' model.

The specific objectives of this research are:

1. To compare the performance of the Logistic regression models, when data is arranged before splitting and when data is split randomly and the Decision Trees models, using performance metrics and validation with reference to the existing work.
2. To establish the percentage of retailers that end up churning, the machine learning algorithm successfully predicts.
3. To establish if the model can correctly identify true positives and false negatives.

1.5 Significance of the study

Urbanization and human population growth are occurring all around the world, but particularly in Africa. These urbanizing and expanding populations require enough wholesome food for a productive existence. Twiga foods have over the past 8 years demonstrated the value of establishing formal wholesale ecosystem.

The key benefits include:

- Access to market for small holder farmers.
- Food safely through professional food handling techniques and traceability.
- Food security through affordable quality food and efficient value chains.
- Economic growth – Twiga has employed over two thousand people, was named the best Taxpayer of the year under the small and medium size companies in 2022.

Customer churn for Twiga foods then negates above efforts since it means that the business would continuously incur high customer acquisition costs and might not reach profitability quickly.

The research therefore will contribute immensely to the e-commerce sector and other similar business models by providing scientific evidence on the most accurate models for predicting customer churn and therefore proactively put measures to retain customers.

Conventionally, Data Scientists and researchers have used various models in prediction of churn such as Cox proportional model which is a regression model checks the association between the survival time of customer and one or more predictor variables such as competitor activities, demographic factors, social factors usage factors, economic fac- tors, etc. Those who have considered using the logistic regression model have not clearly outlined whether the model would perform differently if the data were arranged before partitioning.

Finally, for tech led and data driven companies, this model will allow them to automate the process of customer acquisition and retention by proactively establishing the customers' probability of churning so that specific treatment can be applied to specific customers.

2 CHAPTER 2: LITERATURE REVIEW

Kairanga James Macharia (2012)[2] settled on use of Cox model in predicting customer churn in telecommunication Industry. In his study, he argued that Telecoms have changed focus from growing customer base to customer retention. The research demonstrated the cost implication of acquiring a new customer over retaining a customer. It also demonstrated how disloyal customers could be since switching from one service provider to another is ridiculously cheap. In his study, he compared Cox proportional hazard model -a model based on the theory of survival analysis to the decision tree model, commonly used in data mining. He evaluated the models on Safaricom Limited's pre-paid customers. His finding was that the decision tree model performed better than the Cox proportional model. In summary, Kairanga's study recommends the use of either model in predicting customer churn with one of the independent variables being competitor monthly activities that could contribute churn.

Powers (2001) established that e-commerce platforms suffer significant customer churn. He argues that a churn rate up to 10% is within reasonable range. Failure to retain customers is costly to any business and it caused loss of revenue in the tune of millions of dollars. The study recommends forming new product lines and implementing customer loyalty programs e.g cashbacks and reward points as mechanisms that would help to increase retention of customers. Powers however argues that there is no standard model for calculation of customer churn which is simply the percentage of customers who left by month or year end.

Keaveney (1995)[3], investigated how various events caused customers to switch from one service provider to another. She did a survey with a sample size of over 500 service customers and came up with a list of about 800 events by a service provider that affect customer retention.

She classified the behaviors into eight segments namely, pricing, core service failures, service encounter failures, employee responses to service failures, and competition. In her study, she points out that all these events can be controlled to some extend however, there are some other events that cause customer churn that a business may not have control over.

Dang Van Quynh (2019)[1], predicted churn in the computer software security industry using a comparative approach. He compared the below models

- Logistic regression
- Decision Tree's
- Random Forest

The exploratory data analysis (EDA) was conducted to identify any data gaps before feature engineering was conducted to treat any data issues e.g., the outliers, gaps, and other very random features in the data. Dang encountered a problem of skewed data with a 95% as non-churners and 4.6% as churners. To fix the issue, he suggested the use of over sampling and under sampling. He used the contemporary 7:30 rule where 70% was used to train and 30% was used to evaluate the models. The model performance was measured using the confusion matrix for precision, recall and F-measure.

The outcome of the study was that the Logistic regression model performed better in predicting non-churner in comparison with the other two models. Key to note is that as per the interpretation of the confusion matrix, the Logistic regression model showed bias to the class with most customers which were the churners. To mitigate against the bias, the size of non-churners sample was reduced. As a result, more churners were predicted than before. The research paper established that accuracy is a good measure of model performance, but when predicting churn, data analysts should be cautious.

3 CHAPTER 3: METHODS

Study design and setting

Case study of **Twiga foods** ltd between 1st January 2019 and 31st October 2022.

Search strategies

Using prisma 2020 guidelines Google scholar, Research gate, PubMed, Cochrane library and University of Nairobi Repository were utilized to search for articles.

Eligibility criteria

Timeframe

Assessment of outcome

Evaluated the model based on standard metrics such that comparing the models *accuracy, precision, confusion matrix, ROC* (receiver operating characteristic curve)

Data extraction

SQL queries to extract data from the **BigQuery** data warehouse on GCP

Risk of bias

To mitigate risk of bias, I used model evaluation criteria (overfitting & oversampling)

Machine Learning is a type of Artificial Intelligence (AI) premised on an algorithm that is able to take in data (train data) and apply statistical analysis techniques and models to anticipate an output. It is also able to update outcomes when new data is availed.

Types of machine learning

- **Supervised learning** – when labelled data is ran through a computer algorithm which is able to estimate the mapping function to the extend that it can predict the output variable *(y)* for that data set when you have new input data *(x).*
- **Unsupervised learning** – an example is the clustering and association problems where unlabelled, uncategorized data is run through a computer algorithm without prior training.
- **Reinforcement learning** – through interacting with its surrounding, the reinforcement learning model learns. The logic is that the agent ls rewarded or penalized depending on the computer algorithm learns.

Machine learning is built upon a statistical framework and statistical models e.g. regression models etc.

3.0.1 Logistic Regression Model

Logistic regression is a binary classification model, the goal model. It utilizes a sigmoid function to assign a probability of between 0 and 1 to each retailer.

Logistic regression seeks to:

- Calculate the likelihood that an event will occur based on the values of one or more independent variables, which may be either numerical or categorical.
- calculate the odds of an event happening for a randomly chosen observation versus the odds of the event not happening.
- foresee how several variables will affect a binary answer variable.
- categorize observations by calculating the likelihood that each observation belongs to a specific category.

