Predictive Analytics for customer Retention Using Mobile Phone Data.

Adams Zequi Mohammed

Department of Mathematics and Statistics
University of Limerick

Master Thesis

Supervised by: David J.P O'Sullivan

Submitted: 23rd August, 2021.

Declaration

This thesis is presented in fulfillment of the requirements for the award of Msc. Mathematical Modelling. It is entirely my own work, completed without collaboration with others except my supervisor, David J.P O'Sullivan. Where use has been made of the work of other people it has been fully acknowledged and referenced accordingly.

Signature: _____

Adams Mohammed Zequi

August 2021.

Abstract

Customer churn is perhaps one of the most relevant issues that telecommunication companies face. Churning here simply refers to a customer leaving a company. As service providers, telecommunication companies rely on the loyalty and continued goodwill of customers to stay profitable, i.e., there is a direct relationship between number of customers and company profitability. There is the need then for companies to find a way to front-run this problem by trying to preemptively find customers that are more likely to leave so that these customers could be targeted with retention initiatives. This thesis discusses this phenomenon of customer churn and proposes two machine learning models, Random Forest and Logistic Regression, that will be used to predict customer churn. Model accuracy is measured with some of the most popular performance metrics, i.e., Accuracy (ACC), Balanced Accuracy (BACC), Precision, Area Under Curve (AUC) and Receiver Operating Curve (ROC). We saw that AUC performed higher than ACC and BACC. For ACC, Random Forest outperformed Logistic Regression but underperformed (albeit negligibly) with regards to BACC and AUC.

Contents

List of Figures

List of Tables

Chapter 1

Chapter 1 - Introduction

1.1 Introduction

While the telecommunications industry has long been an oligopoly due to it's high barrier of entry, it's markets have ironically been saturated [RFF+14]. This has created the need for telecommunication firms to focus on finding insights that could help them predict the future decisions of their clients, in this case the probability that the client would leave or stay on the network (customer churn).

Customer churn explicitly refers to the phenomenon of cancellation of service contract by a customer. The rate of customer turnover in a company is called 'churn rate' and is quantitatively categorized using binary variables like '0' and '1', which can represent 'stay' or 'leave' and these are used interchangeably in this thesis.

Customers cancel subscriptions for various reasons. Some of them being lack of satisfaction, expense, change in circumstances etc. Whatever the reasons are, companies have to find those reasons and address them to maintain the customer base as its much more expensive to gain new customers than retain old ones. This problem especially impacts the bottom line of service based companies (example, in the telecommunication sector [HCJ17]) because they deal directly with number of customers at any point in time and therefore have the incentives to make ensure that they either gain more customers than they lose or work on retaining customers. The latter is more plausible in the telecommunications sector as that industry is very competitive in terms

of gaining customer loyalty and improving customer satisfaction [NAR18]. Retainment measures can include aggressive marketing strategies, improvements in service provision, affordable prices and improvement in customer service etc. [Xev05].

Due to the quantity of data produced by any single telecommunication company on the performance of their service and on the different characteristics of individual customers, it is possible to use statistical or machine learning models that consider these variables to reduce the probability that a customer will cancel their contract [ADGD18]. This leads us to discussing previous work on this research topic.

In their paper, Praveen Lalwani et al. [LMCS21] used some of the most popular machine learning algorithms including Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), Adaboost, Xgboost, Catboost and extra tree classifiers.The paper submitted that the best performance scores were achieved by XGBoost and Adaboost classifiers using 5 key accuracy metrics(Precision, Recall, AUC, ACC and F-measure). The Accuracy from their results for RF and LR were 78.04% and 80.45% respectively while their Area Under the Curve is 82% for both. The paper however noted that besides the difficulty in predicting customer churn and genuine customers, machine learning provides the best tools to predict customer churn.

Rodan et al. [RFF$^+$14] have worked on predicting churn rate using negative correlation learning (NCL) and Multilayer Perceptron (MLP) (both feed forward artificial neutral networks) using ensemble learning. They concluded that using MLP ensemble learning via negative correlation learning performed better than other machine learning models.

Kavitha et al.[KKKH20] used XGBoost, Random Forest(RF) and Logisitc Regression (LR) to predict customer churn and noted that RF was the most accurate of the three. Besides the three tree learning classification algorithms used, they also proposed the usefulness of lazy learning artificial intelligence algorithms that track the changing needs and requests of customers to be able to predict the continuous behavioral patterns of customers.

Navid Forhad et al. [FHR14] noted in their paper that to make accurate predictive analytics of churn rate of customers, firms need to mine holistic data of customers covering the complete state of the customer. The paper also noted while determining

if a customer will churn or not is important for retention, it is equally important to determine if the customer provides a net loss or gain to the company if they are retained. This phenomenon is called Life Time Value (LTV) of a customer [HJS04]. It also discusses retention strategies like marketing and the relevant time frame of data used for analysis.

Nabgha Hashmi et al. [HBI13] highlighted some limitations with customer churn prediction in their paper. Data from telecommunication companies are usually imbalanced and suffer from a large number of possible prediction variables. While machine learning classification prediction models result in good predictions when modelling non-linear data sets, they are sometimes not reliable when it comes to huge time series data sets and encounter complications when used the in real world.

Ahmed et al. [AJA19] encountered balancing issues with their customer churn research as the ratio of churn to non-churn was 95:5 so they used tree algorithms like XGBoost, Random Forest, Decision Trees, and Gradient Boosting Method which deals with imbalance issues by under-sampling. The highest AUC levels were achieved by XGBoost, GBM, RF and DT respectively with XGBoost at 93%. When different data sets were used, XGBoost still achieved the highest AUC at 89%. The research also implemented Social Network Analysis (SNA) to increase predictive ability.

Chapter 2

Chapter 2 - The Data

2.1 Exploratory Data Analysis

2.1.1 Size ,Shape and Variables in the Dataset

The data set used for our analysis is a subset of a larger data set simulated by *IBM Cognos Analytics 11.1.3* [tel]. The simulated data set covers telecommunications services in the 3rd quarter for Californian residents. It contains various variables (see Table 2.1) describing customers marital status, gender, phone subscription services etc. The most salient variable is 'Churn' which is a binary variable indicating a customers decision to leave or stay with the fictional telecommunication company. The data set consists of 7032 customers and 22 variables representing customer features.

2.1.2 Data Cleaning and Class Refactoring

The research began with cleaning the data set and refactoring the categorical variables. The variables 'InternetService', 'OnlineBackup' , 'DeviceProtection', 'TechSupport' and 'StreamingTV' were transformed to binary response as the third response 'No internet service' was redundant and was changed to 'No'. The redundant response 'No phone service' was also changed to 'No' in the variable 'MultipleLines'.

The original data set had the response of the 'tenure' (which gives the time that the customer was with the company) variable into months. To make the analysis easier,

Column Name	Category	Data Type
customerID	Alphanumeric	Multiple
gender	Nominal	Binary
SeniorCitizen	Nominal	Binary
Partner	Nominal	Binary
Dependents	Nominal	Binary
tenure	Numerical	Discrete
PhoneService	Nominal	Binary
MultipleLines	Nominal	Binary
InternetService	Nominal	Binary
OnlineBackup	Nominal	Binary
DeviceProtection	Nominal	Binary
TechSupport	Nominal	Binary
StreamingTV	Nominal	Binary
StreamingMovies	Nominal	Binary
Contract	Nominal	Multiple
PaperlessBilling	Nominal	Binary
Payment Method	Nominal	Multiple
MonthlyCharges	Numeric	Continuous
TotalCharges	Numeric	Continous
Churn	Nominal	Binary

Table 2.1: Descriptions of categories of our data set.

'tenure_by_year' was created that delineated the tenure months in years as that is the natural length of contract a customer may have with the company. The variables created are '0-1year', '1-2years', '2-3years', '3-4years' and '>4years'.

The variable 'avg_total_charge' was created to be able to capture the long term costs of a customers staying with the telecommunication firm. With changing circumstances, customers might decide to pause a service, include more services, change service plans etc., therefore this variable was necessary to give us an average monthly

overview of a customer's cost over the tenure of his/her stay on the network. This variable differed from the 'MonthlyCharges' variable as that only captured most recent charges. Average total charges is calculated as:

$$Average\ total\ charge = \frac{total\ charge}{tenure}. \tag{2.1}$$

The next step was to reorder the levels ,i.e., change the reference level of the churn variable from 'No' to 'Yes' as that is our focus factor.To model the data , we partitioned the data into 80:20 splits with 80% to used for training the data and 20% for testing the data. We do this to be able to have two independent data sets to train and fit with.

2.1.3 Descriptive Statistics and Plots

In Figure 2.1, we are provided with a breakdown of proportions of key binary variables in the data set.

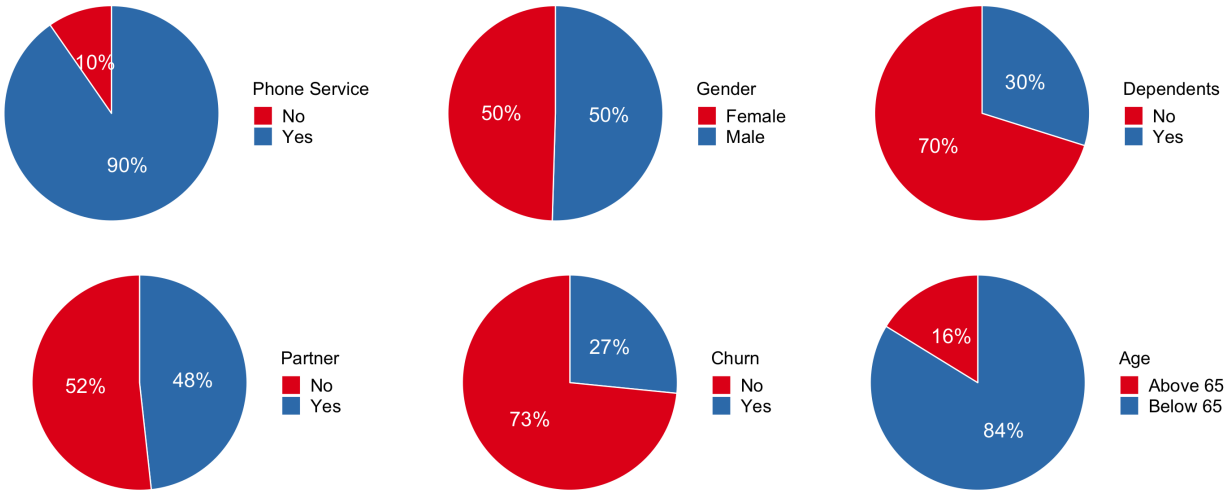

Figure 2.1: Pie plots showing percentages of binary variables in the data set.

Interestingly, we notice that male and females are of equal proportion but their age proportions vary greatly as 'Below 65' are 65 percentage points more than 'Above 65'. While customers with dependants are 30%, customers with partners are 48% compared to customers without partners.

For phone service subscription, customers were largely subscribed at 90%. Non-churners were also overwhelmingly represented at 73%. Due to our goal of trying to predict churners, we can see that the data is imbalanced in favor or non-churners so that tells us to be careful of performance metrics and models used because some error metric will favor classes with higher representative (i.e., the non-churner class). In addition to this, it is the identification of churns that we really care about. So we to use error measure that take this into account.

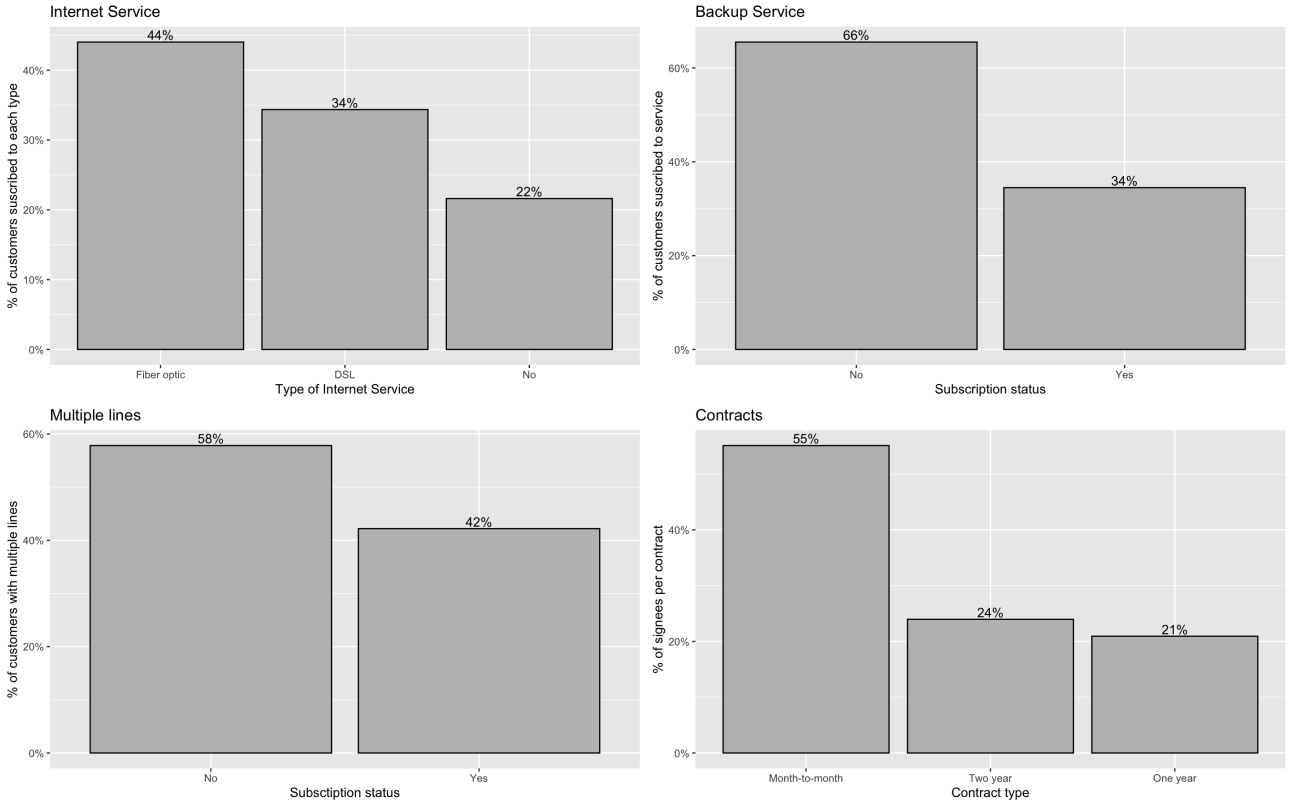

Figure 2.2: Bar graphs showing percentages of customer responses to Internet Service, Backup Service, Multiple Lines and Contract Type.

From Figure 2.2, we see that customers are overwhelmingly subscribed to internet services (78%) with 44% of those preferring fiber optics to DSL and the rest opting not to subscribe to these services. The number of customers that are not subscribed to backup services are approximately twice as many as those subscribed and there are 16% less people subscribed to multiple lines than those that aren't. The number of

people that are on a month to month basis contract are approximately equal to the sum of those on both one year and two year contracts.

A visual analysis of the numerical multivariate variables in data set show that there aren't any large differences between average monthly charges (see Equation 2.1) over the tenure of the individual and the previous monthly charges indicating that customers use the same or slight variation the plan they started with through out their stay with the network. The average monthly charge and most recent monthly charge to maximum number of customers over their tenure on contract are both approximately 20 dollars while the least number of customers were charged approximately 30 dollars. While most of the customers cancelled their contract in the first year, even less cancelled their contract if they made it to the second year. The majority of customers cancelled

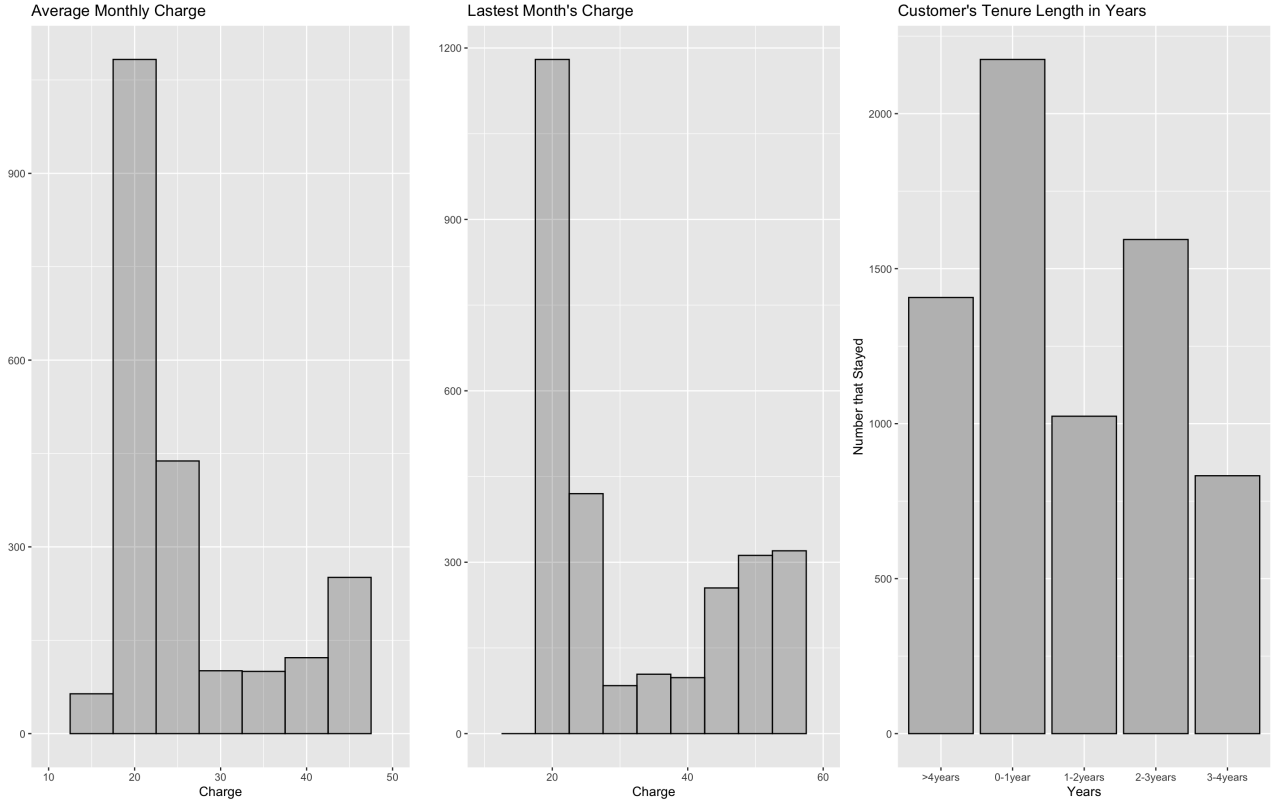

Figure 2.3: Histograms showing breakdown of multivariate numerical variables in the data set.

their contracts between two and three years but if a customer lasted up to the third

year, they tended to stay in the fourth year and beyond.

From Figure 2.4 below, we can see that the below 65 demographic were more likely to churn than the above 65 demographic at 16% and 85% respectively. There wasn't any large difference between customers males and females with regards to churn responses.

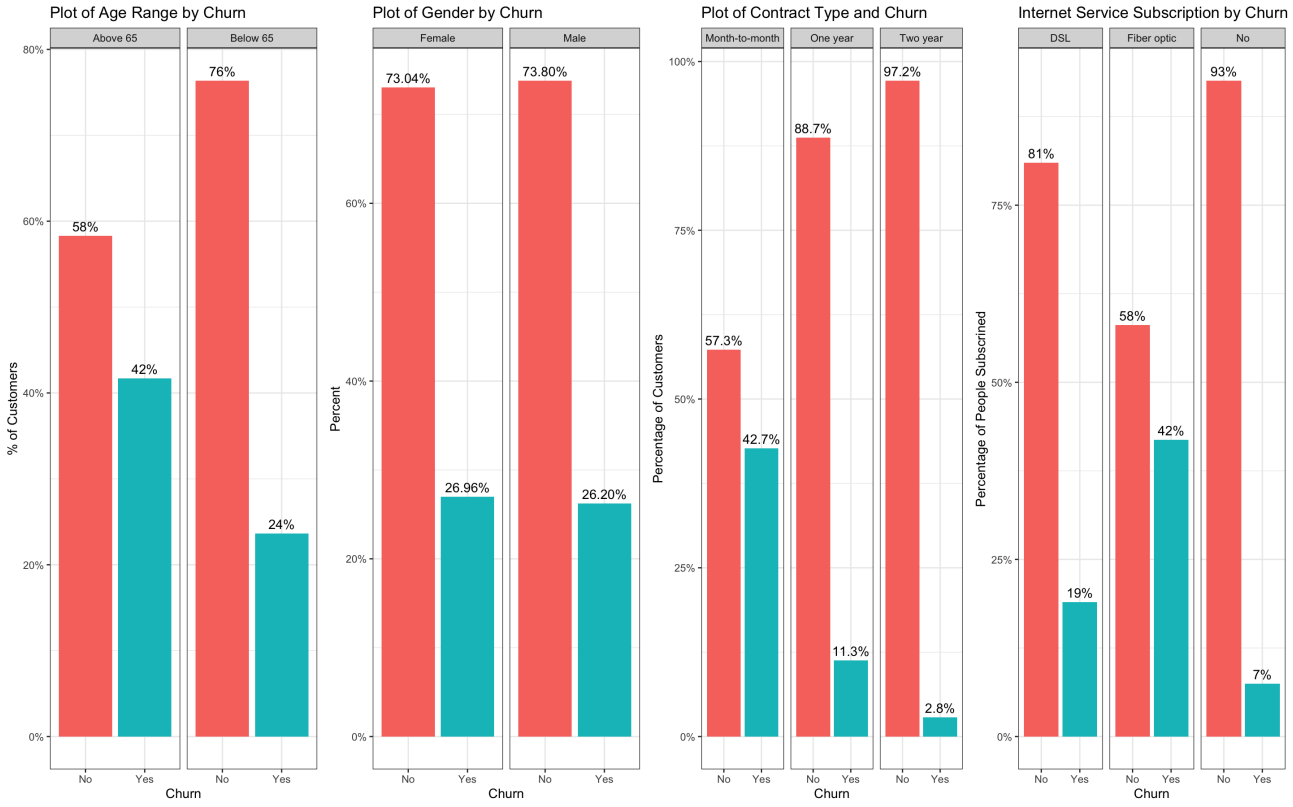

Figure 2.4: Multivariate plots showing 2-way relationships in data set.

Customers that were subscribed to two year contracts tended to stay much longer than customers with other contract types. Regarding the length of stay, 97.2% of customers who signed onto two year contracts stayed on compared to 88.7% and 57.3% for one year and month to month contracts respectively. Customers who weren't subscribed to any form of internet service also overwhelmingly stayed (93%) while customers who were subscribed to fiber optics were twice as likely to leave at 42% compared to customers subscribed to DSL at 19%.

The Figure 2.5 gives us interesting three-way insights into our data set. Younger customers (below 65) had much lower charges and were less likely to churn than their

older counterparts who happen to incur higher charges on average. Older customers who churned however, also incurred lesser charges on average. Regardless of age customers tended to stay much longer than churn but if they churned then older customers stayed longer than younger ones. Non-churners, older customers also stayed longer than younger customers. For customers who had multiple lines,monthly charges for churn-

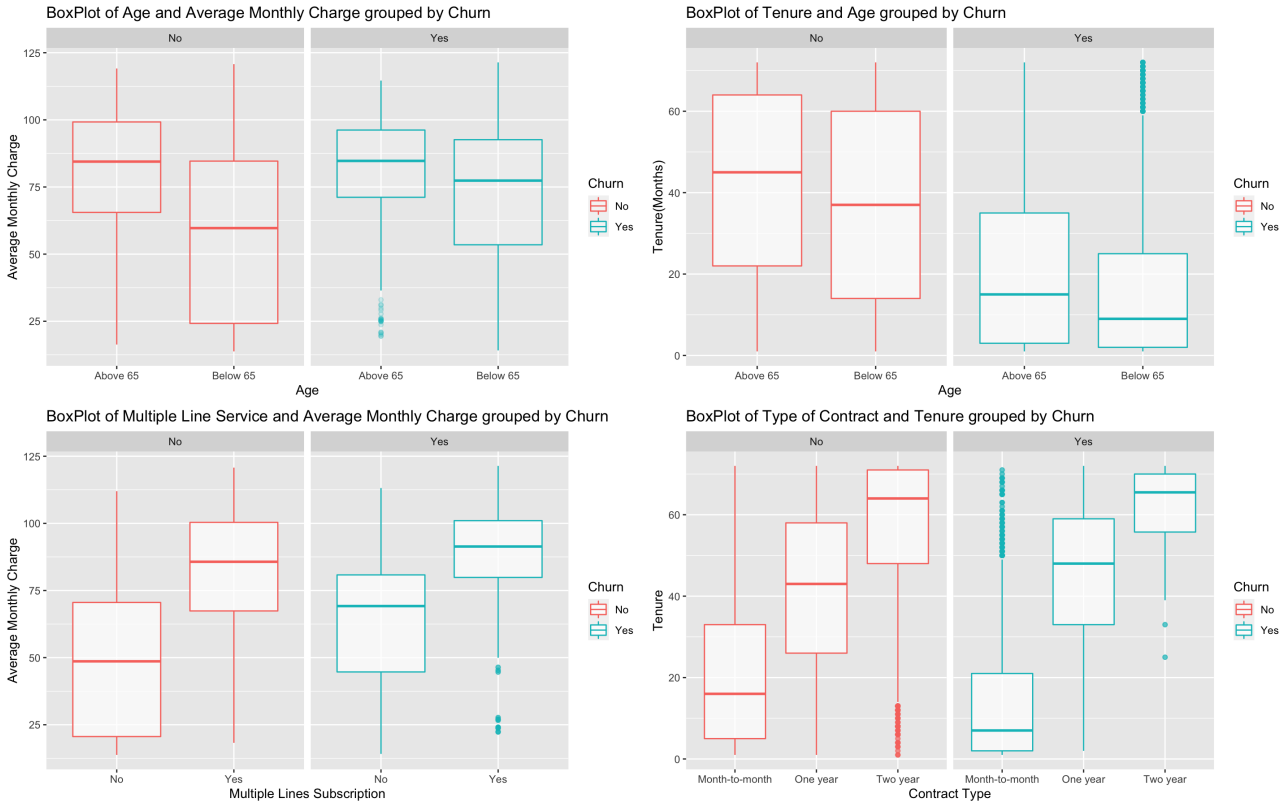

Figure 2.5: Multivariate Plots showing 3-way relationships in the data set.

ers and non-churners were roughly the same. Customers who did not have multiple lines and churned however, had higher charges than those who stayed and had multiple lines. For both churners and non-churners, customers with two years contracts stayed longer than those with one year contracts and they in turn stayed longer than those with month to month contracts.

The next section discusses the summary of everything discussed in this section including some information about what to expect for subsequent sections.

2.2 Summary of Exploratory Data Analysis

This section highlight some characteristics of the data before we move onto the statistical and machine learning methods that we applied to the data in this next chapters.

The raw data was unorganised so we had to do a bit of data cleaning to be able to use the data. We changed redundant categories for certain variables and delineated time variables properly into years.

In the uni-variate analysis of our data sets , we saw that our dependent variable (churn) was imbalanced in favor of non-churners. Customers with phone service were disproportionately more than customer who didn't have phone service. Gender and partner had were fairly proportionately distributed. Customers preferred to use fiber optics and usually signed on month to month contracts. Customers weren't keen on backup services and using multiple lines. Highest charges incurred by the majority of customers were usually approximately 20 dollars. While gender wasn't a significant factor for churn, customers above 65 were much more likely to churn than customers below. Customers that signed on to two year contracts tended to stay much longer than customers on any other type of contracts. Customers who weren't subscribed to internet services stayed much longer than others.

Customers that were subscribed to multiple lines and were non-churners had roughly the same charges as those that churned. Regardless of age, churn customers stayed for fewer months than non-churners.

The key variables to note in the exploratory data analysis were churn, multiple lines, phone service, age, tenure and dependants while gender and partner were fairly benign.

From the exploratory data analysis carried out earlier in this chapter, we will list some key variables that stood out: churn, multiple lines, senior citizen (age), tenure, dependants and internet service.We will ascertain if these variables are consistent with significant variables during our modeling process.

In the next chapter ,we discuss the models used ,how to use them and how to measure their efficacy.

Chapter 3

Chapter 3 - Modeling and Model fit

3.1 Performance and Error Metrics

Our goal in this section is to describe the models that we are going to fit to the data to predict which customers are most likely to churn. All the model will yield a predicted label yes or no that we can compare to whether they actually churn or not. Using this we can measure a given models accuracy, so first, we describe the methods that use to access model fit and then move on to describe the models themselves.

3.1.1 Confusion Matrix

A confusion matrix (also known as a contingency table) is a table that defines the instances of a classification model [Faw06]. For a 2x2 classification model, we intuitively have 2 classes (Actual or True Class and Predicted or Hypothesized Class) and 2 instances (Positive/Yes and Negative/No). A positive instance that is inaccurately classified as negative is termed a false negative. A positive instance that is accurately classified as positive is termed a true positive. A negative instance that is inaccurately classified as positive is termed a false positive. A negative instance that is accurately classified as negative is termed a true negative.

Table 3.1 describes all classes and instances of a 2x2 matrix. The left to right diagonal gives us the accurate predictions while the right to left diagonal shows us the 'confused' predictions. In explaining performance metrics in later chapters, we will

		Predicted Class		TOTAL
		No	Yes	
Actual Class	No	*True Negative (TN)*	*False Positive (FP)*	Actual Negative Numbers
	Yes	*False Negative (FN)*	*True Positive (TP)*	Actual Positive Numbers
TOTAL		Predicted Negative Numbers	Predicted Positive Numbers	Total Customers

Table 3.1: Confusion matrix table detailing actual and predicted class differences.

have to describe briefly the True Positive Rate(alternatively known as TPR, Recall or Sensitivity) and False Positive Rate(alternatively known as FPR or False Alarm Rate) [JST14]. These can be mathematically expressed as:

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

, and

$$FPR = \frac{FP}{FP + TN}. \tag{3.2}$$

Accuracy (ACC), Balanced Accuracy (BACC) and Precision are also frequently used metrics in assessing the performance of model.Accuracy (also known as the ACC) is an evaluation criteria that evaluates how accurately instances of both classes are classified. Its defined by [BV09] :

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.3}$$

The denominator in Equation 3.3 refers to the total number of customers while the numerator refers to the accurately predicted classes of customers.

For balanced data sets, the ACC is a fairly reliable measure however for imbalanced data sets, it has some drawbacks therefore it's biased towards the most frequently occurring class (imbalance data) therefore if the classification model keeps predicting the class with the highest occurrence, the accuracy level will reflect this class and vice

versa. Since real world data sets tend to be imbalanced [Kot13], researchers prefer to use the Balanced Accuracy metric or both BACC and ACC.

While Accuracy measures how accurately both classes have been classified, balanced accuracy measures how accurately each class has been classified [TRMB$^+$18]. It's defined mathematically as:

$$Balanced\ Accuracy = (\frac{TP}{P} + \frac{TN}{N}) \times 0.5. \tag{3.4}$$

where P and N stand for the total number of positives predicted and total number of negatives predicted respectively.We can see from Equation 3.4 that BACC balances out the ACC by segregating positives and negatives, giving equal weight to both classes.

Precision is a measure that tells us how accurate predicted positive instances are relative to total number of predicted positive instances [PA11].

$$\frac{TP}{TP + FP}. \tag{3.5}$$

The Accuracy, Balanced Accuracy and Precision will play different roles in assessing the performance of our models. ACC will tell us how accurately the model as a whole performed while BACC will tell us the unbiased performance of our model relative to each class.Precision tells us how well out model performed relative to our goal of predicting churners, i.e, positive class.

The next chapter discusses another commonly used metric in machine learning, the Receiver Operating Curve (ROC) and Area Under the Curve (AUC).

3.1.2 Receiver Operating Characteristics (ROC) and Area Under Curve (AUC)

The ROC curve plots the true positive rate (see Formula 3.1) against the false positive rate (see Formula 3.2) on the Y and X axis respectively. The ROC curve displays the relationship between false alarms and accurate predictions.

In the ROC graph from Figure 3.1, we are mainly concerned with the upper part of the diagonal line.

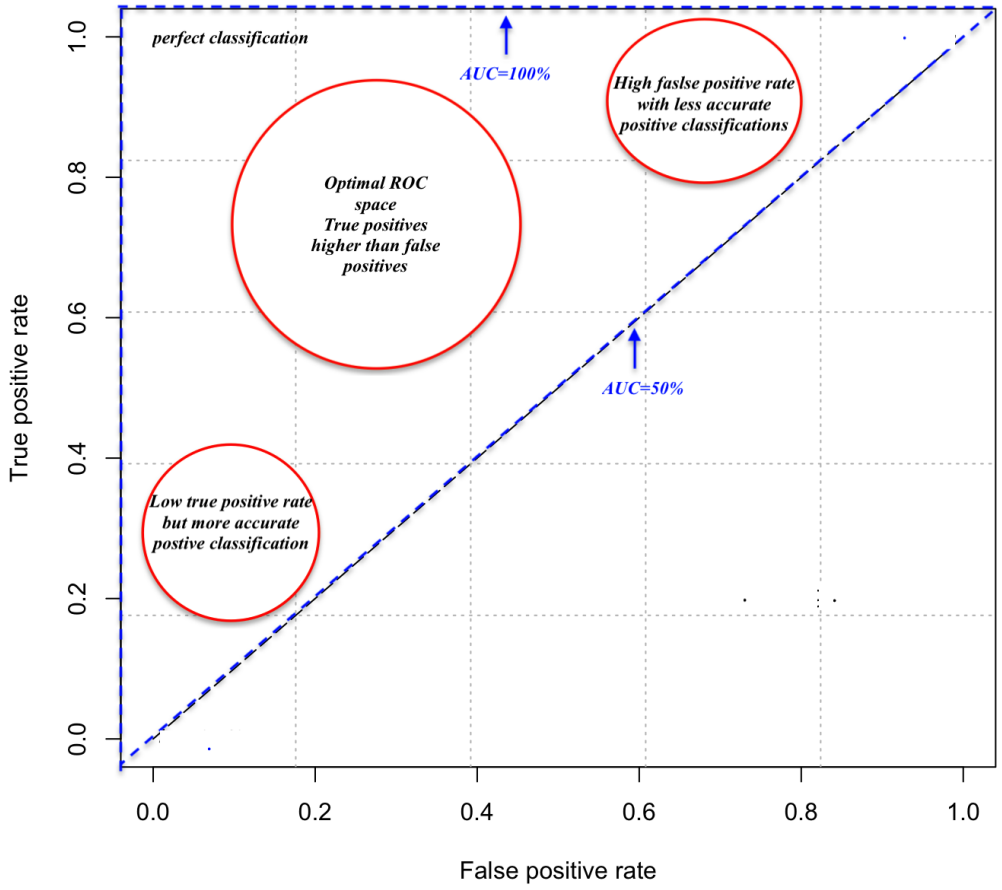

Figure 3.1: The receiver operating characteristics (ROC) space.