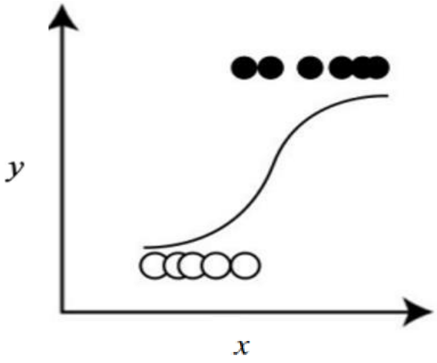


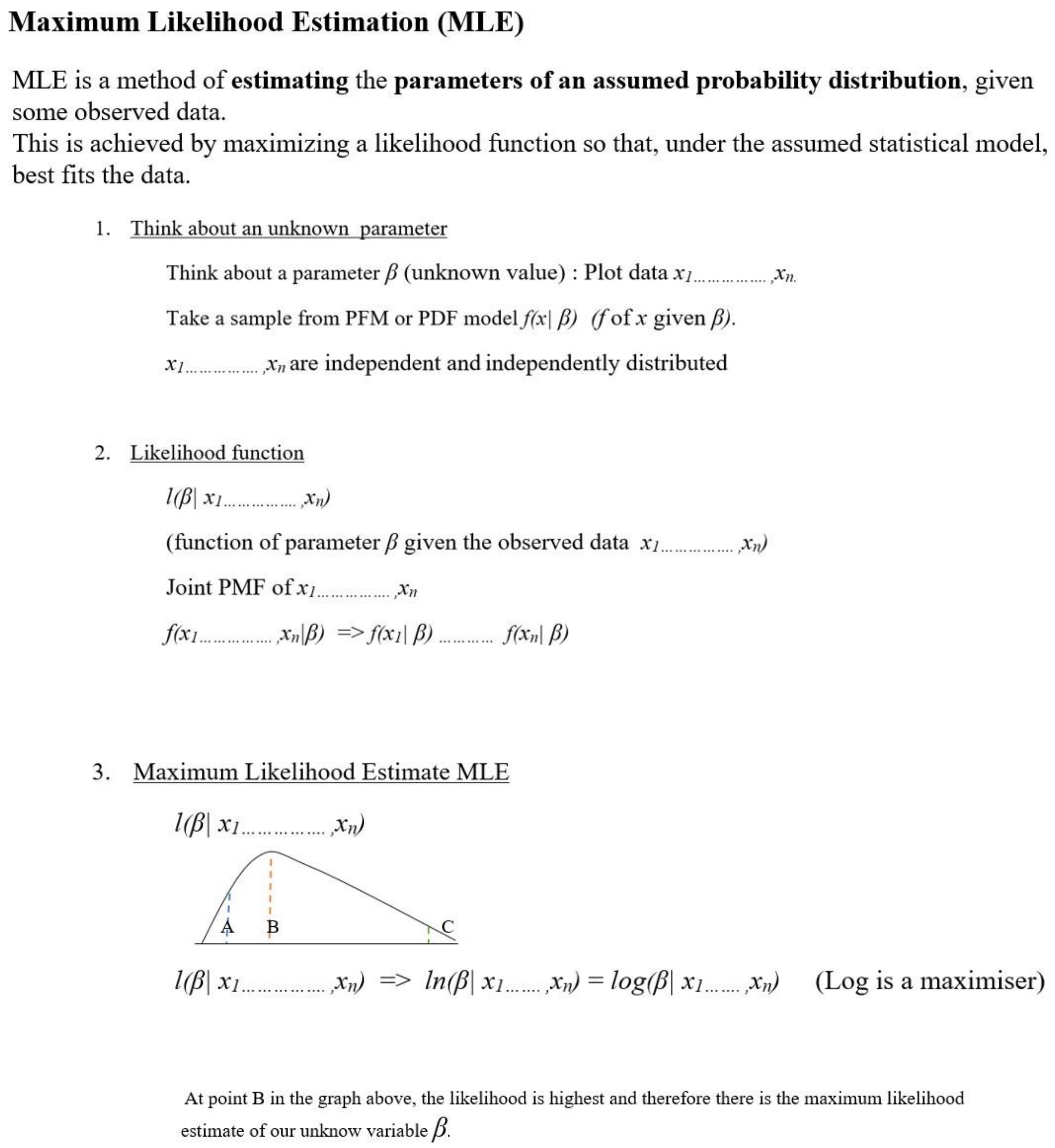
Figure 1. Logistics Regression Model

x = *the* ***independent*** *variable* ***(numeric)*** *for our case the are; (days* ***since registration/customer lifetime)*** *(****delivery_date****), products ordered, customer_type, order_value, no_of _returns & order_frequency)*

**This study used a single model variable the only significant variable.*

y = *the* ***dependent*** *variable* ***(Boolean/binary)*** *for our case 0,1 (Churn or Non-churn retailers)*

Historically, one of the first uses of regression-like models for binary data was bioassay results, (Finlay, 1997). In Finlay’s study, the outcomes were the proportions or percent- ages of success; for example, the proportion of experimental responses to given doses of drugs to treat a condition. The objective was to describe the probability of responding positively to treatment.



Maximum Likelihood Estimation (MLE)

MLE is a method of **estimating** the **parameters of an assumed probability distribution**, given some observed data.
This is achieved by maximizing a likelihood function so that, under the assumed statistical model, best fits the data.

1. Think about an unknown parameter

   Think about a parameter β (unknown value) : Plot data $x_1 \ldots\ldots\ldots\ldots ,x_n.$

   Take a sample from PFM or PDF model $f(x|\ \beta)$ (f of x given β).