The lower part of the diagonal line intuitively across the x-axis indicates negative bias and increasingly higher false positive rates.Negative bias here means that the model predicted more 'No' classes than actual 'No' so that the more bloated the curve is around the (0,0) axis, the more negative bias there is with the model. The opposite is true for the right hand of the upper part of the diagonal. The more bloated the curve is in that region (0,1), the more positively bias our model is, meaning our model predicted more positives than there are actual positives. The upper part of diagonal line can be divided into three parts and described hence. Expanding on our ROC space, To the left, we see that while the true positive is low, there is more accurate

predictions of positives as its a negative bias zone. Typically the optimal point of the upper-diagonal space is where we get a fairly accurate representation of positive and negative classification so that true positive rate is balanced with positive rate. The third space shows us that while positive classifications are loosely made we can see, there is also a high false positive rate.

The Area Under Curve (AUC) describes the area under the ROC curve. For a binary classification problem, the AUC is expressed as is a plot of the True positive rate (TPR) against the False positive rate (FPR).Ideally the larger the area under the curve, the higher the AUC. Looking at the triangle above the diagonal line from Figure 3.1, the diagonal line (also called chance diagonal) gives an AUC of 50% [DvS+07] which is essentially the area under the graph. An AUC equal to that of the chance diagonal means that the the predictive ability of the model is purely by chance, like a coin toss making it useless. If we atain perfect discriminator between the two classes, regardless of the threshold applied, the area will be 1 which corresponds to a TPR and FPR of 100% each.

3.2 Logistic Regression

The initial classification algorithm we use is the logistic regression (LR). It's used to establish relationships between a (or multiple) predictors and outcome variable. It is an extension of the popular ordinary linear squares regression (OLS) and is popularly used in social sciences research for the means of binary classification [cel].

The main difference between LR and OLS is their outputs. While the OLS has outputs a continuous value, the LR's output is discrete. A continuous value, like the name implies can be any value within a range of values example, marks in an exam, height of individuals etc. A discrete value is fixed example, the number of players in a football team (because there cant be half a player in a football team).

As mentioned earlier, the logistic regression is an extension on the OLS model. More accurately, the OLS is passed through the sigmoid function (a classification function that results in binary values of 0 and 1) to get the logistic function. Keeping this in mind, we intuitively realise that this process can alternatively be understood as

converting continuous variables (from OLS) to continuous probability between 0 and 1(LR), i.e., probability of an event happening (p) or not ($1-p$). We will prove this by computing the probability and odds of an event happening and comparing that to our final logistic function.

The odds are calculated from probability estimates, i.e., If the probability of an event happening is p , then the odds of that event happening is:

$$odds = \frac{p\,(probability\,of\,event\,happening)}{1-p\,(probability\,of\,event\,not\,happening)}. \tag{3.6}$$

Before we summarise our logistic regression process , lets state the sigmoid function :

$$sigmoid\,function(s) = \frac{1}{1+e^{-y}}. \tag{3.7}$$

The three steps below summarise the logistic function creation process :

- Given an OLS function, $y = mx + c$.
- y is passed through sigmoid function (see Equation 3.7).
- Resulting equation =
$$In(\frac{s}{1-s}) = mx + c.$$

On the left hand side of the equation, we see a similar equation to our odds ratio (see Equation 3.6). To test our logistic function, if we have a slope and y-intercept of a regression function are 1 and 0 respectively, plotting a range of x-values between -5 and 5 show an S-shaped (see Figure 3.2) curve representing the sigmoid curve.

The next chapter discusses how the model for the results of logistic regression should be interpreted and how to tune the models using step-wise regression.

The next section discusses step-wise regression and why and how its important to improve the logistic regression model using step-wise regression.

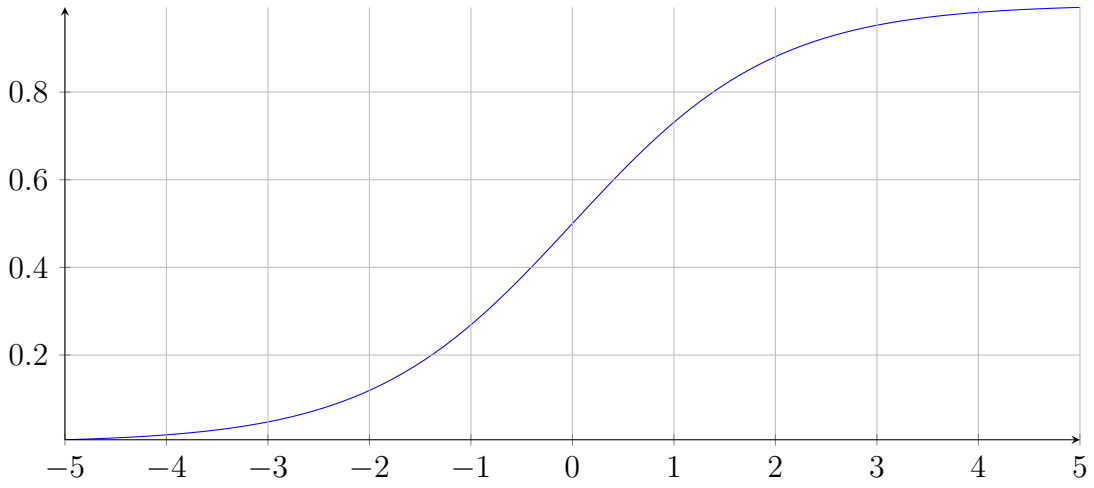

Figure 3.2: Sigmoid function of range of x-values between -5 and 5.

3.2.1 Improving a logistic regression model

Our first logistic regression model will contain all the variables discussed in the logisitc regression section. However, all of these are unlikely to be predictive of churn. To isolate the important varaible we will fit a second LR model with few variable and compare its accuracy to the first and that of the Random forest model. We will do this variable selection processes via foward stepwise.

The step-wise regression helps us make that decision by using the model building process. It us AIC as a measure of model fitness. It initially does a multiple regression of all the predictors and picks the best predictors. The problem with this multiple regression process is there is high collinearity among the predictors. The step-wise regression solves this by also undertaking univariate regressions and selecting the significant predictors.After presenting our significant variables, we can gain insights into how to deal with variables that are deemed important by calculating our accuracy metrics on them and comparing the figures obtained.

3.3 Random Forest

Random forest (RF) is a classification algorithm that uses both decision trees and ensemble learning for it's learning and prediction purposes. The decision trees, as the

name suggest are used to create the trees and the ensemble learning combines them to create the forest, more accurately, it is a combination of weak learners ,i.e., aggregating the single trees and creating a prediction on the basis of majority vote.

This classification of 'trees' are built by randomly selecting data that is then voted on so that the majority voted is categorized into a particular class. The process behind the model is thus: There is a bagging or bootstrapping process that is done with replacement. The out-of-bag sample is constructed from non-bagged data. The bagged or bootstrapped sample is used to create a tree. Contrary to normal classification tree method, a sample X is randomly selected from our variables. X is then used to select the best predictors. Due to the random sampling, overfitting is avoided. Also, the process tends to be uncorrelated as a result of the randomization process. This is key because strong variables are not over-selected and the forest doesn't have an information-bias (giving us the same information). This same process is carried out repetitively until we get our desired trees in the forest. The probability for each class is then the average votes it got.

The algorithm then checks errors of each tree using the variables that weren't chosen in the bagging process. This is called out-of-bag (OOB) errors. The errors are checked by iteratively reducing the gini node impurity anytime a predictor is split so that the lowest impurity variable is listed as most important.

The algorithm then lists the most important features based on how small the errors of that feature is albeit care must be taken when interpreting variable importance scores as shown in.

Random forest is particularly favored because, as stated above, it results in correlations which keep errors isolated to trees. It can handle a huge number of variables without requiring the user to do variable selection.

With respect to the accuracy metrics, area under curve (AUC), random forest outperforms most classification machine learning algorithms with the exception of gradient boosting (GDBT) where accuracy between RF and GDBT converge as number of data sets tested increases [ZLZA17].

3.3.1 Improving a Random Forest model.

Like the logistic regression model, there is a way to improve the efficacy of a random forest model. This process is called 'parameter tuning' of a random forest model. During model tuning, we are concerned with 'mtry' and 'ntrees'.

The 'ntree' parameter specifies the number of trees we want our model to grow. Ideally like in nature, the larger the number of trees, the denser the forest , hence the more computationally demanding our model is. Depending on our hardware setup, its optimal to find a balance between our specified number of trees and our setup.

The 'mtry' parameter specifies how many nodes should be split after the initial training process. As a rule of thumb, we should avoid using smaller values of node to avoid over fitting.

Specifying these parameters carefully usually results in a much better model than our inital random forest model.

The next chapter discusses the results from our both our models.We assess both the training and data set in this regard.We also present the performance metrics and compare them across models for all both training and test datasets.

Chapter 4

Chapter 4 - Results

4.1 Logistic Regression

We began the result analysis with our logistic regression model. Here we initially model all 22 variables (see Table 2.1) with logistic regression. We do this initially to be able to tell which variables were significant and which weren't during our step wise model and to also be able to compare the results of a model populated by all variables and models populated by variables deemed important by the step-wise. The function used, generalized linear model (glm), is the 'stats' package in R [cra].

The model we run above outputs each feature, its corresponding coefficients, standard error, z-value and p-test results.

For the results presented in Figure 4.1, we focus on the columns or rows indicated in red. The coefficient estimates (first column in red) tell us two things, the direction and magnitude. The direction is indicated by the sign and magnitude by the number. The significance estimates (last column in red) tells us two things too, i.e., if the variable is significant and which significant level (alpha) value used. The significance keys (row in red) corresponding to different alpha levels are displayed in the row below.

Interpreting a categorical variable like PaperBillingYes will be: The positive sign indicates that, all things being equal, customers who were billed with paper were more like to churn than customers who didn't. Besides that, the effect of a customer being billed by paper versus by other means has more than twice an effect on the model

Results from our logistic regression model

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.209870	0.300161	-7.362	1.81e-13	***
PaperlessBillingYes	0.360656	0.082374	4.378	1.20e-05	***
OnlineSecurityYes	-0.450670	0.095086	-4.740	2.14e-06	***
MultipleLinesYes	0.165301	0.090521	1.826	0.067835	.
genderMale	-0.039640	0.072022	-0.550	0.582055	
DependentsYes	-0.082967	0.099861	-0.831	0.406071	
SeniorCitizenBelow 65	-0.301990	0.093156	-3.242	0.001188	**
MonthlyCharges	0.006629	0.012266	0.540	0.588895	
avg_total_charge	-0.002357	0.011666	-0.202	0.839871	
tenure_by_year0-1year	1.767704	0.183477	9.634	< 2e-16	***
tenure_by_year1-2years	0.911414	0.181164	5.031	4.88e-07	***
tenure_by_year2-3years	0.504285	0.167218	3.016	0.002564	**
tenure_by_year3-4years	0.350482	0.182522	1.920	0.054830	.
InternetServiceFiber optic	0.695122	0.151075	4.601	4.20e-06	***
InternetServiceNo	-0.837966	0.175356	-4.779	1.76e-06	***
PaymentMethodCredit card (automatic)	0.014661	0.125045	0.117	0.906668	
PaymentMethodElectronic check	0.368496	0.103602	3.557	0.000375	***
PaymentMethodMailed check	-0.106403	0.126140	-0.844	0.398932	
PartnerYes	-0.038667	0.085971	-0.450	0.652881	
ContractOne year	-0.764297	0.118166	-6.468	9.93e-11	***
ContractTwo year	-1.750631	0.200375	-8.737	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 4.1: Results from our logistic regression model.

than a customer having multiple lines(0.36 to 0.16). Another interpretation of the magnitude would be, the relative odds of churn for someone with paperless billing to someone with paperless billing is 0.36. The corresponding significant value estimate also shows that PaperlessBillingYes is significant at an alpha value of 0.001 (which makes it automatically significant at alpha values $<$ 0.05(our chosen alpha level)). Interpreting numerical variables like average total charges will be: All else being equal, customers with higher average monthly charges are less likely to have churned.

We can see that PaperlessBillingYes, OnlineSecurityYes, SeniorCitizenBelow65, Tenure, InternetServiceFiber optic, InternetServiceNo, PaymentMethodElectronic Check, Contract are all significant at our alpha level of 0.05 which is consitient with the finidngs from our exploratory data analysis.

The next step after analysing our variables is to create our predictions. This process is undertaken using the predict function from the same package 'stats' used for the modeling. The prediction gives us predicted values of each customer. To be able to determine when a positive class is predicted, we need to specify a value that will be used for the predicted probability. This value is the cutoff point or threshold.

Since we would like to be able to identity an optimal threshold, we do this by exploring two criteria that we will try to maximize on the training data. We improvised our algorithm to be able to cover a range of values(see Algorithm 1) and used a range of values from 0.1 to 0.8 in ranges of 0.01 as potential thresholds.

Algorithm 1 Deriving Confusion Matrix and Accuracy Metric of models across thresholds

modeled_data =logistic function(*dataset*)
thresholds_list =range of *thresholds*
for each *threshold* in *thresholds_list* **do**
  predicted_probability =predict(*modeled_data*)
  if *predicted_probability* > *threshold* **then**
    classify as *churner*
  else
    classify as *non churner*
  end if
  predicted_class = dataframe of *churners* and *non churners*
  confusion_matrix =table of *predicted* and *actual class*
  accuracy_level =accuracy level calculated from *confusion_matrix*
  data_frame =data frame of *threshold* and *accuracy_level*
end for

As stated earlier, each threshold determines a predicted positive and that information is used to create a confusion matrix (see Figure 3.1). The optimal threshold selected is the threshold that provides the confusion matrix that maximises measured accuracy. The next steps are to extract our performance and error metrics for the confusion matrix(see Algorithm 1).

This introduction detailed the process of running our model, predicting probabilities, determining positive and probabilities and creating our confusion matrix. The next section discusses the results obtained using this process.

4.1.1 Model Results

We begin analysing our results by comparing the churn data set with our model predicted probabilities. A density plot is used for this analysis. From Figure 4.2 , we can see the probability density plot of the training set which shows the distribution of predicted probabilities from our model.

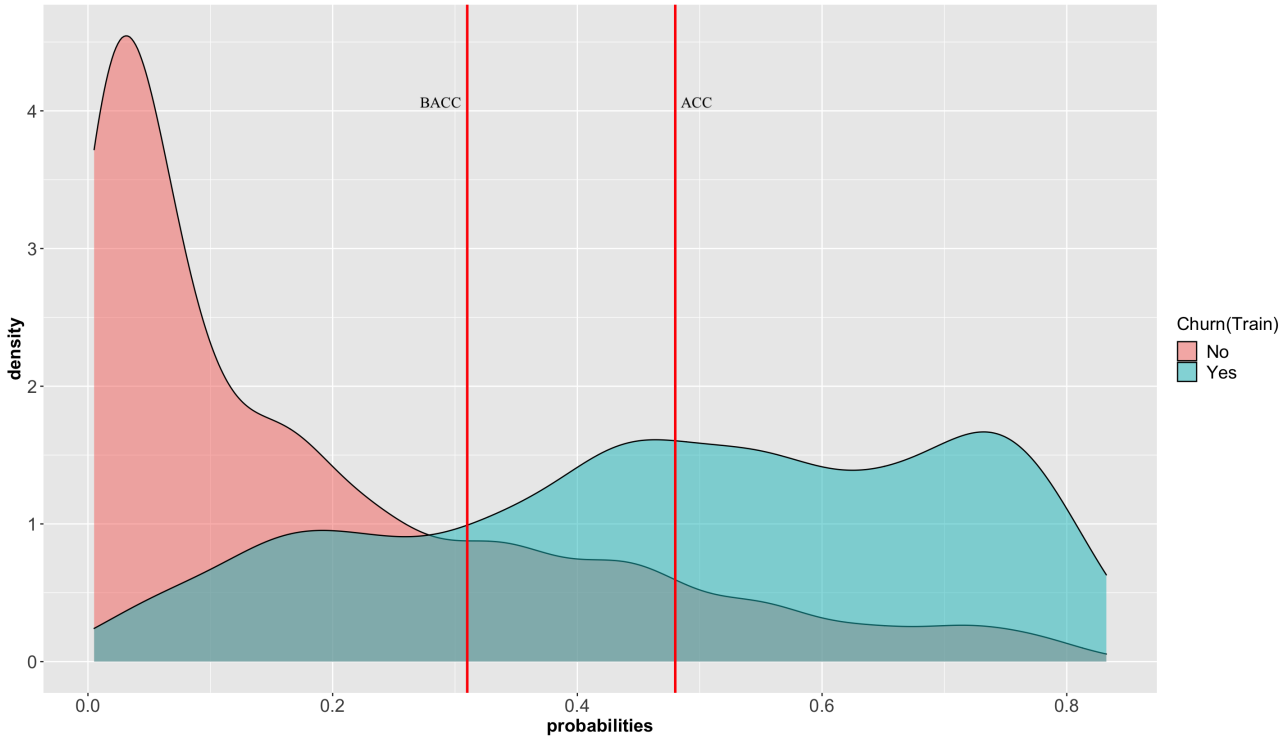

Figure 4.2: Probability density plot of churn grouped by actual training data.

While customers have higher probability of churning, most of the predicted customers are non-churners (which is understandable due to the imbalance of our data set) with a density greater than 4. The number of churners grow however with increasing probability while non-churners reduces with increasing probability. The same phenomenon occurs with our test data set with the slight difference of density of churn-

ers being a little lower (Appendix .1). The vertical lines show our ACC and BACC which will be explained in later paragraphs.

Before we move to our confusion matrix, let's look at an overview of all the accuracies we considered in this study using a range of thresholds. Figure 4.3 plots a series of thresholds of our training set (0.1 to 0.8) in equal ranges of 0.01. BACC starts off higher than Accuracy, at approximately 62%, peaks at around 73% but falls steeply to around 52% as soon as it hits peak, Accuracy starts off lower, at around 60% peaks at 79% and ends at 74%. We see a similar trend for the test data set(see Appendix .2). While there is a higher ACC for the test data for performance at a different threshold, our current accuracy is not far behind.

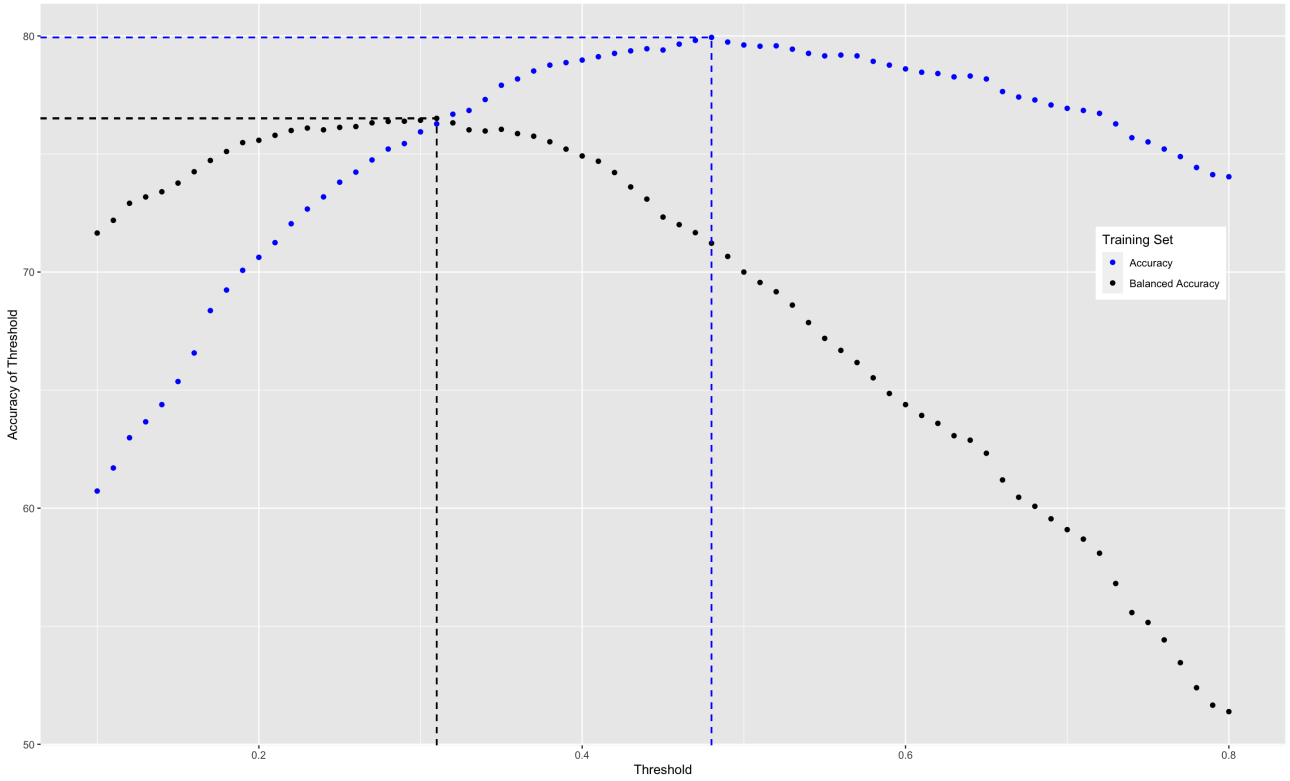

Figure 4.3: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (Training Set).

Table 4.1 displays the 2x2 confusion matrix of accuracy and balanced accuracy for both our training and test data sets (see Appendix .2). We can see immediately

that the false positives for the balanced accuracy is twice the size of false positives for accuracy for both training and test data sets. This is by default the side effect of using balanced accuracy on imbalanced data set such as ours .We can confirm that quantitatively by looking at our true positive rate(TPR) and false positive rate(FPR). With regards to ACC, the two metrics, TPR and FPR for our training sets are 0.52 , 0.10 and for our test sets, they are, 0.56 and 0.10. To expand on the TPR and FPR of our training set, when the actual class is a 'Yes', the probability that our model predicts a 'Yes' is 0.52 and when the actual class is 'No' the probability that our model predicts a 'Yes' is 0.10. Even though our TPR isn't necessarily good, our model is relatively good at predicting our true instances accurately. For BACC, the TPR and FPR for our training and test data sets are 0.77, 0.23, 074 and 0.21. Immediately we can see that both the TPR and FPR for our balanced accuracy is much higher than that of the accuracy metric. While the TPR and FPR are good positive measures, especially for our goals in this paper (predicting churners), ACC and BACC give more encompassing measures by including both classes.

	Accuracy				Bal. Accuracy		
	Predicted Class			Total	Predicted Class		Total
		No	Yes		No	Yes	
No	Training Set	*3711*	*420*	*4131*	*3140*	*991*	*4131*
Yes	Training Set	*709*	*787*	*1496*	*344*	*1152*	*1496*
	Total	*4420*	*1207*	*5627*	*3484*	*2143*	*5627*

Table 4.1: Confusion matrix of optimal cutoff point.

		Optimal Threshold	Accuracy Level	Precision
Accuracy	Training Set	*0.48*	*79.4%*	*65.9%*
	Test Set	*0.48*	*80.7%*	*66.1%*
Balanced Accuracy	Training Set	*0.31*	*76.5%*	*53.7%*
	Test Set	*0.31*	*76.4%*	*55.7%*

Table 4.2: Performance and Precision of logistic regression model.

Looking at Table 4.2, we can see that the Accuracy(ACC) metric performs better

with better precision (the model is correct approximately 66% of the time it predicts a 'Yes' for compared to 54% for BACC) than Balanced Accuracy for both training and test data sets. This is expected because, as discussed earlier, while ACC measures performance of both classes, BACC essentially measures the average of each class hence has a 'balancing' effect. The optimal thresholds for ACC are also higher than BACC. The differences in accuracy level for both training and test data sets are negligible even though the test data's ACC accuracy level is 'quantifiably' higher than the training data and vice versa for BACC.

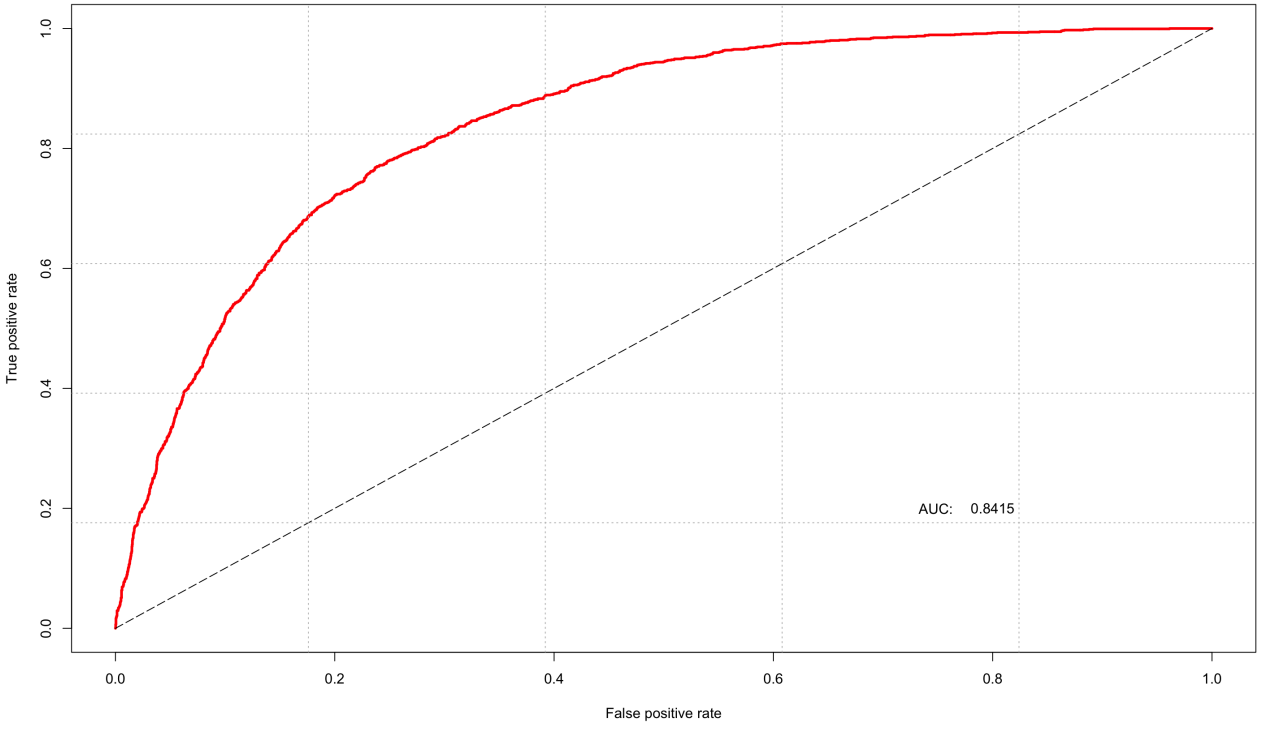

Figure 4.4: ROC curve and AUC of training set

Area	*0.9-1.0*	*0.8-0.9*	*0.7-0.8*	*0.6-0.7*	*0.6-0.5*	*>0.5*
Diagnostic Curve	Excellent	Very Good	Good	Sufficient	Bad	Test not useful

Table 4.3: Interpretation of Area Under Curve in ROC space.

The next metric to discuss will be the ROC and AUC.To discuss that we will need

to introduce the concept of Yould's Index.This is a table that matches the given area of any curve to an accuracy diagnostic [UIUA12]. Table 4.3 gives us an insight into how to interpret the area calculated under the curve in an ROC space.

The ROC curves and AUCs of training and test data set in Figure 4.4 and Appendix .3 show us the ROC space and its calculated area. Using the Yould's index we can see that the AUC's of both plots are 'very good'.Visually , we also see that most of our values are concentrated around the optimal ROC space (see Figure 3.1) and there's essentially no difference between the performance our test and training sets.

The next section discusses the results obtained from our step-wise logistic regression model and the insights we got from them.

4.1.2 Step-wise LR Model Results

We begin analysis of our step-wise model by first comparing the variables that were highlighted by our step-wise model to our logistic regression model and EDA. From Figure 4.5, we can see that variables like multiple lines, senior citizen, internet service and tenure are present in both our exploratory data analysis and logistic regression model which indicates that they have high predictive ability of churn.

Looking at the confusion matrix(see Table 4.4 and Appendix **??**) for optimal cut-off points for both Accuracy(ACC) and Balanced Accuracy metrics(BACC), we see a similar phenomenon with the initial logistic regression model .We see that the false positives for BACC are significantly more (sometimes twice) those of the Accuracy metric across both training and test sets and that is again, a testament to the balancing effect of the metric. The TPR and FPR for ACC across the training set are 0.53,0.10 respectively while the BACC's TPR and FPR stand at 0.77 and 0.24.

We can clearly see that the BACC performs much better than the ACC in terms of predicting positives correctly (24 percentage points more) but we must be wary that both these measures are only concerned with positives. The density plots of the step-wise models also tell a similar story to the initial logistic regression model (see Appendix .4 and .5). While probabilities of churn are spread across with higher probabilities mostly associated with churn, non-churners are much denser hence more

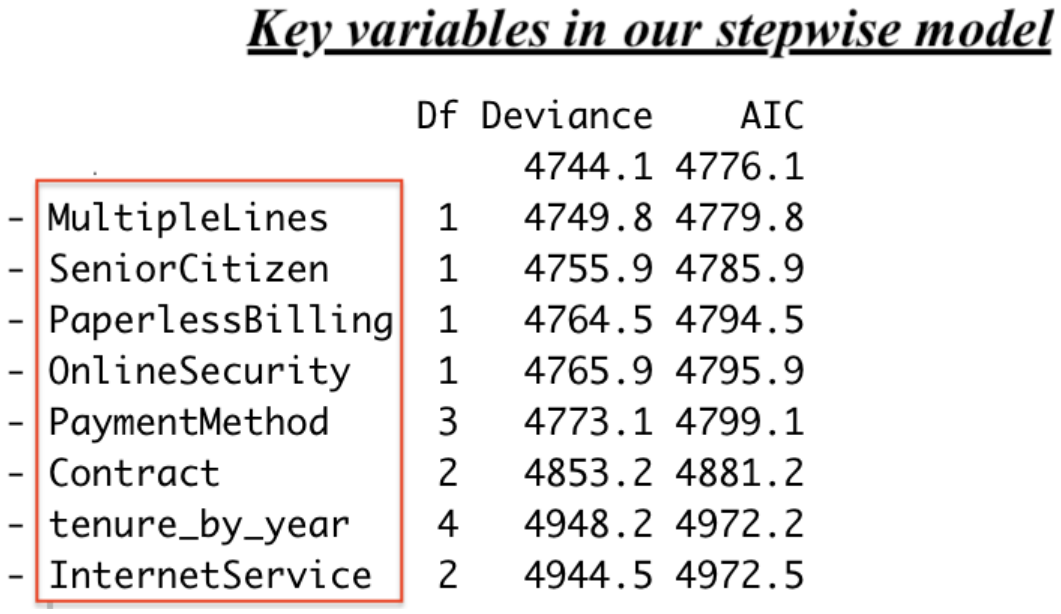

Figure 4.5: Results of our step-wise model.

	Accuracy				Bal. Accuracy		
	Predicted Class			Total	Predicted Class		Total
		No	Yes		No	Yes	
No	Training Set	*3688*	*443*	*4131*	*3105*	*1026*	*4131*
Yes	Training Set	*695*	*801*	*1496*	*332*	*1164*	*1496*
Total		*4383*	*1244*	*5627*	*3437*	*2190*	*5627*

Table 4.4: Confusion matrix of optimal cutoff point.

non-churners are predicted than churners. This is consistent with the structure of the original data itself. The vertical line indicating BACC and ACC on the density plot also confirm the balancing effect of BACC as it moves further to the left where the are more concentrated number of non-churners (as evidenced by our dataset) and less concentrated number of churners (as also evidenced by our dataset).

The Table 4.5 gives the figures for performance and precision of both ACC and BACC for step-wise logistic regression model. Again while the test data sets performed negligibly better than the training for ACC, they performed negligibly less than for

		Optimal Threshold	Accuracy Level	Precision
Accuracy	Training Set	*0.48*	*79.7%*	*64.3%*
	Test Set	*0.48*	*79.8%*	*64.3%*
Balanced Accuracy	Training Set	*0.30*	*76.4%*	*53.1%*
	Test Set	*0.30*	*75.4%*	*52.8%*

Table 4.5: Performance and Precision of Step-wise logistic regression model.

BACC. The precision for ACC is also higher than that of BACC meaning that when the model predicts yes, its correct 64% of the time compared to 54% for BACC.Optimal threshold for ACC is also higher than BACC due to the balancing effect.

It is also apparent that there aren't any significant differences between the step-wise and original logistic regression model where in some cases, the logistic regression model performed negligibly better than the step-wise. We can reasonably conclude that the variables (example, partner, dependants, monthly charge, gender and average total charge) that were excluded due to the step-wise model had no(or negligible) effect on the performance of the model and certainly did not improve it. Ideally the step-wise model did a good job considering that it reduced the number of variables and potentially decreased run time but didn't improve the model significantly.

When the BACC and ACC were plotted across a range of thresholds, the plot was similar to that of our original logistic regression model (see Appendix .6 and Appendix .7). While ACC for balanced data starts at higher performance points for similar thresholds, it quickly peaks and declines while ACC endures and peaks higher with slower declines to levels higher than ACC. The test set also has better BACC points at different thresholds than the optimal threshold chosen.

From Figure 4.6 shows us that the area under the curve in the ROC space for our step-wise test set(see Appendix .8) is higher than that of our training set. Albeit being roughly equal to the AUCs for our original data set , the original AUC's are slightly(or negligibly) higher.This confirms the assertion from our previous performance metrics that while the step-wise model did not improve performance of our official model , it reduce the number of feature variables which might help increase run time for the same results.The performance refernce from the Yould's index in Table 4.3 also tell us that

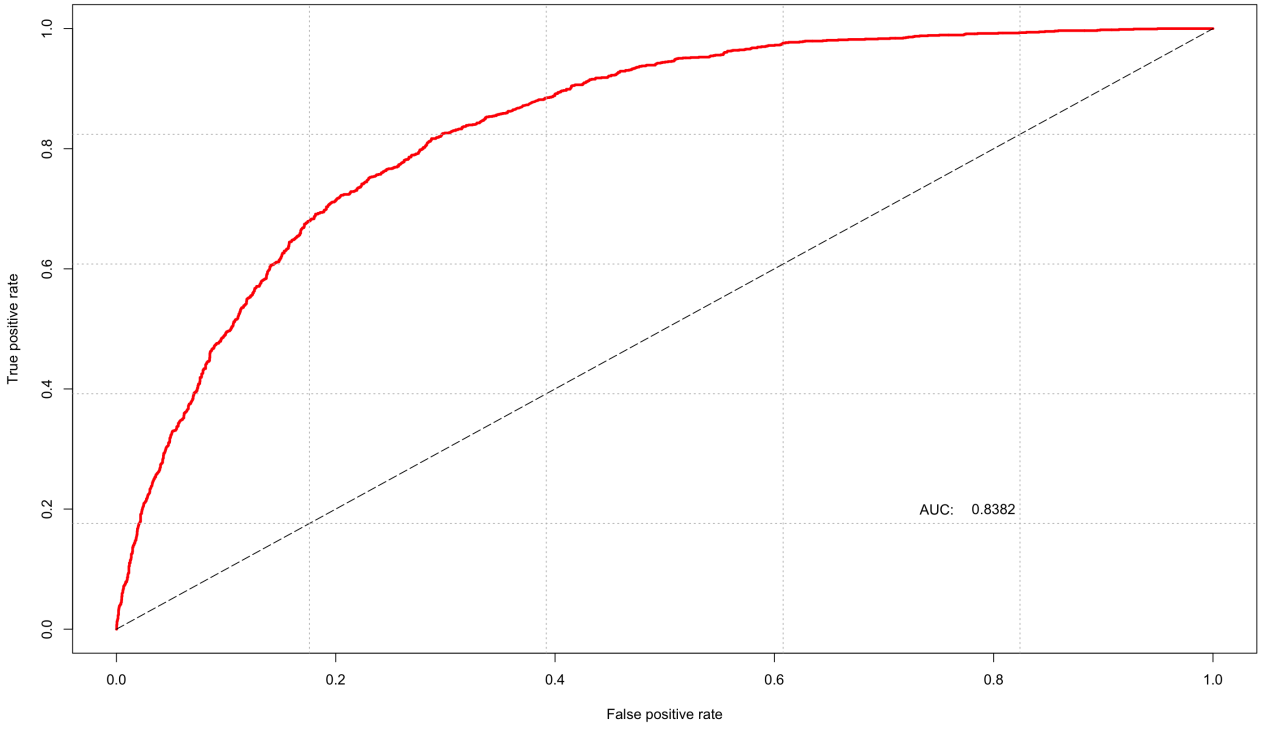

Figure 4.6: ROC curve and AUC of training set

the our model performed 'very good'.