   $x_1 \ldots\ldots\ldots\ldots ,x_n$ are independent and independently distributed

2. Likelihood function

   $l(\beta|\ x_1 \ldots\ldots\ldots\ldots ,x_n)$

   (function of parameter β given the observed data $x_1 \ldots\ldots\ldots\ldots ,x_n)$

   Joint PMF of $x_1 \ldots\ldots\ldots\ldots ,x_n$

   $f(x_1 \ldots\ldots\ldots\ldots ,x_n|\beta) \;=>\; f(x_1|\ \beta) \ldots\ldots\ldots\ \ f(x_n|\ \beta)$

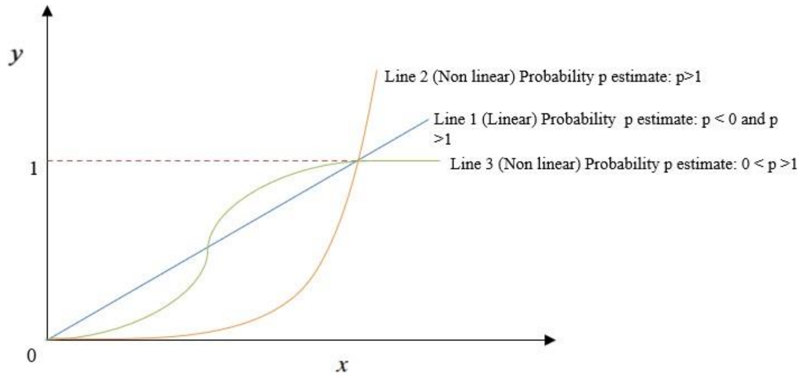
3. Maximum Likelihood Estimate MLE

   $l(\beta|\ x_1 \ldots\ldots\ldots\ldots ,x_n)$

   $l(\beta|\ x_1 \ldots\ldots\ldots\ldots ,x_n) \;=>\; ln(\beta|\ x_1 \ldots\ldots ,x_n) = log(\beta|\ x_1 \ldots\ldots ,x_n)$ (Log is a maximiser)

   At point B in the graph above, the likelihood is highest and therefore there is the maximum likelihood estimate of our unknow variable β.

Deriving the Logistic regression model



STEP 1: Selection of the correct line

$$line\ 2 \Rightarrow \hat{p} = e^{(b_0+b_1x)}$$

This equation (non-linear/ *p* estimate as an exponential function of *x*) However, this line only works when *p* estimate is greater than zero, but does not work when *p* estimate is capped at 1(100%) as it is in our case.

$$line\ 3 \Rightarrow \hat{p} = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}$$

This is the perfect line that represents our problem. The estimated probabilities of *p* is greater or equals to *0* and less or equals to *1 i.e.* $[0 \le \hat{p} \le 1]$

STEP 2: Simplifying the selected line

$$\hat{p} = \frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}} \Rightarrow \hat{p} = \frac{1}{1+e^{-(b_0+b_1x)}}$$

STEP 3: Prepare a model that we can fit into a machine learning model as a a linear model

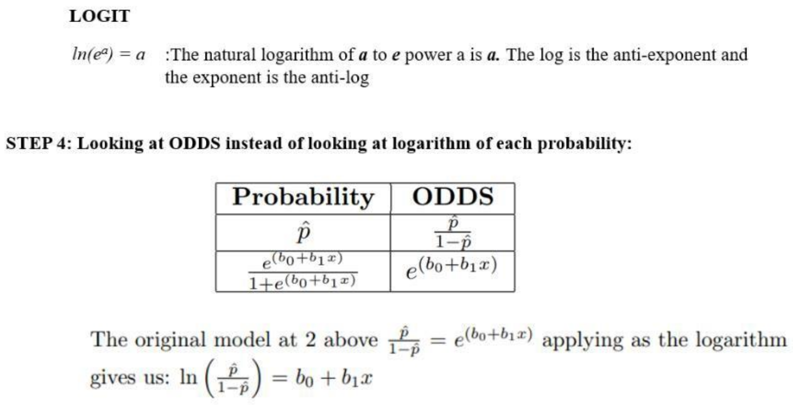
This brings us to two concepts

i. ODDS

ii. LOGIT

ODDS

$\text{ODDS (A)} = \frac{p(A)}{1-p(A)}$ The odds of A is equal to probability of A happening divided by the probability of A does not happen.



LOGIT

$ln(e^a) = a$:The natural logarithm of ***a*** to ***e*** power a is ***a.*** The log is the anti-exponent and the exponent is the anti-log

STEP 4: Looking at ODDS instead of looking at logarithm of each probability:

Probability	ODDS
$\hat{p}$	$\frac{\hat{p}}{1-\hat{p}}$
$\frac{e^{(b_0+b_1x)}}{1+e^{(b_0+b_1x)}}$	$e^{(b_0+b_1x)}$

The original model at 2 above $\frac{\hat{p}}{1-\hat{p}} = e^{(b_0+b_1x)}$ applying as the logarithm gives us: $\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = b_0 + b_1x$

3.0.2 Decision Trees Model

Decision trees is premised around dividing data into groups.

Think of a tree that has the roots, main stem, branches, and leaves. Assuming the roots is the main data set, stem, branches, and leaves are rules, stem being the first rule, the branches being the second rule and the leaves being the third rule, the first rule splits the entire data set into some number of pieces, and then the second rule may be applied to each piece of the first rules outcome. The different pieces will form a second generation of pieces. In another scenario, a piece in the preceding layer may either be split or left to form its own final group.

For this study, the decision tree is significant as it helps to check the performance of the Logistic regression model.
There are two types of categories of Decision Trees, regression trees and Classification trees.

In the case of predicting a continuous outcome, regression trees are use whereas
the classification trees are used in the case of predicting a binary outcome. For our case of retailer churn, we will use classification decision tree.

Building The Decision Tree

Even though the basic concept of a decision tree is built around choosing how to construct the tree is the most interesting part of the process. The sample dataset is divided into two subsets, a training set, and a test set, with a split of 80% and 20%, respectively. To one of the two subsets, the retailers are apportioned at random. To construct our decision tree, the training dataset is required. The test set will be necessary in the future to evaluate the predictive model's precision. trees that make decisions