The next section discusses our random forest model and its results. We will explore the same metrics measured for our logistic and step-wise regression.

4.2 Random Forest

4.2.1 Model Results

This section discusses the results of the random forest model used for this paper. We begin, as usual, by presenting the confusion matrices of training (see Table 4.6) and test data (see Table .14) to assess the model's performance on unseen data and discuss some metrics associated.

We see from Table 4.6 that on the 'Yes' column of our predicted class, there are lot more predictions for BACC than ACC and that's consistent with the logistic regression

	Accuracy				Bal. Accuracy		
	Predicted Class			Total	Predicted Class		Total
		No	Yes		No	Yes	
No	Training Set	*3717*	*414*	*4131*	*3182*	*949*	*4131*
Yes	Training Set	*702*	*794*	*1496*	*378*	*1118*	*1496*
	Total	*4419*	*1208*	*5627*	*3560*	*2067*	*5627*

Table 4.6: Confusion matrix of optimal cutoff point.

and step-wise model. The TPR and FPR for our training set is similar to that of our logistic and step-wise model at 0.53 and 0.10 respectively, meaning that when the random forest model does predict a 'Yes', its right 53% of the time and when it predicts a 'Yes', it is actually a 'No' 10% of the time (see Equation 3.1 and Equation 3.2). While the TPR for BACC was much higher, its FPR was higher too at 0.74 and 0.22. These values are similar to both the logistic and step-wise logistic model.

		Optimal Threshold	Accuracy Level	Precision
Accuracy	Training Set	*0.48 and 0.50*	*80.1%*	*65.7%*
	Test Set	*0.48 and 0.50*	*80.0%(higher)*	*67.0%*
Balanced Accuracy	Training Set	*0.30*	*75.87%*	*54.0%*
	Test Set	*0.30*	*73.9%*	*51.6%*

Table 4.7: Performance and Precision of Random Forest model.

The probability density plot of the random forest model also has similar distribution and behaves the same way as our initial logistic regression and step-wise model (see Appendix .9 and Appendix .10).

Table 4.7 gives the performance and precision of the random forest model(RF). We can immediately see that for ACC, RF performed better with higher precision (some of the time) than logistic regression (LR) and step-wise however it under-performed LR and SW for BACC. For the training data set, we see that there are two optimal thresholds but after applying both to our model, 0.50 provided higher accuracy there we settles for that.

An analysis of different points thresholds plotted against accuracy levels show that

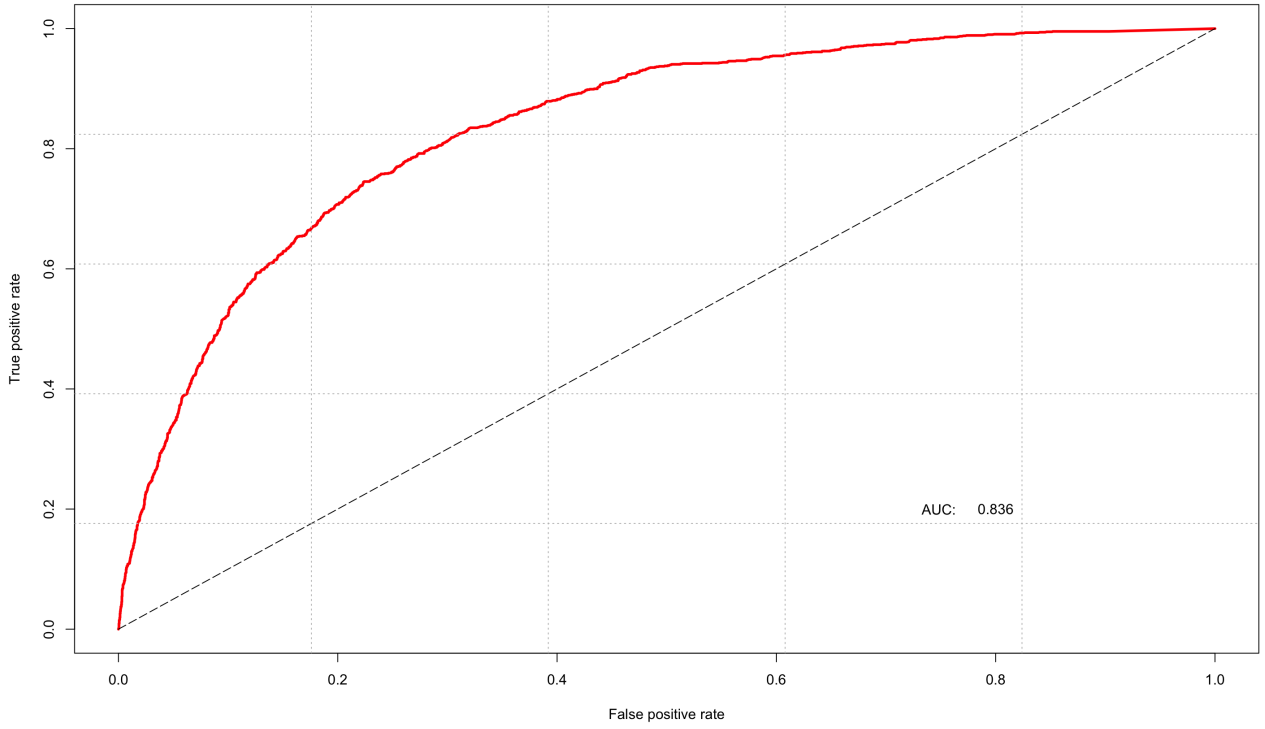

Figure 4.7: ROC curve and AUC of training set

the scatter plot behaves that same way as those for the LR and SW model. See Appendix .11 and Appendix .12 for more information. We also see that the BACC line goes to the left of ACC line which, as stated earlier, is a testament to the balancing effect of BACC.

Figure 4.7 displays the ROC space of our random forest model. we can see that it outperforms the ACC and BACC of both our LR and SW model. The area under the curve in the ROC space can also be interpreted to be 'very good' using the Yould's index(see Table 4.3 for interpretation hints).

The next section discusses our final model which is the tuned RF model. This model is tweaks the parameters of our random forest model to produce a more accurate estimate od our variables.

4.2.2 Tuned RF Model Results

We begin by running our training data through the RFtuned function in R [tun]. The results of this process show us that an 'mtry' of 2 has an OOB error of 0.19 which is the lowest therefore we tune our random forest 'mtry' parameter to 2.

To analyse the results from our tuned RF model and determine if it does better than our original RF model (and our other models), lets analyse the confusion matrix on Table 4.8 for our training set (see Appendix .17 for confusion matrix of our test set).

	Accuracy				Bal. Accuracy		
	Predicted Class			Total	Predicted Class		Total
		No	Yes		No	Yes	
No	Training Set	*4066*	*65*	*4131*	*2979*	*1152*	*4131*
Yes	Training Set	*1234*	*262*	*1496*	*307*	*1189*	*1496*
	Total	*5300*	*327*	*5627*	*3286*	*2341*	*5627*

Table 4.8: Confusion matrix of optimal cutoff point.

The TPR and FPR of ACC of our tuned RF model are very different from what we've seen so far.Even though the model makes the mistake of predicting a 'Yes' when its actually 'No' only 1.5% of the time ,it only predicts 'Yes' when its actually 'Yes' 17% of the time and that's quite low too.The BACC performs much better in this regard , similar to our previous models at 74% for TPR and 22% for FPR and that is why as stated earlier that it is advisable for BACC to be paired with ACC.

We also see that there have been mire accurate true negative (TN) and false negatives (FN) predictions than all other models used and for ACC .The BACC however gives us a balanced view of the negatives to positives.

Looking at a scatter plot of thresholds versus accuracy from Figure 4.8, we see that even though its similar to all our other models ,the distance between accuracy levels for ACC and BACC get wider as thresholds increase. See Appendix .13 for scatter plot of test set.

Analysing the performance and precision of our tuned model from Table 4.9, we see

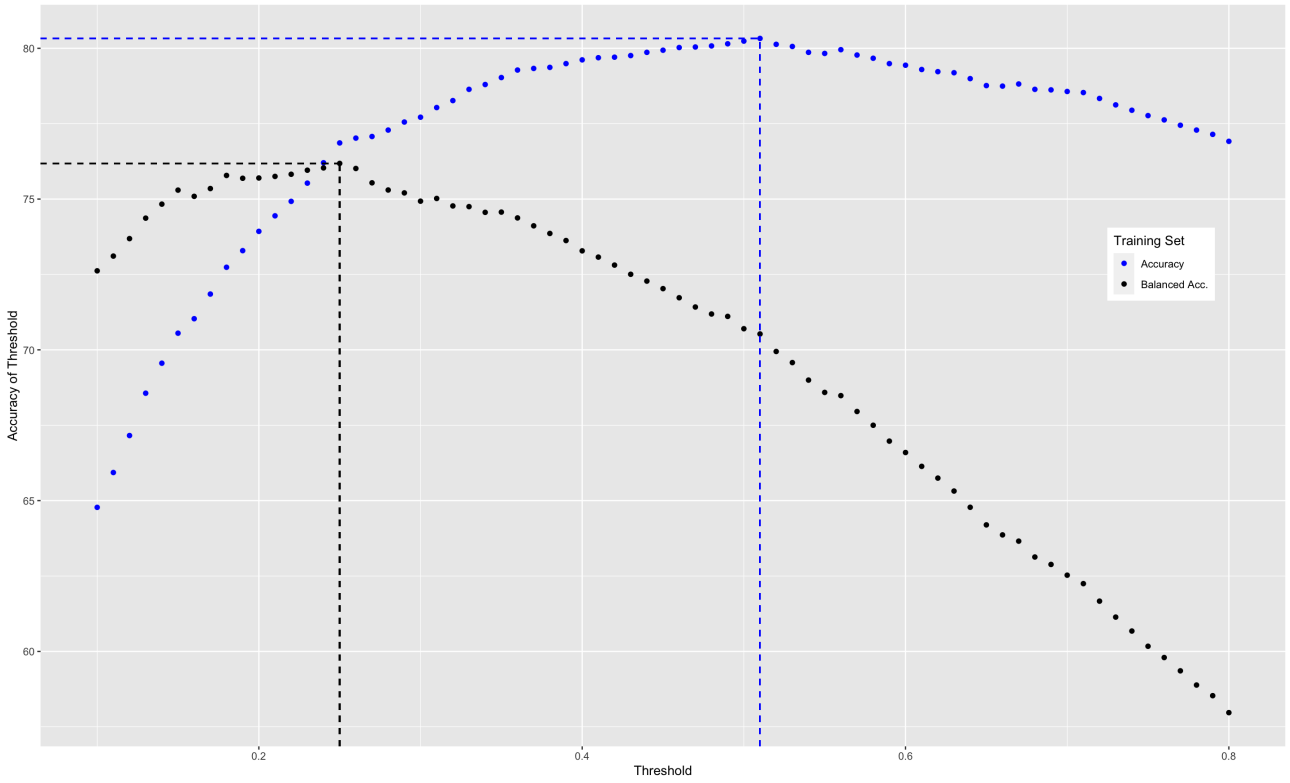

Figure 4.8: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (Training Set).

		Optimal Threshold	Accuracy Level	Precision
Accuracy	Training Set	*0.51*	*80.3%*	*80.1%*
	Test Set	*0.51*	*80.0%*	*68.4%*
Balanced Accuracy	Training Set	*0.25*	*75.87%*	*54.0%*
	Test Set	*0.25*	*73.3%*	*50.0%*

Table 4.9: Performance and Precision of Tuned RF model.

that the precision of ACC stands out as higher(80.1%) than all other previous models with and accuracy level 0f 80% but all other metrics are quite similar to our previous models.

The AUC for our tuned model is however higher than BACC and ACC at 83.29% for training data set and 81.94% for test data set(see Appendix .14 and Appendix .15).

4.2.3 Feature Importance

Feature importance is a key part of predictive modelling. It basically tells us what variables are important across all models used in the analysis.

Step-wise LR model	Feature importance(RF model)	Feature importance (tuned R.F model)
Multiple lines	Total charges	Total charges
Senior citizen	Average total charges	Tenure
Paperless billing	Monthly charges	Average total charges
Online security	Tenure	Monthly charges
Payment method	Contract	Contract
Contract	Payment method	Tenure(by year)
Tenure(by year)	Internet service	Payment method
Internet service	Tenure(by year)	Internet service
	Paperless billing	Paperless billing
	Gender	Online security

Table 4.10: Important variables across step-wise logistic regression , random forest and tuned random forest variables ranked from highest to lowest probability of churn.

Looking at Table 4.10, we see the most important variables across the step-wise logistic regression, random forest and tuned random forest model.

While some variables are consistent across the RF and Tuned RF model, only a few run across step-wise and RF or step-wise and tuned RF. Variables like payment method, paperless billing,contract, tenure (by year) and internet service cut across all three of our models which is an indication that they are important in predicting churn.

Referring back to our exploratory data analysis, we can see the consistency between our key noted variables and some of the features in our feature importance. Gender was deemed benign from our EDA and sure enough, its the last feature in our RF model. Variables like tenure and multiple lines show up in our key models and feature importance list.

Chapter 5

Chapter 5 - Conclusions and Further work

		Training Set			Test Set	
Model	Acc.	Bal. Acc.	AUC	Acc.	Bal. Acc	AUC
Logistic regression	*79.4%*	*76.5%*	*84.1%*	*80.7%*	*76.4%*	*84.2%*
Stepwise L.R	*79.7%*	*76.4%*	*83.2%*	*79.8%*	*75.4%*	*84.1%*
Random forest	*80.1%*	*75.8%*	*83.6%*	*80.0%*	*73.9%*	*82.1%*
Tuned R.F	*80.3%*	*75.87%*	*83.2%*	*80.0%*	*73.3%*	*81.9%*

Table 5.1: Comparative table describing accuracy levels across all models.

Customer churn is one of the major problems that service companies face. Telecommunication companies in particular have an incentive to find out reasons why customers cancel contracts or leave their services. This paper addresses this issue by trying to predict customers that are more likely to churn so they can be targeted with aggressive marketing advertisements, bonus products etc.

The paper began exploring the proposed data set by undertaking a cleaning and exploratory data analysis. The importance of this is to have an idea of what the data looks like and if there are any problems with the data that should be 'cleaned'. Real life data often are imbalanced so it is important to determine the structure of our data before we work on it. Another reason for exploratory data analysis is that helps us

know how to approach modelling the data. For our data set in particular, we realise that our predictor variable, churn, is imbalanced in favour of non churners so the use of certain performance metrics will be inadequate. This will be discussed in the next paragraph

The next chapter discussed models to be used, how to asses these models and how to check the precision of these models. We looked at Confusion matrix and its associate metrics (which focused on only churns), Accuracy(ACC) , Balanced Accuracy(BACC), Receiver Operating Curve(ROC) and Area Under Curve(AUC).

Chapter 4 deals with the results of our analysis which have been summarised as table 5.1. It gives us a summary of all ACC,BACC and AUC across our model. We see that while the AUC metric shows higher performance figures than others, Random Forest generally outperforms Logistic regression for ACC but Logistic Regression generally outperformed Random forest for BACC and AUC. We also discussed the most important variables across all our models.

Some interesting research to consider for the future would be regress feature variables across models to see why the deviate from model to model.

Applications of continuous mathematical modelling to customer churn (continuous customer behavioral pattern modelling which is also also known as concept drift) would also be an interesting area for further research because customers tend to have ever changing needs and situations so that whatever model employed should be able to capture these needs and changes in real time.