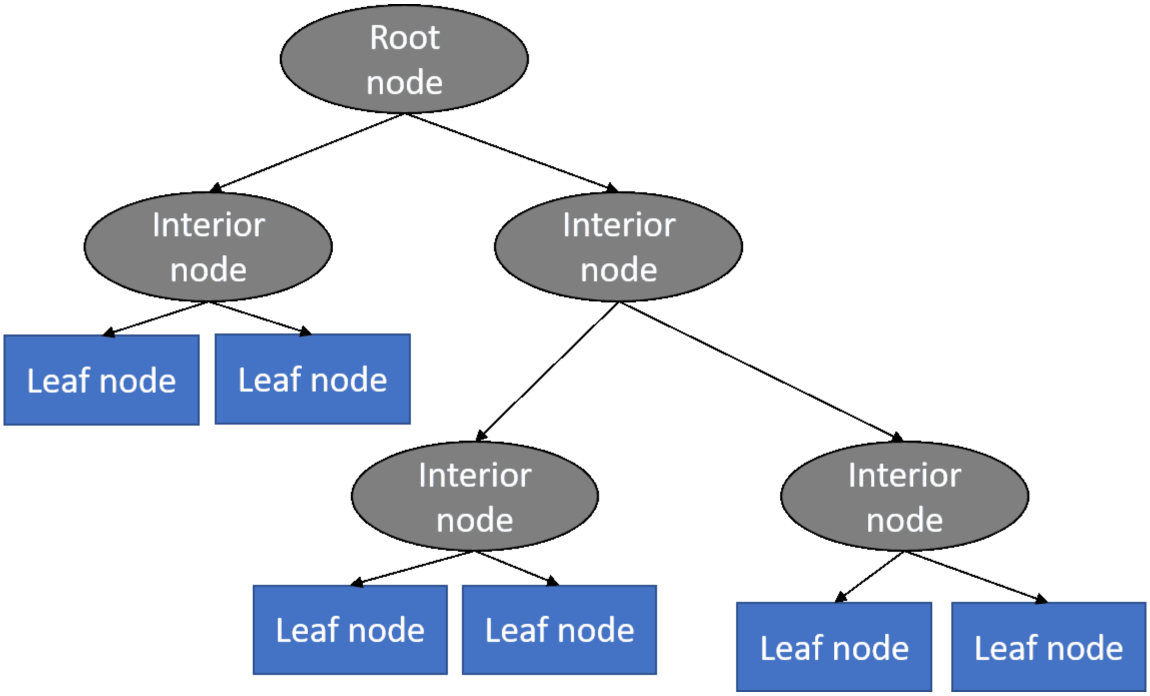


Figure 2. Decision Tree

are quite simple to build. It requires data cleaning and feature engineering then you aregood to fit the model.

There arenumerous measures to purity the decision tree (supervise). In e-commerce, the measure used entropy (entropy is a measure of lack of predictability). The formula of entropy isgien as follows:

$$entropy = -p_1 \cdot \ln(p_1) - p_2 \cdot \log(p_2) - \cdots = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$$

p_i is the proportion of property i in the set.

Entropy is then applied to the dependent variable ‘Churn’ which is categorized to 0,1. A bucket that’s pure will have entropy 0 while a bucket in which at least half the retailers would churn have entropy 1. The objective is to keep on making the decisions until you arrive at a scenario where the bucket has an entropy that is as close to zero as possible.
Another key concept in Decision Trees model is the information gain (IG).
The IG helps to establish the most optimal way to create a bucket i.e to split the data.

$$IG = entropy(root\ node) - p(c_1) \cdot entropy(c_1) + p(c_2) \cdot entropy(c_2)$$

c_1 and c_2 are the two resulting nodes after the split.

$p(c_1)$ the proportion of the customers that end up in node c_1.

When all possible splits are examined, the split which results in the highest informationgain is chosen.

4 CHAPTER 4: DATA ANALYSIS AND RESULTS

4.1 Data Analysis

The data for this study was obtained from Twiga foods limited. A research strategy is a plan for how a researcher will approach addressing his or her research topic, according to Saunders, Lewis, and Thornhill (2016)[5]. This study uses the case study approach, one of the eight categories of strategies mentioned by Saunders (Saunders, Lewis, and Thornhill, 2016 [5]), since it enables researchers to comprehend complicated social events while preserving the comprehensive and significant elements of actual events. (Yin, 2014). Yin (1984:23) [10] , "An empirical inquiry that analyses a contemporary phenomenon inside its real-life setting; when the boundaries between phenomenon and context are not readily visible; and in which numerous sources of evidence are utilized," is how the case study research technique is defined. This research emphasized:

1 Retailer registration data – Retailer registration data contains the registration date, retailer unique id and other Know Your Customer details.

2 Orders – the data set contains the orders generated by the company. It has the orderdate, retailer_id, order_id and order revenue.

3 Delivery data – Delivery data contains the details of an order that has been fulfilled.

4.1.1 Data Curation

The organizing and integration of data gathered from multiple sources is known as **data curation**. The data must be annotated, published, and presented in such a way that its value is sustained through time and that it is still accessible for reuse and preservation (M Stonebreaker et al. - Cidr, 2013)[7]. The Twiga Foods Technology team's e-commerce platform is where the three data sets are gathered.

4.1.2 Exploratory Data Analysis

A preliminary analysis of the data was conducted for the following reasons:

1 To investigate for any incomplete, incorrect, or inaccurate data.

2 To understand and summarize the main characteristics of the data.

3 Provide a basis for the type of models to be selected for churn prediction.

4.1.3 Feature Engineering

Any measurable input that can be employed in a predictive model is referred to as a "feature." The act of choosing, modifying, and changing raw data into features that may be utilised in supervised learning is known as feature engineering.

Features used and their description

Features/Variables Used	Description
days_ordered	No of days in a particular month that the customer placed an order in a month
ordered_value	Value of order in KES per Month
products_ordered	No of unique products ordered by the customer in a month
days_delivered	No of days in a particular month that the delivery was made
delivered_value	Value of delivery in KES per month
products_delivered	No of unique products delivered to the customer per month
order_fulfillment_drops	No of days ordered/days delivered per month
order_fulfillment_product	No of products ordered/No of products delivered per month
order_fulfillment_value	Value ordered/Value delivered per month
days_since_last_order	Current_date - last_order_date (in days)
customer_type	Customer segmentation - by products they purchase
churned_status (Target)	YES = Customer churned. NO= Customer was retained
Avg_days_delivered_week	Avg no of Days deliveries were made in a week
Avg_delivered_value_week	Avg value of deliveries were made in a week
Avg_products_delivered_week	Avg products delivered in a week
vendor_segment	Vendor segmentation based on basket size, frequency of order and value of order (platinum, gold, silver, bronze)
latitude	Location: Latitude:
Longitude	Location: Longitude

Table 1. Feature Engineering

Under Sampling

By retaining all the data from the minority class and reducing the size of the majority class, random under sampling is used to balance out the uneven dataset.