Appendices

.1 Appendix A

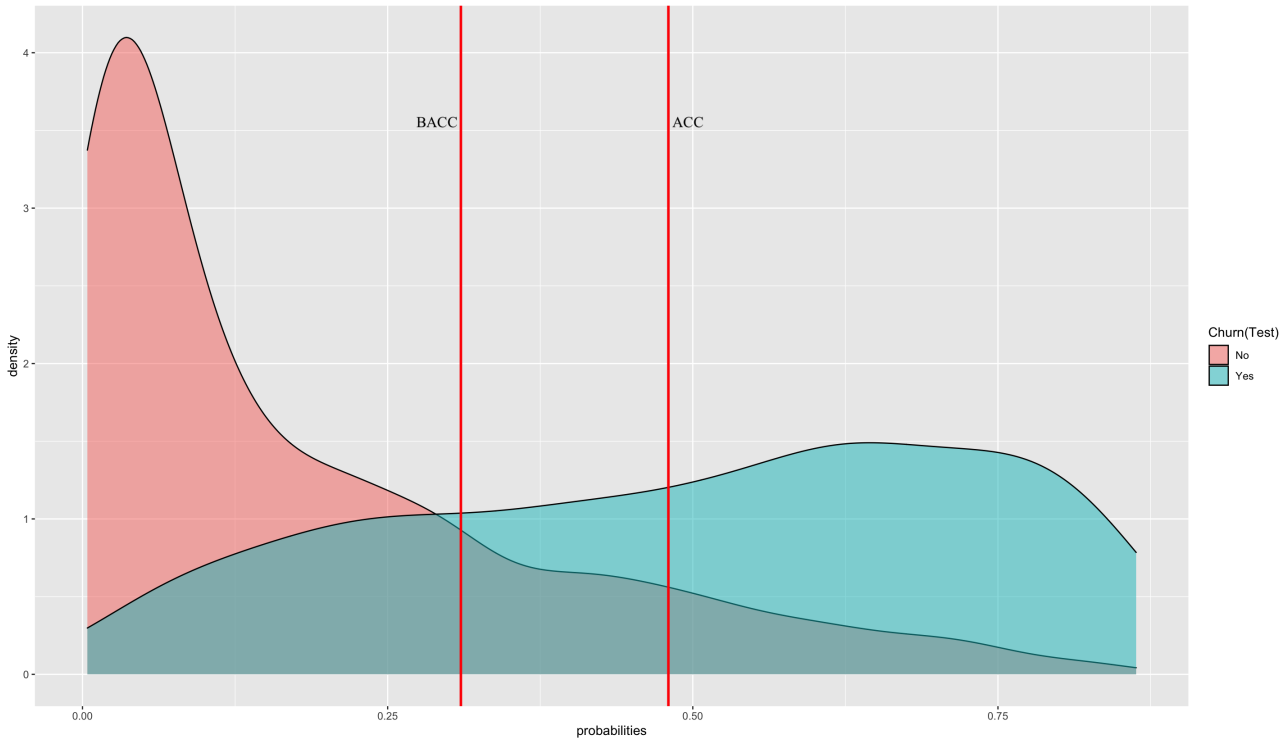

Figure .1: Probability density plot of churn grouped by actual test data.

.2 Appendix A.1

	Accuracy					Bal. Accuracy		
	Predicted Class			Total		Predicted Class		Total
		No	Yes			No	Yes	
No	Test Set	*924*	*108*	*1032*		*812*	*220*	*1032*
Yes	Test Set	*162*	*211*	*373*		*96*	*277*	*373*
	Total	*1086*	*319*	*1405*		*908*	*497*	*1405*

Table .2: Confusion matrix of optimal cutoff point.

.3 Appendix B

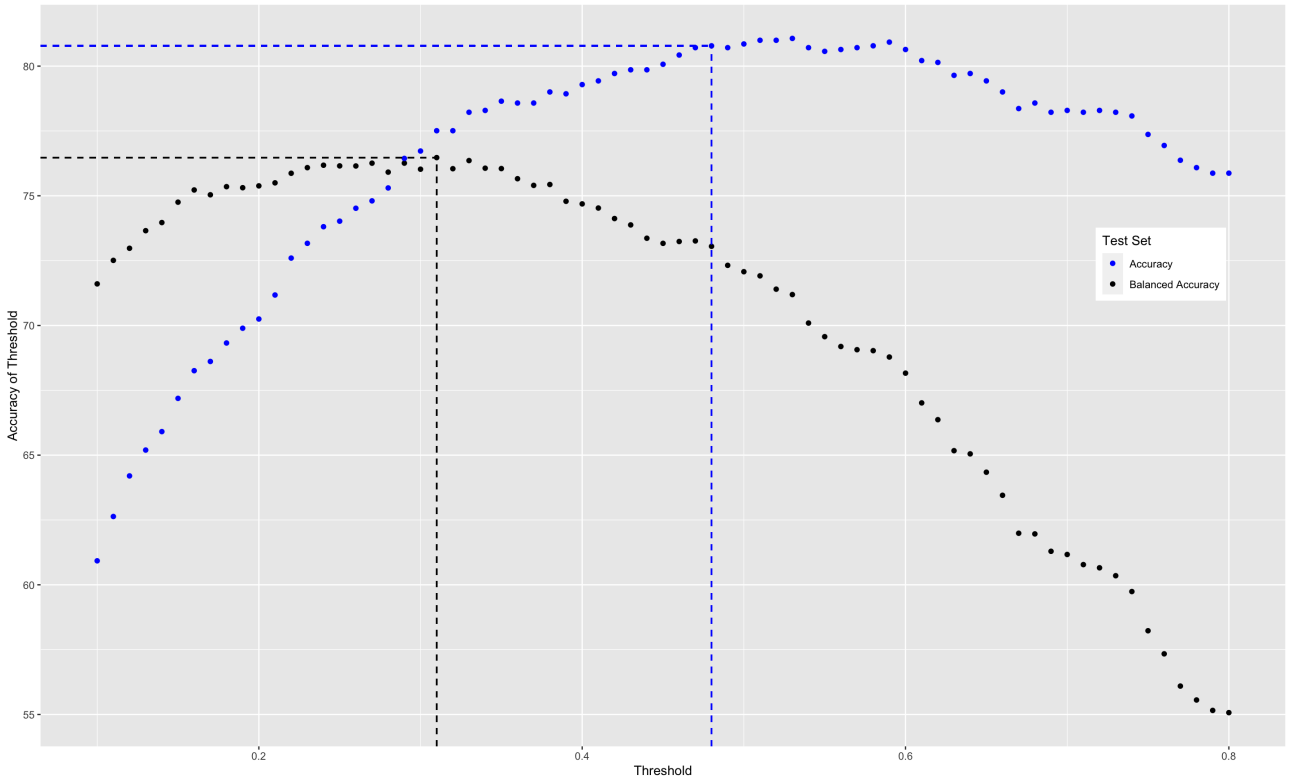

Figure .2: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (test set).

.4 Appendix C

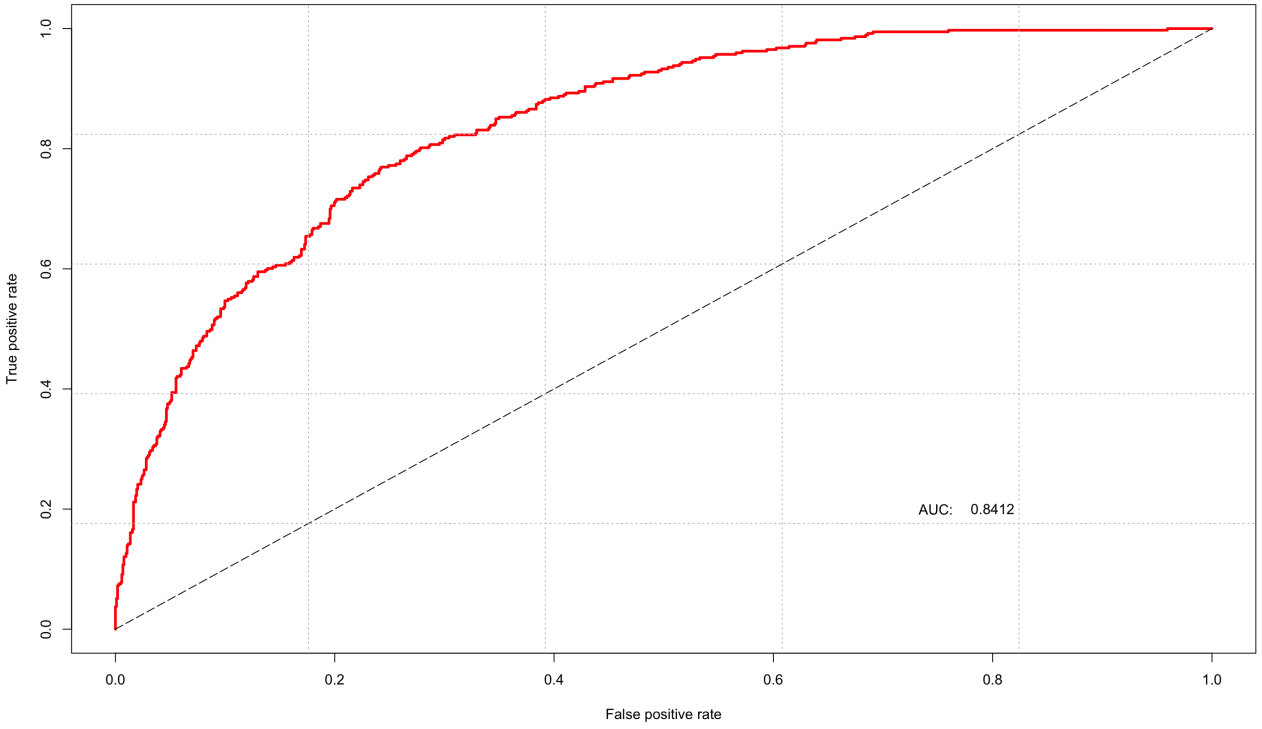

Figure .3: ROC curve and AUC of test set.

.5 Appendix D

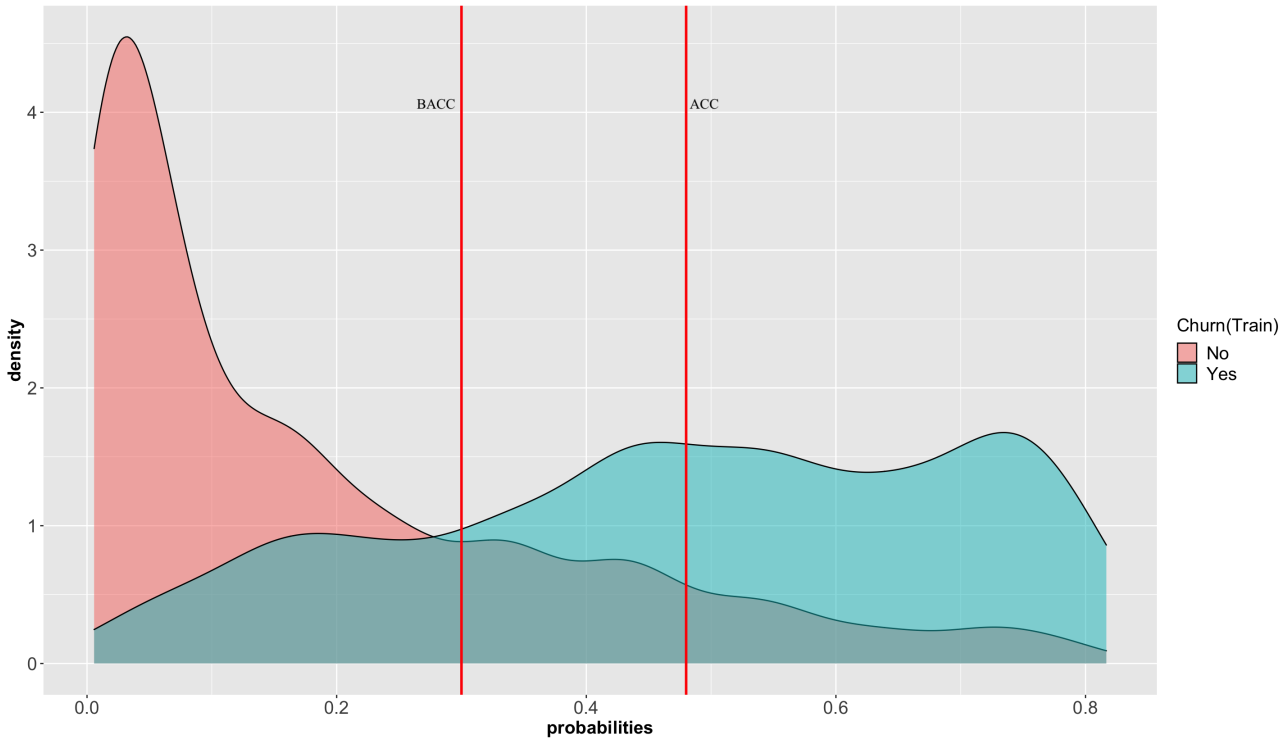

Figure .4: Probability density plot of churn grouped by actual training data.

.6 Appendix D.1

	Accuracy					Bal. Accuracy		
	Predicted Class			Total		Predicted Class		Total
		No	Yes			No	Yes	
No	Test Set	*920*	*112*	*1032*		*782*	*250*	*1032*
Yes	Test Set	*171*	*202*	*373*		*93*	*280*	*373*
	Total	*1091*	*314*	*1405*		*875*	*530*	*1405*

Table .3: Confusion matrix of optimal cutoff point.

.7 Appendix E

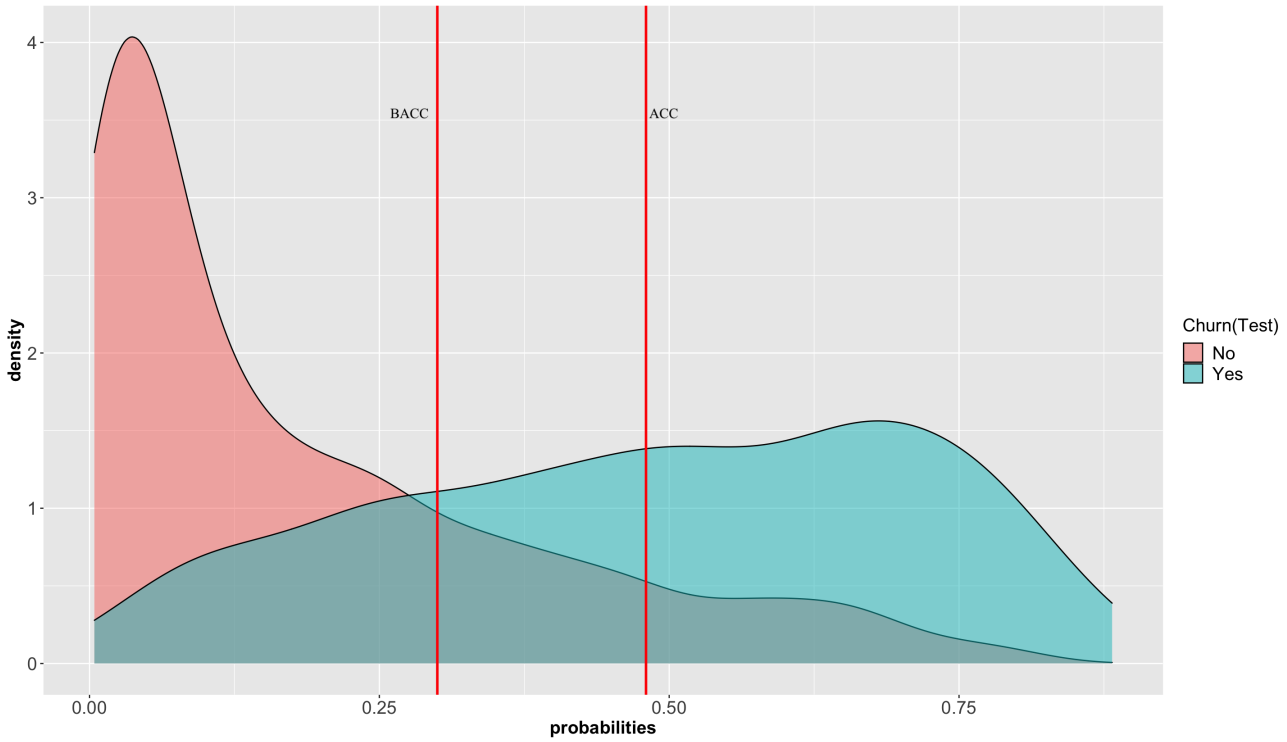

Figure .5: Probability density plot of churn grouped by actual test data.

.8 Appendix F

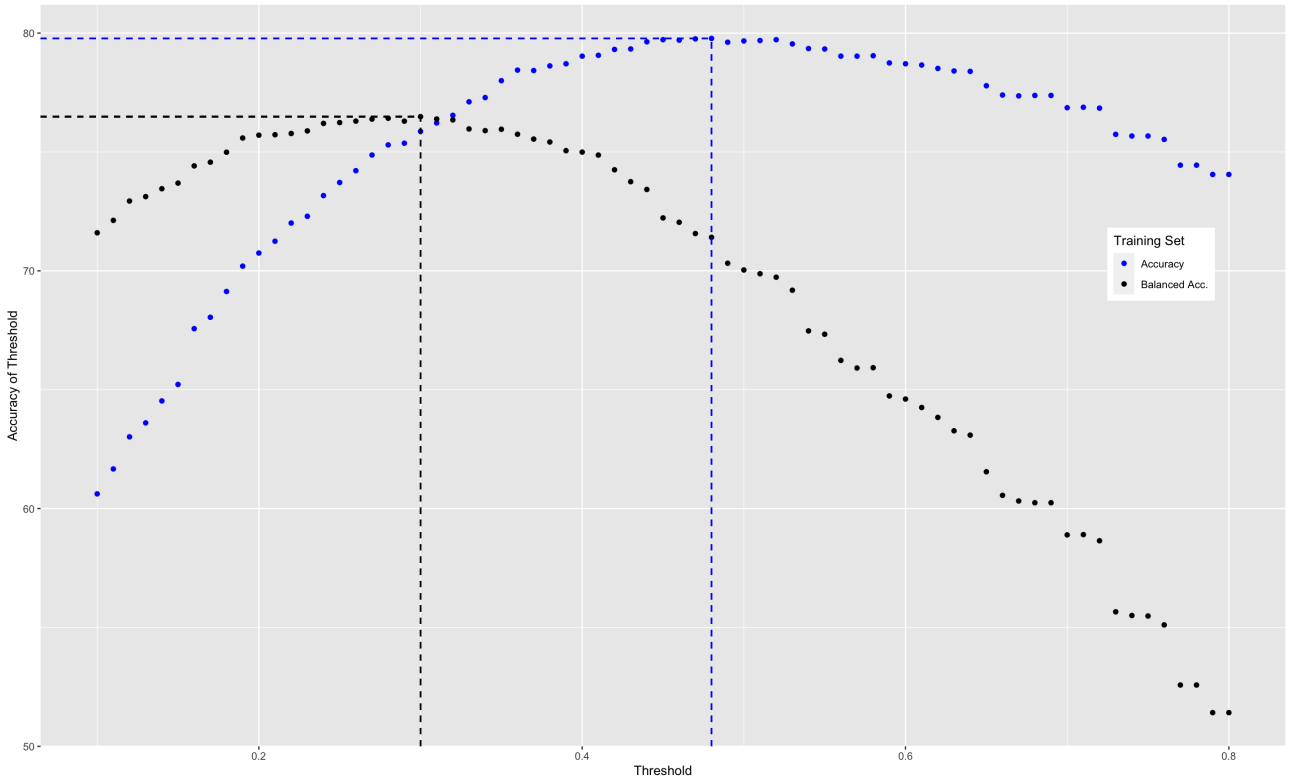

Figure .6: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (training set).

.9 Appendix G

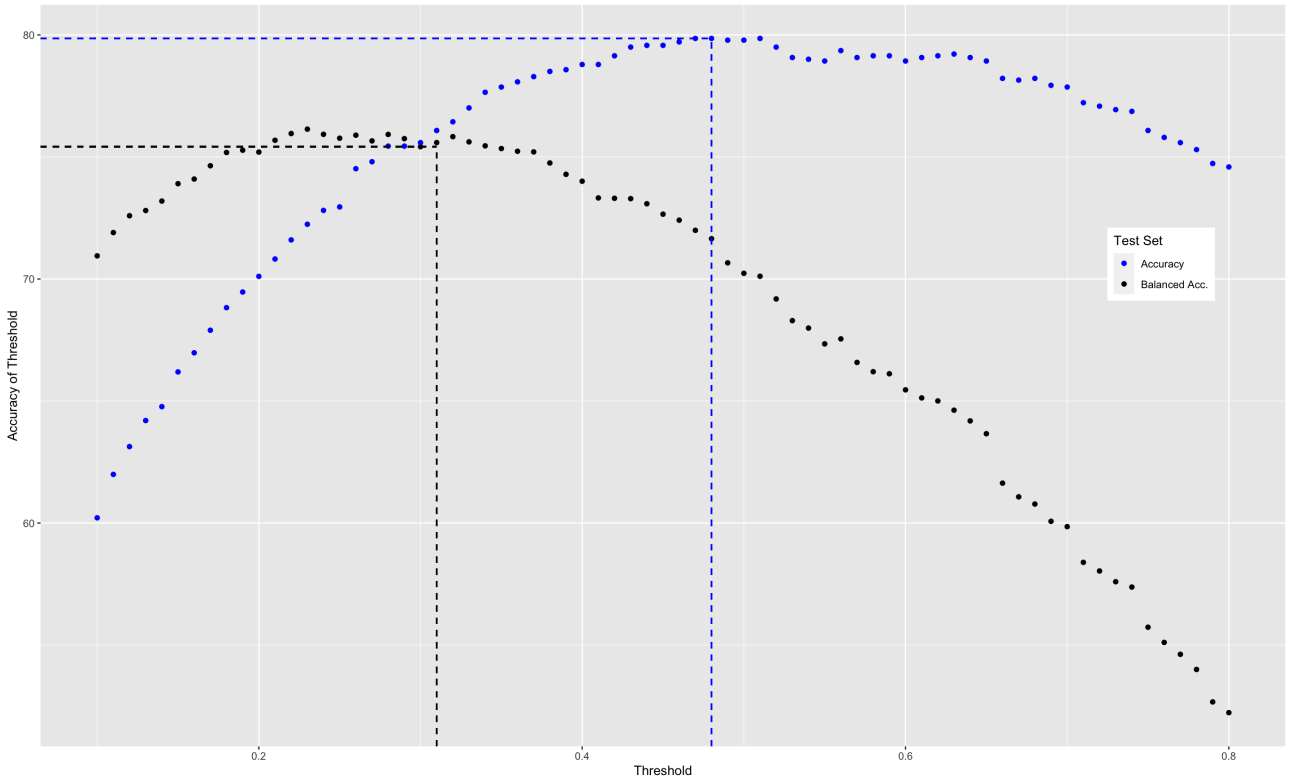

Figure .7: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (test set).

.10 Appendix H

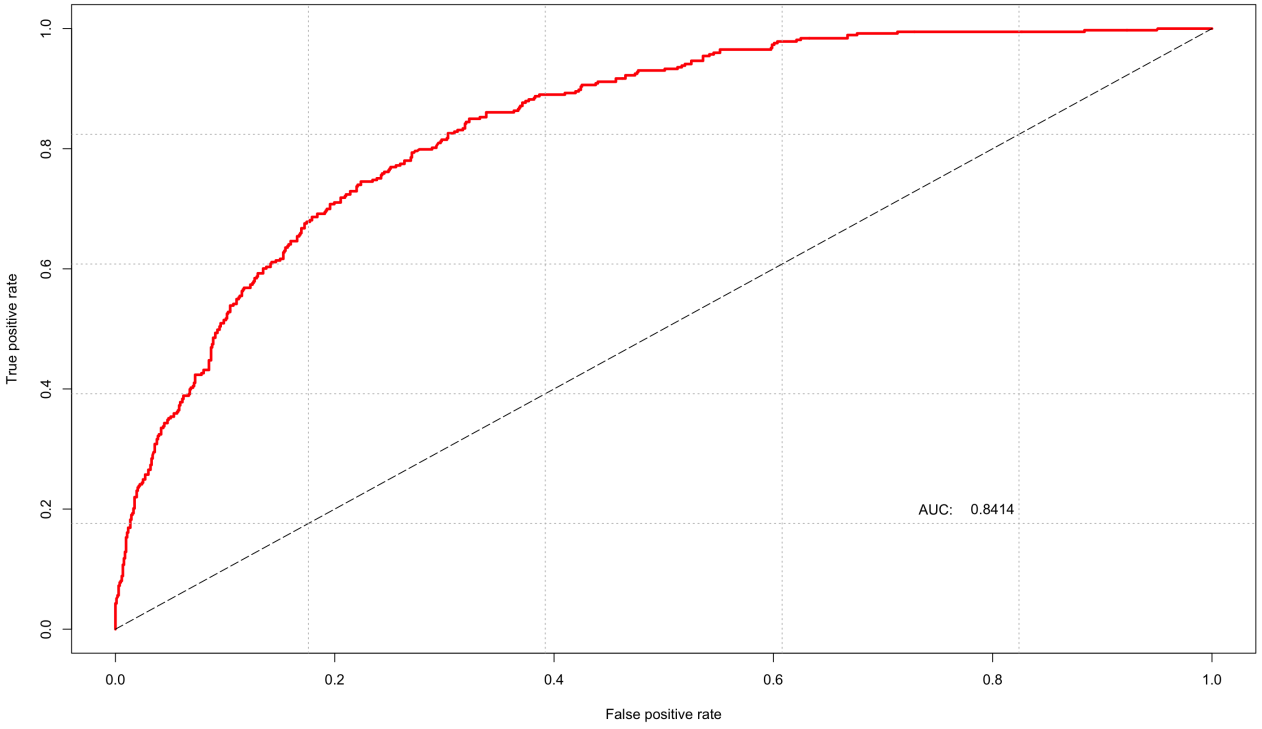

Figure .8: ROC curve and AUC of test set.

.11 Appendix I

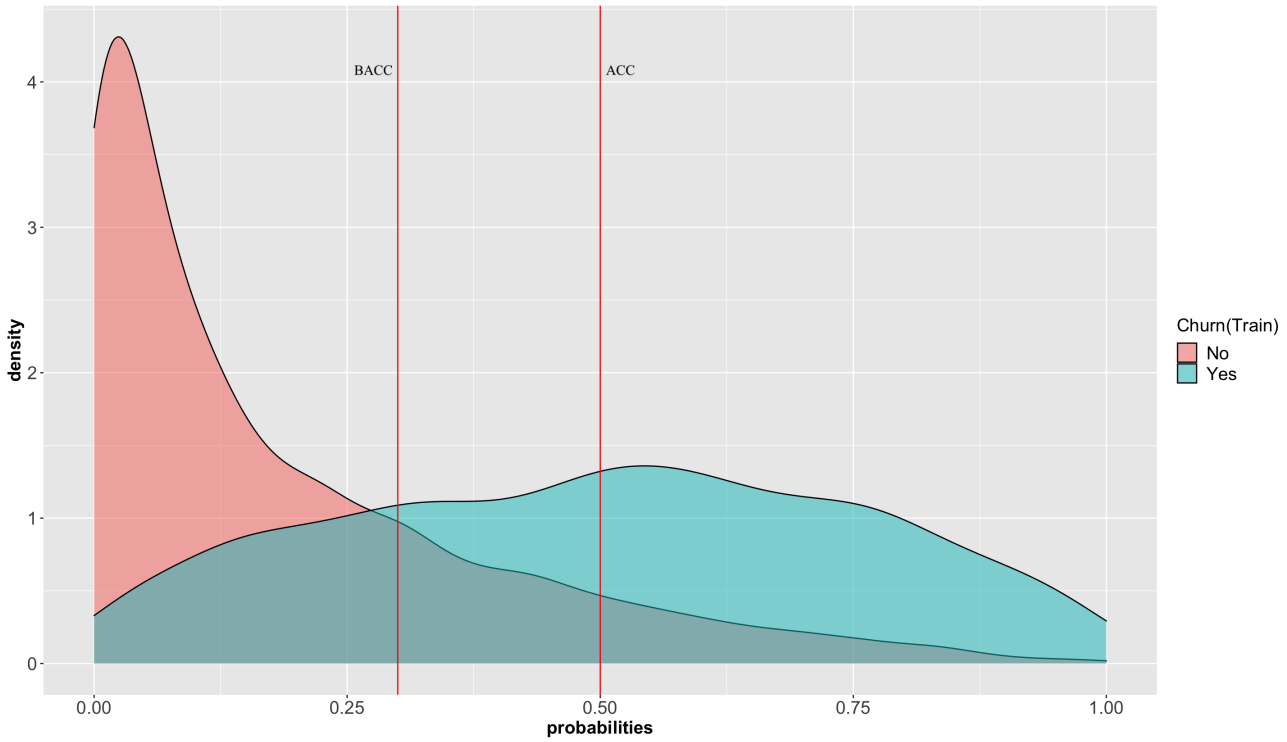

Figure .9: Probability density plot of churn grouped by actual train data.

.12 Appendix J

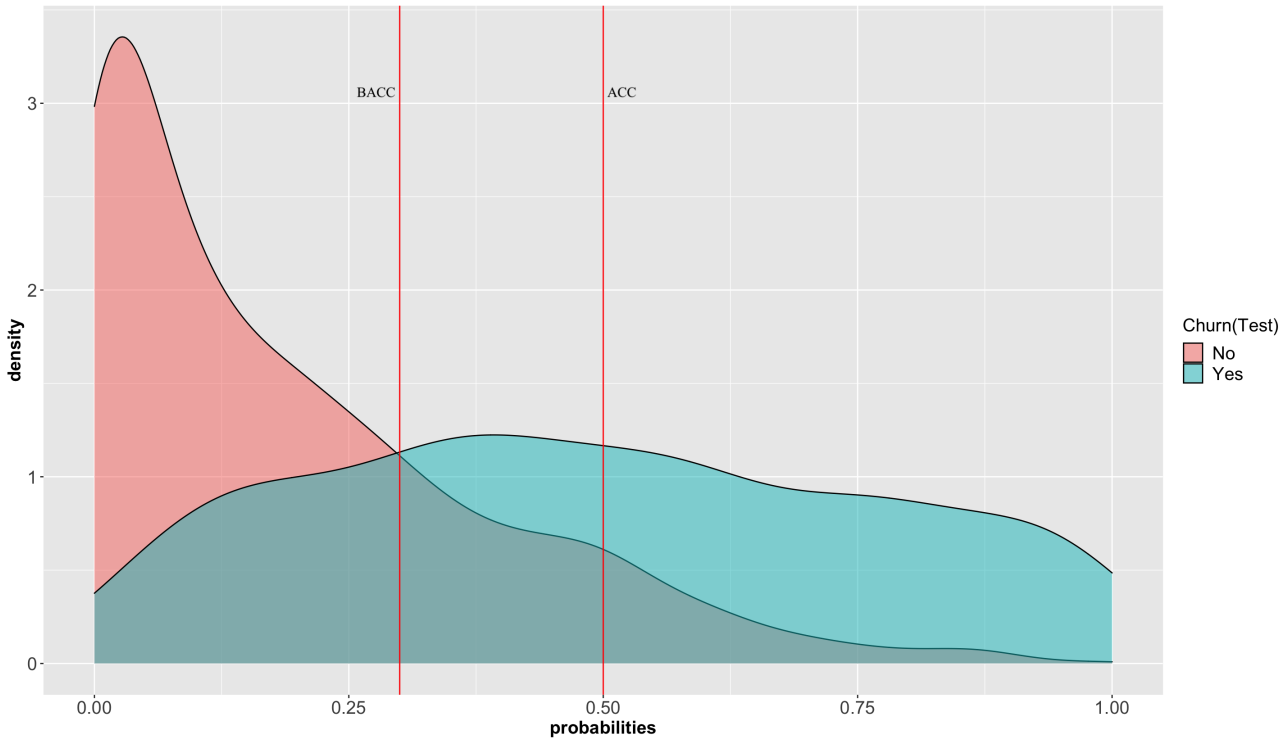

Figure .10: Probability density plot of churn grouped by actual test data.

.13 Appendix K

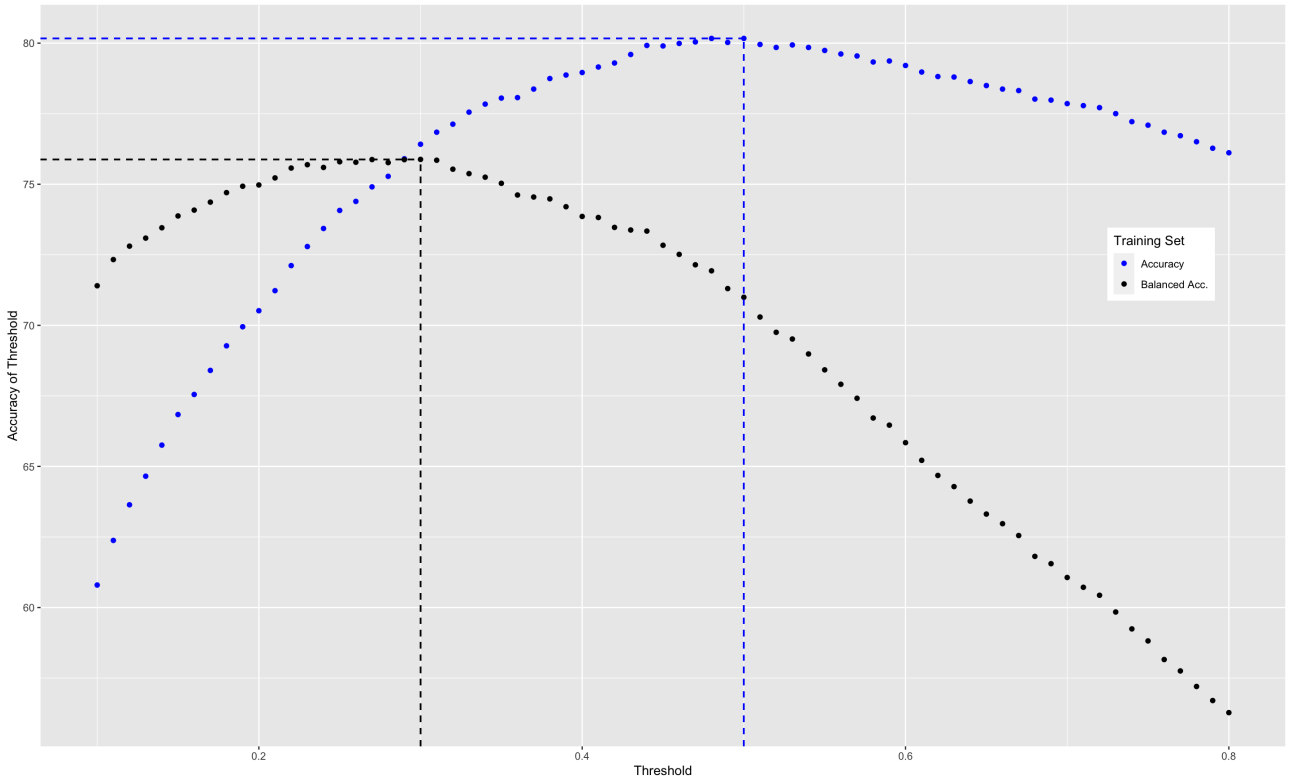

Figure .11: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (training set).

.14 Appendix K.1

	Accuracy					Bal. Accuracy		
	Predicted Class			Total		Predicted Class		Total
		No	Yes			No	Yes	
No	Test Set	*975*	*57*	*1032*		*791*	*241*	*1032*
Yes	Test Set	*218*	*155*	*373*		*104*	*269*	*373*
	Total	*1193*	*212*	*1405*		*895*	*510*	*1405*

Table .4: Confusion matrix of optimal cutoff point.