Data Pre-processing

When a vendor did not make a purchase for a month for the first time, I labeled them as churned. Vendors that received a delivery at least once per month were labeled as retained. I removed the columns for shop id, depot name, route name, minimum order date, current date, last delivery date, and last order date that would not be helpful for churn prediction. We used a label encoder to convert any non-numeric features, such as customer type and vendor segment, into numeric data.

Scaling

I standardized features by removing the mean and scaling to unit variance using standard scaler. This allows us to have standard normally distributed data.

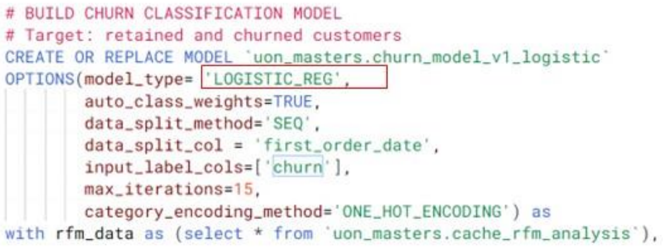
Splitting the Data

I split the data into 80% training, this is what I would use to train the model and 20% test data, which I used to evaluate how well the model trained.

4.1.4 Model Selection

Model Selection: Fitting of Logistic Regression

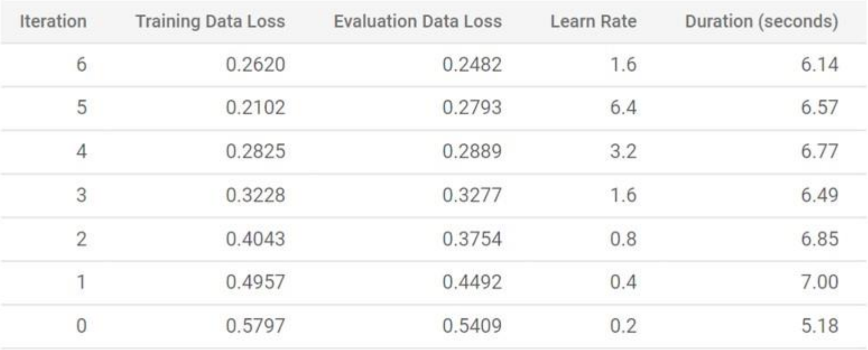
I used Google cloud platform's Artificial Intelligence and machine learning (AI/ML) as my tool of analysis



```
# BUILD CHURN CLASSIFICATION MODEL
# Target: retained and churned customers
CREATE OR REPLACE MODEL `uon_masters.churn_model_v1_logistic`
OPTIONS(model_type= 'LOGISTIC_REG',
        auto_class_weights=TRUE,
        data_split_method='SEQ',
        data_split_col = 'first_order_date',
        input_label_cols=['churn'],
        max_iterations=15,
        category_encoding_method='ONE_HOT_ENCODING') as
with rfm_data as (select * from `uon_masters.cache_rfm_analysis`),
```

- **auto_class_weights TRUE** means that the model will balance class labels using weights for each class in inverse proportion to the frequency of that class.
- **data_split_col** – column to split **train 80% & test 20%** data sets (for $model_1$ the data is **ordered asc** by this column before splitting and for $model_{2,}$ data is split randomly.
- **input_label_column – the dependent variable** p **being predicted** i.e 'churn' a binary column with the features 1 = Churn and 0 = non churn.
- **category – the independent variables** $x_{1,}\ x_{2,}\ x_{3,}\ x_{4.....,}\ x_n$ i.e the numeric variables including;
  - ✓ ***(days since registration/customer lifetime) (delivery_date),***
  - ✓ *products ordered,*
  - ✓ *customer_type,*
  - ✓ *order_value, no_of_returns & order_frequency)*

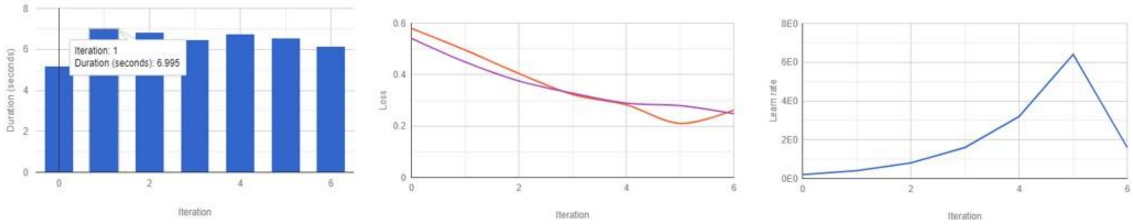
Model outcomes - Logistic Regression *(data is ordered)* $model_1$



Iteration	Training Data Loss	Evaluation Data Loss	Learn Rate	Duration (seconds)
6	0.2620	0.2482	1.6	6.14
5	0.2102	0.2793	6.4	6.57
4	0.2825	0.2889	3.2	6.77
3	0.3228	0.3277	1.6	6.49
2	0.4043	0.3754	0.8	6.85
1	0.4957	0.4492	0.4	7.00
0	0.5797	0.5409	0.2	5.18

- The model ran 6 sampling iterations.
- It was able to use at least 80% of the data to train and reserved the remaining 20% for model validation/testing.
- The 5th Iteration had the highest learn rate of 6.4.
- On average the model takes about 6 seconds to run 6 iterations, meaning, for a new registered customer, it would take 6 seconds to profile the customer as either **going to churn** or **will be trained.** This allows this model to be used for Sales force Automation.

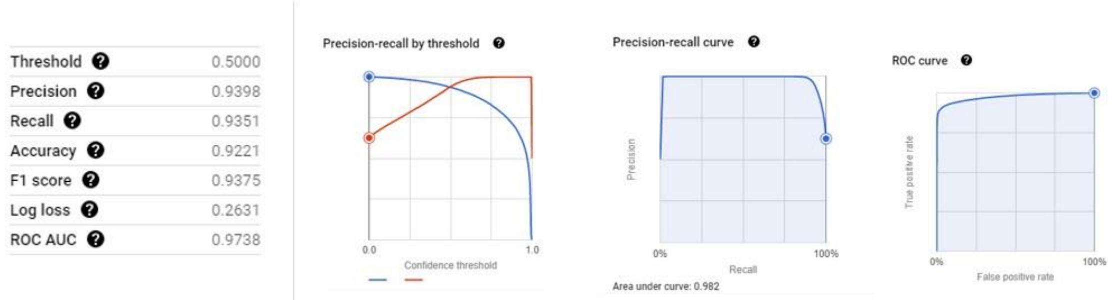
Model outcomes - Logistic Regression ***(data is ordered)*** *model$_1$*



- The model ran 6 sampling iterations.
- It was able to use at least 80% of the data to train and reserved the remaining 20% for model validation/testing.
- The 5th Iteration had the highest learn rate of 6.4.
- On average the model takes about 6 seconds to run 6 iterations, meaning, for a new registered customer, it would take 6 seconds to profile the customer as either **going to churn** or **will be trained.** This allows this model to be used for Sales force Automation.

Model outcomes - Logistic Regression ***(data is ordered)*** *model$_1$*

We evaluate the **model on accuracy and confusion** matrix.

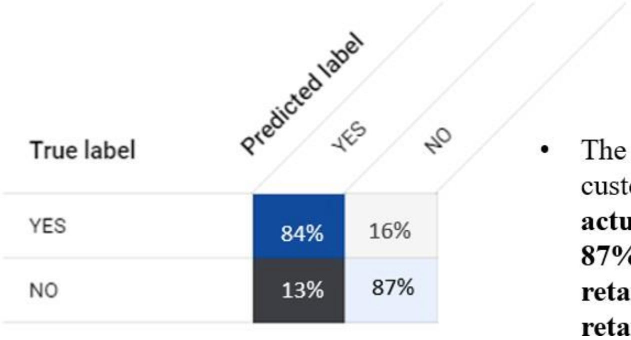


Threshold ❓	0.5000
Precision ❓	0.9398
Recall ❓	0.9351
Accuracy ❓	0.9221
F1 score ❓	0.9375
Log loss ❓	0.2631
ROC AUC ❓	0.9738

- **Model accuracy** – *true positives (correctly identified as churn) 92%*
- **Precision (specificity)** - of all the Vendors that the algorithm predicts will churn, how many of them do churn? Based on test data- 93%
- **Recall (Sensitivity)** - *What percentage of vendors that end up churning does the algorithm successfully find? – 94%*

Model outcomes - Logistic Regression *(data ordered)* *model$_1$*

Confusion matrix reports the number of **true positives, false negatives, false positives, and true negatives.** This allows more detailed evaluation of the model than simply observing the proportion of correct classifications (accuracy). Accuracy can yield misleading results if the data set is unbalanced i.e. that is, when the numbers of observations in different classes vary greatly.

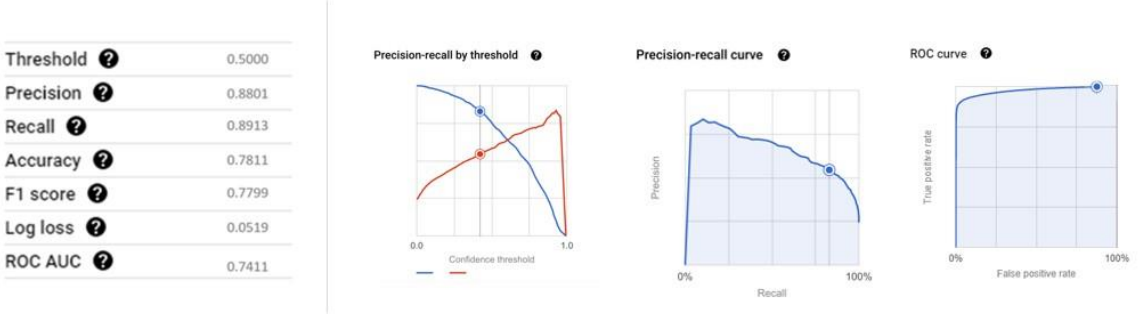


True label	Predicted label YES	Predicted label NO
YES	84%	16%
NO	13%	87%

- The model predicted that **84%** of customers **would churn** and they **actually churned**. It predicted that **87%** of the customers would be **retained** and they were **actually retained.**

Model outcomes - Logistic Regression *(data is random)* *model$_2$*

We evaluate the **model on accuracy and confusion** matrix.

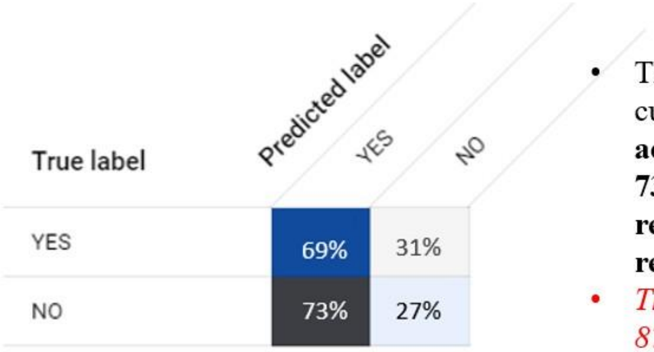


Threshold ❓	0.5000
Precision ❓	0.8801
Recall ❓	0.8913
Accuracy ❓	0.7811
F1 score ❓	0.7799
Log loss ❓	0.0519
ROC AUC ❓	0.7411

- **Model accuracy** – *true positives (correctly identified as churn) 78% compared to 92% when data is ordered.*
- **Precision (specificity)** - of all the Vendors that the algorithm predicts will churn, how many of them do churn? Based on test data- 88% compared to *93% when data is ordered.*
- **Recall (Sensitivity)** - *What percentage of vendors that end up churning does the algorithm successfully find? – 89%* compared to *94% when data is ordered.*

Model outcomes - Logistic Regression *(data is random)* *$model_2$*

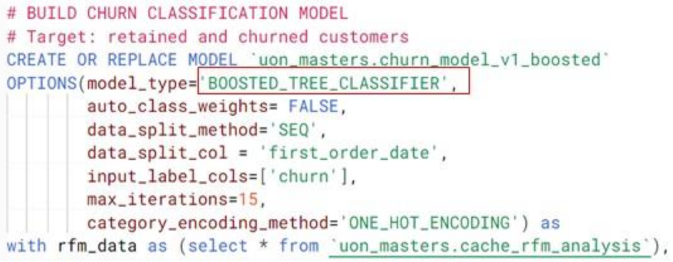
Confusion matrix



True label	Predicted label YES	NO
YES	69%	31%
NO	73%	27%

- The model predicted that **69%** of customers **would churn** and they **actually churned**. It predicted that **73%** of the customers would be **retained** and they were **actually retained.**
- *This is in comparison to 84% and 87% respectively when data is ordered.*