.15 Appendix L

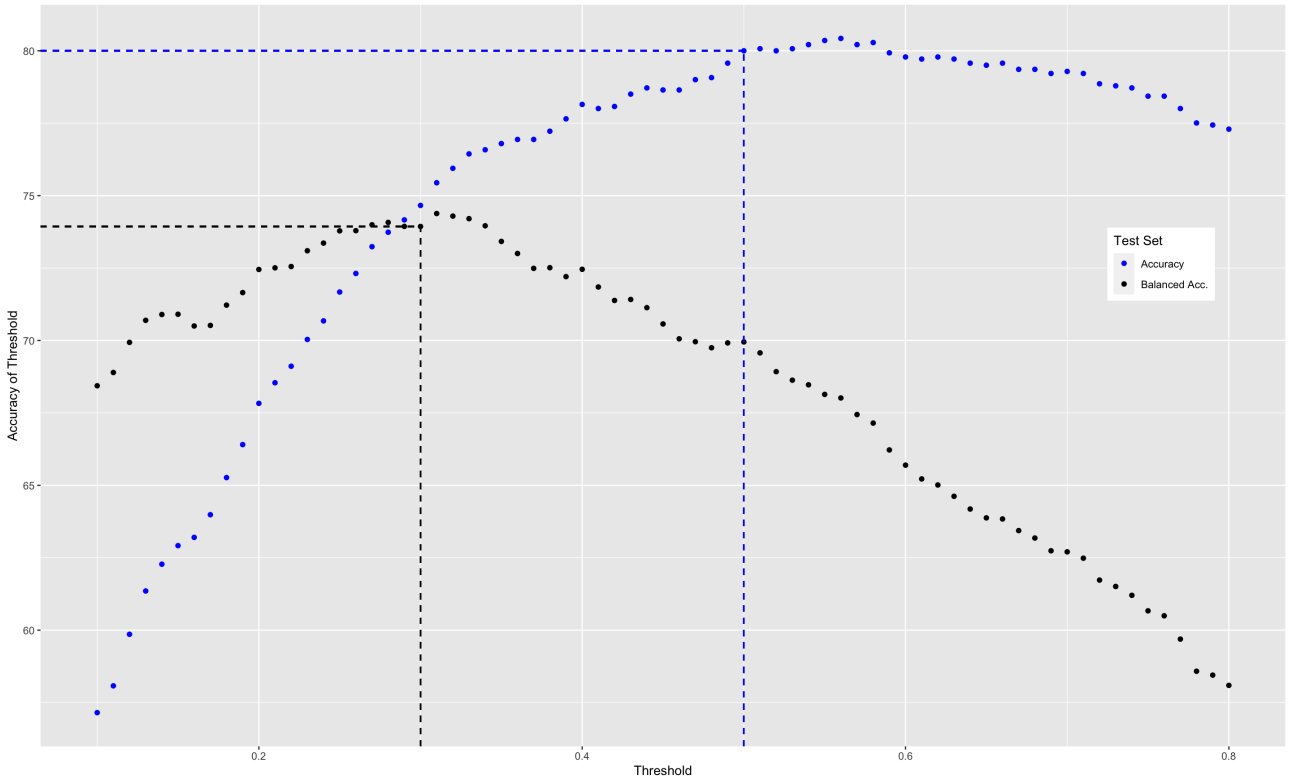

Figure .12: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (test set).

.16 Appendix M

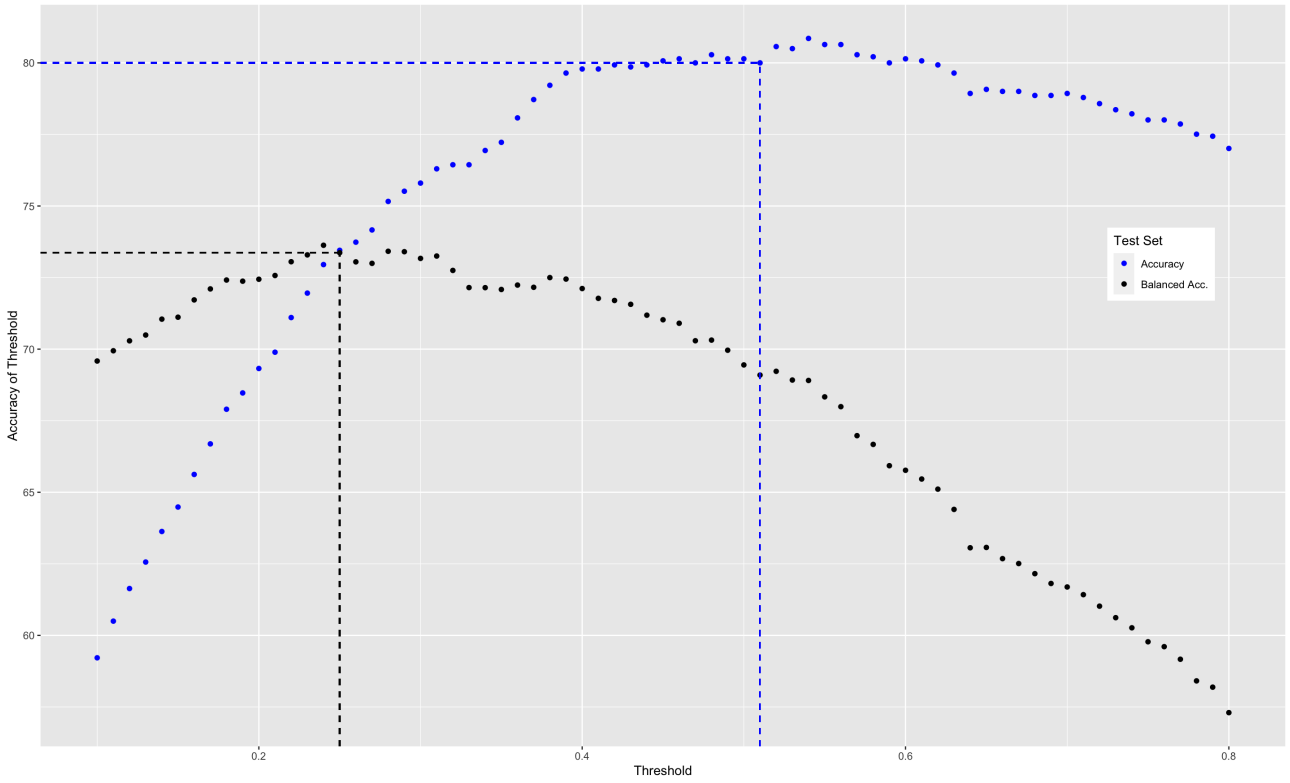

Figure .13: Scatter plot of accuracy and balanced accuracy showing different thresholds and their corresponding accuracy (test Set).

.17 Appendix M.1

	Accuracy					Bal. Accuracy		
	Predicted Class			Total		Predicted Class		Total
		No	Yes			No	Yes	
No	Test Set	*953*	*79*	*1032*		*759*	*273*	*1032*
Yes	Test Set	*202*	*171*	*373*		*100*	*273*	*373*
	Total	*1155*	*250*	*1405*		*859*	*510*	*1405*

Table .5: Confusion matrix of optimal cutoff point.

.18 Appendix N

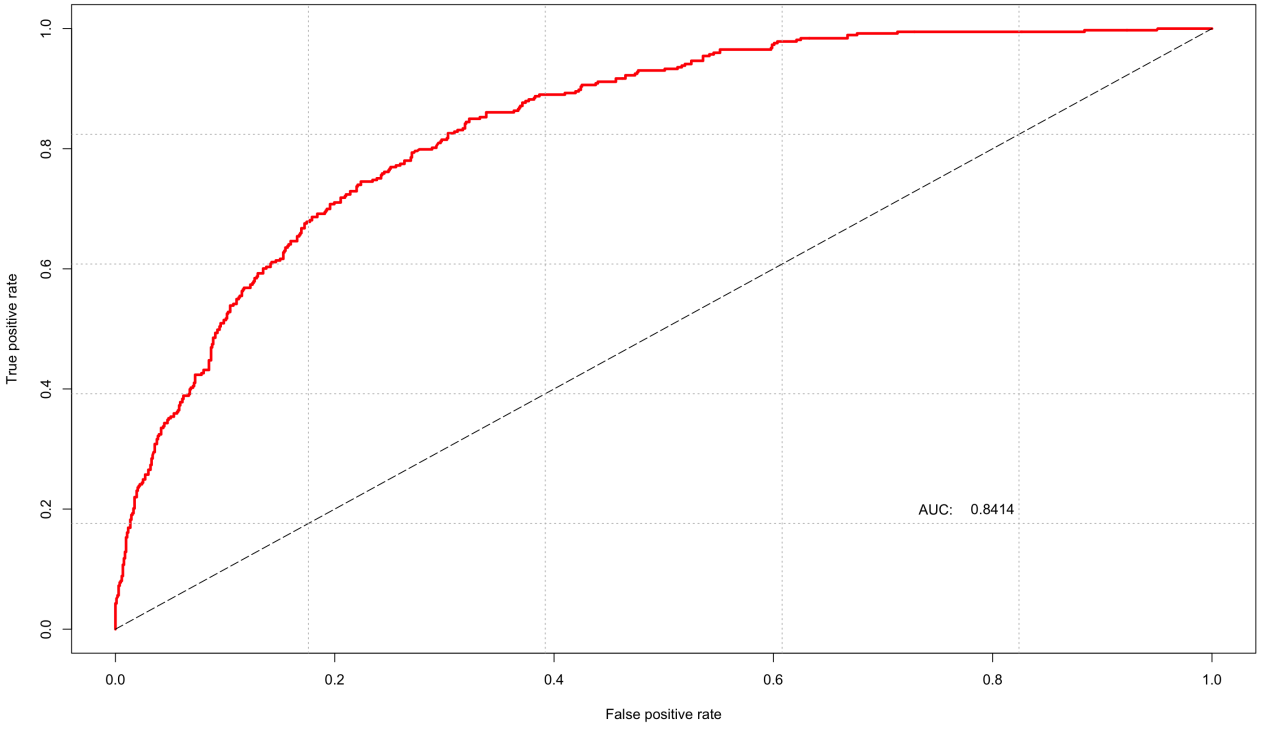

Figure .14: ROC curve and AUC of training set.

.19 Appendix O

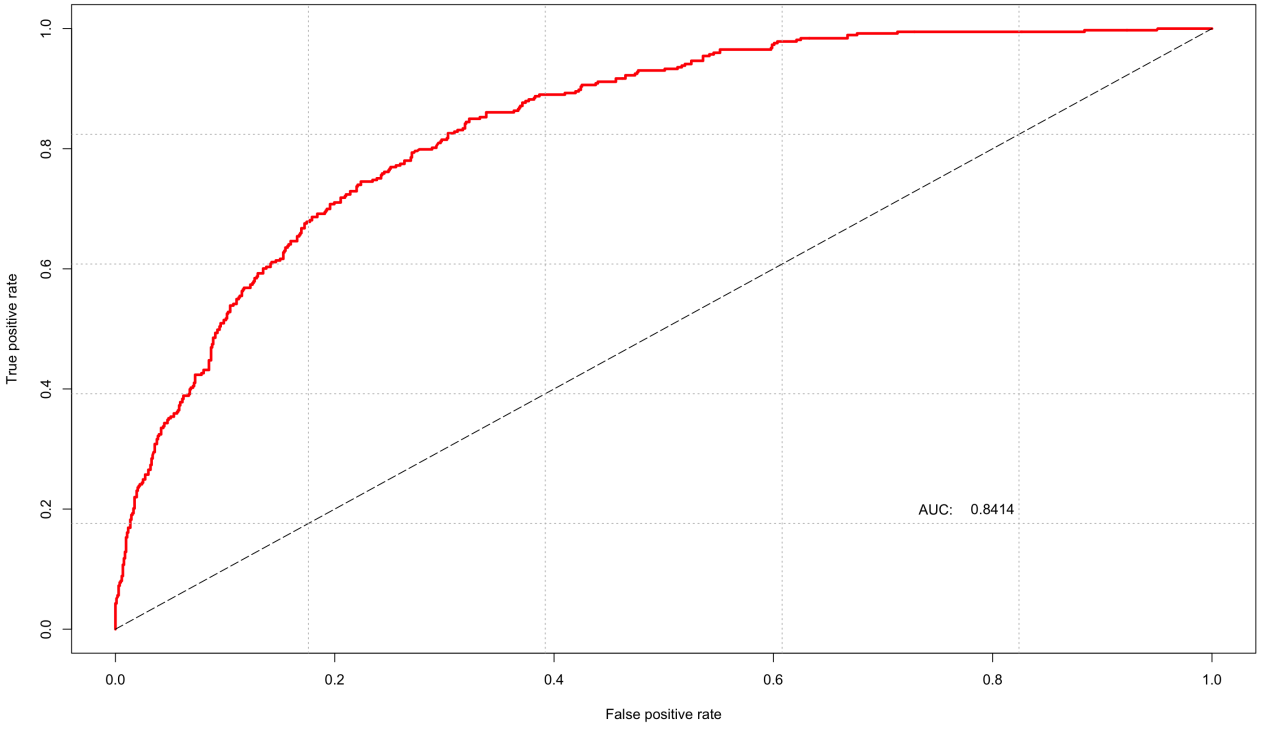

Figure .15: ROC curve and AUC of test set.

Bibliography

[ADGD18] Sanket Agrawal, Aditya Das, Amit Gaikwad, and Sudhir Dhage. Customer churn prediction modelling based on behavioural patterns analysis using deep learning. pages 1–6, 07 2018.

[AJA19] Abdelrahim Ahmad, Asef Jafar, and Kadan Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6, 03 2019.

[BV09] J. Burez and D. Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3, Part 1):4626–4636, 2009.

[cel]

[cra] glm: Fitting generalized linear models.

[DvS+07] Andy.K Devos, Sabine van Huffel, Arjan W. Simonetti, Marinette van der Graaf, Arend Heerschap, and Lutgarde M.C. Buydens. Chapter 11 - classification of brain tumours by pattern recognition of magnetic resonance imaging and spectroscopic data. In Azzam F.G. Taktak and Anthony C. Fisher, editors, *Outcome Prediction in Cancer*, pages 285–318. Elsevier, Amsterdam, 2007.

[Faw06] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[FHR14] Navid Forhad, Md Hussain, and Mohammad Rahman. Churn analysis: Predicting churners. pages 237–241, 09 2014.

[HBI13] Nabgha Hashmi, Naveed Anwer Butt, and Dr.Muddesar Iqbal. Customer churn prediction in telecommunication a decade review and classification. *IJCSI*, 10:271–282, 09 2013.

[HCJ17] Md Hossain, Md Chowdhury, and Nusrat Jahan. Customer retention and telecommunications services in bangladesh. *International Journal of Asian Social Science*, 7:921–930, 01 2017.

[HJS04] Hyunseok Hwang, Taesoo Jung, and Euiho Suh. An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry. *Expert Systems with Applications*, 26(2):181–188, 2004.

[JST14] Sadaf Hossein Javaheri, Mohammad Mehdi Sepehri, and Babak Teimourpour. Chapter 6 - response modeling in direct marketing: A data mining-based approach for target selection. In Yanchang Zhao and Yonghua Cen, editors, *Data Mining Applications with R*, pages 153–180. Academic Press, Boston, 2014.

[KKKH20] V. Kavitha, G. Kumar, S. Kumar, and M. Harish. Churn prediction of customer in telecom industry using machine learning algorithms. *International Journal of Engineering Research and*, V9, 05 2020.

[Kot13] Sotiris Kotsiantis. Decision trees: A recent overview. *Artificial Intelligence Review*, pages 1–23, 04 2013.

[LMCS21] Praveen Lalwani, Manas Mishra, Jasroop Chadha, and Pratyush Sethi. Customer churn prediction system: a machine learning approach. *Computing*, pages 1–24, 02 2021.

[NAR18] Md Nekmahmud Argon and Shafiqur Rahman. *Measuring the Competitiveness Factors in Telecommunication Markets*, pages 339–372. 05 2018.

[PA11] David Powers and Ailab. Evaluation: From precision, recall and f-measure to roc, informedness, markedness correlation. *J. Mach. Learn. Technol*, 2:2229–3981, 01 2011.

[RFF+14] Ali Rodan, Ayham Fayyoumi, Hossam Faris, Jamal Alsakran, and Omar Al-Kadi. Negative correlation learning for customer churn prediction: A comparison study. *The Scientific World Journal*, 2015, 09 2014.

[tel]

[TRMB+18] Antti Tolonen, H.F.M. Rhodius-Meester, Marie Bruun, Juha Koikkalainen, Frederik Barkhof, Afina Lemstra, Teddy Koene, Philip Scheltens, Charlotte Teunissen, Tong Tong, Ricardo Guerrero, Andreas Schuh, Christian Ledig, Marta Baroni, Daniel Rueckert, Hilkka Soininen, Anne Remes, Gunhild Waldemar, Steen Hasselbalch, and Jyrki Lötjönen. Data-driven differential diagnosis of dementia using multiclass disease state index classifier. *Frontiers in Aging Neuroscience*, 10:111, 04 2018.

[tun] tunerf: Tune randomforest for the optimal mtry parameter.

[UIUA12] Okeh UM, Ogah IE, Okeh US, and Agwu A. The use of receiver operating characteristic (roc) analysis in the evaluation of the performance of two binary diagnostic tests of gestational diabetes mellitus. *International Journal of Asian Social Science*, 2:34–43, 01 2012.

[Xev05] Evangelos Xevelonakis. Developing retention strategies based on customer profitability in telecommunications: An empirical study. *The Journal of Database Marketing Customer Strategy Management*, 12:226–242, 04 2005.

[ZLZA17] Chongsheng Zhang, Changchang Liu, Xiangliang Zhang, and George Almpanidis. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82, 04 2017.