Fitting of Decision Tree Model (*$model_3$*)

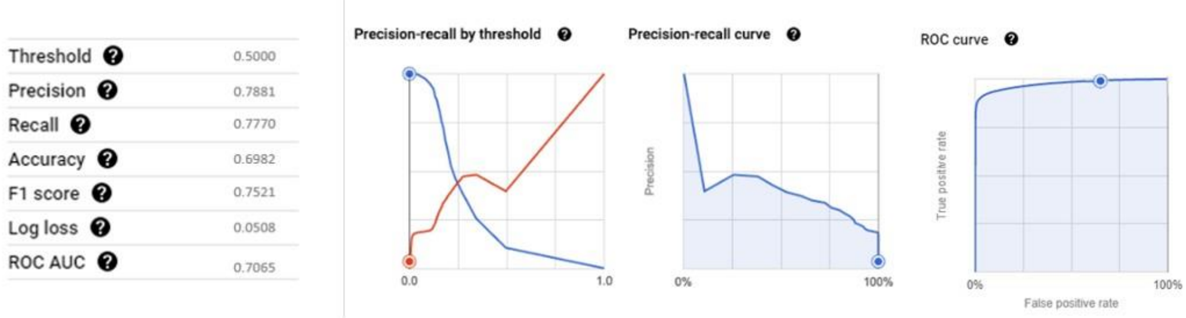


```
# BUILD CHURN CLASSIFICATION MODEL
# Target: retained and churned customers
CREATE OR REPLACE MODEL `uon_masters.churn_model_v1_boosted`
OPTIONS(model_type='BOOSTED_TREE_CLASSIFIER',
        auto_class_weights= FALSE,
        data_split_method='SEQ',
        data_split_col = 'first_order_date',
        input_label_cols=['churn'],
        max_iterations=15,
        category_encoding_method='ONE_HOT_ENCODING') as
with rfm_data as (select * from `uon_masters.cache_rfm_analysis`),
```

- **auto_class_weights TRUE** means that the model will balance class labels using weights for each class in inverse proportion to the frequency of that class.
- **data_split_col** – column to split **train 80% & test 20%** data sets
- **input_label_column – the dependent variable** p **being predicted** i.e 'churn' a binary column with the features 1 = Churn and 0 = non churn.
- **category – the independent variables** $x_1, x_2, x_3, x_4 ....., x_n$ i.e the numeric variables including;
  - ✓ ***(days since registration/customer lifetime) (delivery_date),***
  - ✓ *products ordered,*
  - ✓ *customer_type,*
  - ✓ *order_value, no_of_returns & order_frequency)*

Model outcomes – Decision Tree $model_3$

We evaluate the **model on accuracy and confusion** matrix.

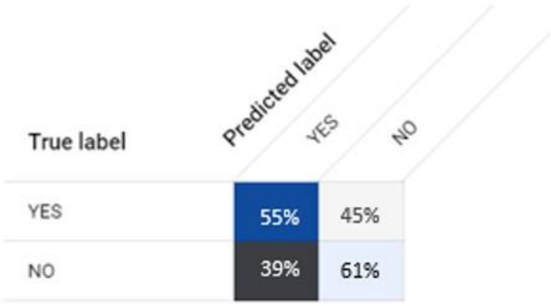


Threshold	0.5000
Precision	0.7881
Recall	0.7770
Accuracy	0.6982
F1 score	0.7521
Log loss	0.0508
ROC AUC	0.7065

- **Model accuracy** – *true positives (correctly identified as churn) 70% which is lower than the above Logistic regression models*
- **Precision (specificity)** - of all the Vendors that the algorithm predicts will churn, how many of them do churn? Based on test data- 79% *which is lower than Logistic regression models*
- **Recall (Sensitivity)** - *What percentage of vendors that end up churning does the algorithm successfully find? – 78% which is lower than the above Logistic regression models*

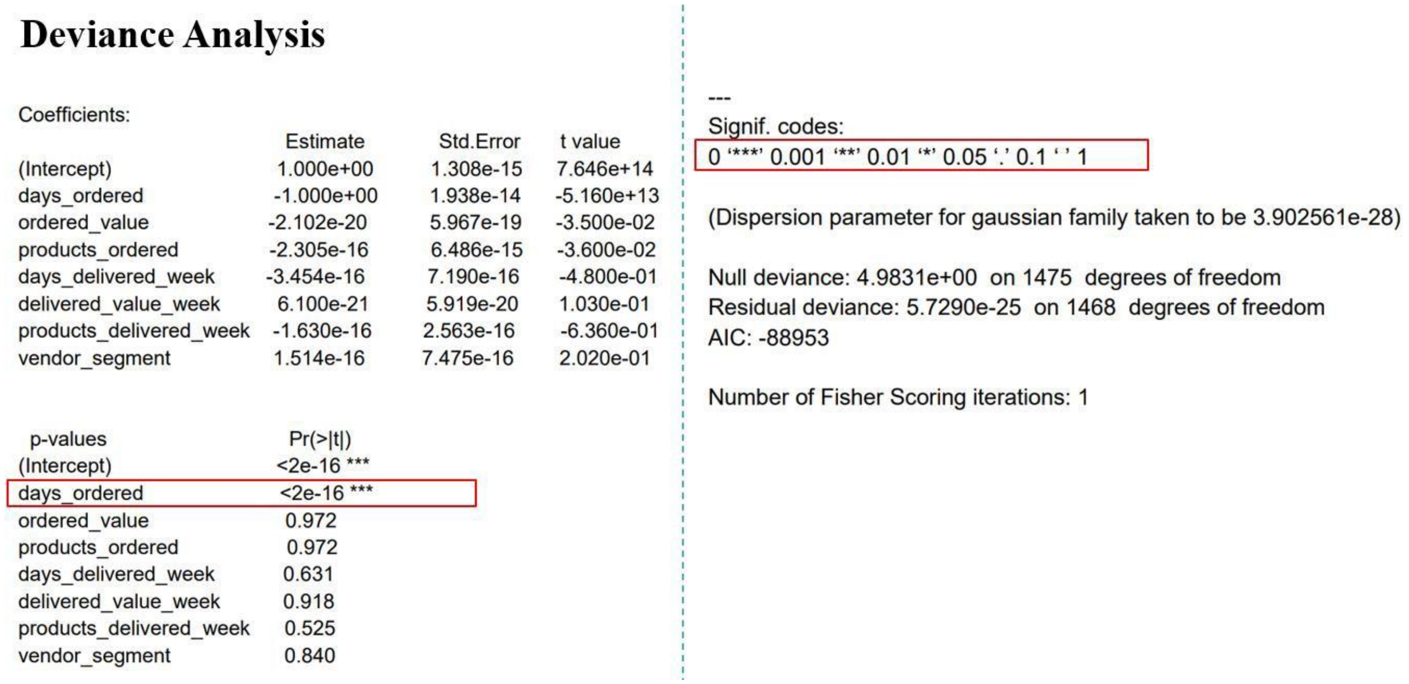
Model outcomes – Decision Tree $model_3$

Confusion matrix



True label	Predicted label YES	NO
YES	55%	45%
NO	39%	61%

- The model predicted that 55**%** of customers **would churn** and they **actually churned**. It predicted that **61%** of the customers would be **retained** and they were **actually retained.**
- *This is lower in comparison to the logistic regression models above.*



Deviance Analysis

```
Coefficients:
                         Estimate      Std.Error    t value
(Intercept)             1.000e+00     1.308e-15   7.646e+14
days_ordered           -1.000e+00     1.938e-14  -5.160e+13
ordered_value          -2.102e-20     5.967e-19  -3.500e-02
products_ordered       -2.305e-16     6.486e-15  -3.600e-02
days_delivered_week    -3.454e-16     7.190e-16  -4.800e-01
delivered_value_week    6.100e-21     5.919e-20   1.030e-01
products_delivered_week -1.630e-16    2.563e-16  -6.360e-01
vendor_segment          1.514e-16     7.475e-16   2.020e-01

  p-values               Pr(>|t|)
(Intercept)             <2e-16 ***
days_ordered            <2e-16 ***
ordered_value            0.972
products_ordered         0.972
days_delivered_week      0.631
delivered_value_week     0.918
products_delivered_week  0.525
vendor_segment           0.840
```

```
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3.902561e-28)

Null deviance: 4.9831e+00  on 1475  degrees of freedom
Residual deviance: 5.7290e-25  on 1468  degrees of freedom
AIC: -88953

Number of Fisher Scoring iterations: 1
```

days_ordered (time period between the time the customer is registered and when the customer places the first order) is the only significant variable.

The study used this column since it is the only significant X variable

The intercept is also significant.

4.1.5 Monitoring the Model

We closely monitor the model performance to identify and eliminate a variety of issues, including inferior quality predictions and poor technical performance. We use data of new vendors from August 2021 to December 2021. Below are the metrics

- Model accuracy – true positives (correctly identified as churn) 69%
- Precision (specificity) - Of all the Vendors that the algorithm predicts will churn, howmany of them do churn? Based on test data- 76%
- Recall (Sensitivity) - What percentage of vendors that end up churning does the al-gorithm successfully find? – 75%

4.1.6 Model Improvements

- Improve the model monitoring metrics by checking at the data distribution shifts, performance shifts.
- Track performance by segment, this will allow us to have a deep understanding of themodel quality on specific customer segments for Example, FMCG, FFV, FMCG/FFV.
- **Build Decision tree model**- They are interpretable and can manage complex feature relationships. We can get the features that had high influence in predicting churn.
- **Build Neural Networks** – Can model large and complex datasets.
- **Build Ensemble** – Combining different models to improve predictive outcomes.
- **Dashboarding/Monitoring** the models in production with current data to avoidmodel decay.

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The Logistic regression model performs better when data is arranged before splitting than when data is split randomly however on overall, compared to other methods such as useof decision trees, the logistic regression model overall, performs better in prediction ofcustomer churn.

40% to 60% of retailers acquired customers/retailers are churn customers. This is a very significant number bearing in mind the customer acquisition costs. In essence, e-commercecompanies should adopt automation of customer churn prediction models and tools to ensure that they spotlight and mitigate against any churn of customers

5.2 Recommendations

E-commerce platforms should use AI and ML to automate their customer acquisition and retention process and in general the Sales force initiatives to ensure that they giveeach customer special attention through customized adverts and communications. All the engagements should be automated to ensure that they are able to scale. These models should be embedded in their sales or customer acquisition tools/systems. The models will however require continuous maintenance, evaluation, and improvement, at least oncea month. To better leverage on the insights on whether a new customer will churn or not upon registration, the company customer acquisition/commercial or sales strategyshould be anchored around data.

Bibliography

[1] Quynh Dang. Customer churn prediction in computer security software. 2019.

[2] James M Kairanga. *Churn prediction in mobile telecommunications industry: Acase study of Safaricom Ltd*. PhD thesis, 2012.

[3] Susan M Keaveney and Madhavan Parthasarathy. Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors. *Journal of the academy of marketing science*, 29(4):374–390, 2001.

[4] Marcin Owczarczuk. Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications*, 37(6):4710–4712, 2010.

[5] Mark NK Saunders, Philip Lewis, Adrian Thornhill, and Alexandra Bristow. Understanding research philosophy and approaches to theory development. 2015.

[6] DongBack Seo, C Ranganathan, and Yair Babad. Two-level model of customer re- tention in the us mobile telecommunications service market. *Telecommunications policy*, 32(3-4):182–196, 2008.

[7] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cher- niack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. Data curation at scale: the data tamer system. In *Cidr*, volume 2013, 2013.

[8] Chih-Ping Wei and I-Tang Chiu. Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, 23(2):103–112, 2002.

[9] Lian Yan, Michael Fassino, and Patrick Baldasare. Predicting customer behavior via calling links. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2555–2560. IEEE, 2005.

[10] Shih Jiun Yin, William F Bosron, Leslie J Magnes, and Ting Kai Li. Human liver al-cohol dehydrogenase: purification and kinetic characterization of the. beta. 2. beta. 2,. beta. 2. beta. 1,. alpha.. beta. 2, and. beta. 2. gamma. 1" oriental" isoenzymes. *Biochemistry*, 23(24):5847–5853, 1984.