Telecom Customer Churn Project

HarvardX Data Science Professional Certificate: PH125.9x
Capstone project(2)

Ahmed G. Alastal

21 June, 2021

Contents

Executive Summary

The data we will use is part of Kaggle competition https://www.kaggle.com/radmirzosimov/telecom-users-dataset Dataset includes churn data from the telecom Operator. The dataset has 22 variables and 5,987 observations. The main object is to predect the customers with a high probability of churn. We will analysis the dataset and focuses on the behavior of telecom customers who are more likely to leave the platform, and then we will use several techniques of Machine Learning to build a model focused on maximizing the true predictions of customers that will stay with the company. We will use various technical performance measures to compare the models with each other on the basis of : accuracy, sensitifity, specificity, MCC, ROC curve, AUC ...

The machine-learning models which constructed to predicts whether customer is at risk of churn are: Logistic Regression, k-Nearest Neighbours, Decision Tree and Random Forest.

1. Introduction

Any business wants to maximize the number of customers. To achieve this goal, it is important not only to try to attract new ones, but also to retain existing ones. Retaining a client will cost the company less than attracting a new one. In addition, a new client may be weakly interested in business services and it will be difficult to work with him, while old clients already have the necessary data on interaction with the service. Accordingly, predicting the churn, we can react in time and try to keep the client who wants to leave. Based on the data about the services that the client uses, we can make him a special offer, trying to change his decision to leave the operator. This will make the task of retention easier to implement than the task of attracting new users, about which we do not know anything yet.

1.1 Objective

The main goal of this projects to explore through survival analysis techniques the variables and their influence on the customer churn rate in order to propose an action plan to improve customer retention using several techniques of Machine Learning

1.2 Dataset overview

The used data is part of Kaggle competition https://www.kaggle.com/radmirzosimov/telecom-users-dataset, Dataset includes churn data from a Telecom Operator. The raw data contains 5,986 rows (observation) and 22 columns (features). The column "churn" is the outcome we want to predict, y.

The structure of the Telecom data set is shown below:

```
## Rows: 5,986
## Columns: 22
## $ X1               <dbl> 1869, 4528, 6344, 6739, 432, 2215, 5260, 6001, 1480, ~
## $ customerID       <chr> "7010-BRBUU", "9688-YGXVR", "9286-DOJGF", "6994-KERXL~
## $ gender           <fct> Male, Female, Female, Male, Male, Female, Female, Fem~
## $ SeniorCitizen    <fct> 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,~
## $ Partner          <fct> Yes, No, Yes, No, No, Yes, No, No, No, No, Yes, Yes, ~
## $ Dependents       <fct> Yes, No, No, No, No, No, No, No, No, No, Yes, Yes, No~
## $ tenure           <dbl> 72, 44, 38, 4, 2, 70, 33, 1, 39, 55, 52, 30, 60, 50, ~
## $ PhoneService     <fct> Yes, Yes, Yes, Yes, Yes, No, Yes, No, No, Yes, Yes, N~
## $ MultipleLines    <fct> Yes, No, Yes, No, No, No phone service, Yes, No phone~
## $ InternetService  <fct> No, Fiber optic, Fiber optic, DSL, DSL, DSL, Fiber op~
## $ OnlineSecurity   <fct> No internet service, No, No, No, Yes, Yes, Yes, No, N~
## $ OnlineBackup     <fct> No internet service, Yes, No, No, No, No, No, No, No,~
## $ DeviceProtection <fct> No internet service, Yes, No, No, Yes, Yes, No, No, Y~
## $ TechSupport      <fct> No internet service, No, No, No, No, Yes, No, No, Yes~
## $ StreamingTV      <fct> No internet service, Yes, No, No, No, No, No, No, No,~
## $ StreamingMovies  <fct> No internet service, No, No, Yes, No, Yes, Yes, No, N~
## $ Contract         <fct> Two year, Month-to-month, Month-to-month, Month-to-mo~
## $ PaperlessBilling <fct> No, Yes, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, N~
## $ PaymentMethod    <fct> Credit card (automatic), Credit card (automatic), Ban~
## $ MonthlyCharges   <dbl> 24.10, 88.15, 74.95, 55.90, 53.45, 49.85, 90.65, 24.9~
## $ TotalCharges     <dbl> 1734.65, 3973.20, 2869.85, 238.50, 119.50, 3370.20, 2~
## $ Churn            <fct> No, No, Yes, No, No, No, No, No, No, No, No, Yes, No,~
```

Description of the variables

- Customer ID: Customer ID
- Gender: Whether the customer is a male or a female
- Senior Citizen: Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- Tenure: Number of months the customer has stayed with the company
- Phone Service: Whether the customer has a phone service or not (Yes, No)
- Multiple Lines: Whether the customer has multiple lines or not (Yes, No, No phone service)
- Internet Service Customer's: internet service provider (DSL, Fiber optic, No)
- Online Security: Whether the customer has online security or not (Yes, No, No internet service)
- Online Backup: Whether the customer has online backup or not (Yes, No, No internet service)

- Device Protection: Whether the customer has device protection or not (Yes, No, No internet service)
- Tech Support: Whether the customer has tech support or not (Yes, No, No internet service)
- Streaming TV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- Streaming Movies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-to-month, One year, Two year)
- Paperless Billing: Whether the customer has paperless billing or not (Yes, No)
- Payment Method: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- Monthly Charges: The amount charged to the customer monthly
- Total Charges: The total amount charged to the customer
- Churn: Whether the customer churned or not (Yes or No)

The metadata can be divided into the following groups:

1. Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
2. Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
3. Demographic info about customers – gender, age range, and if they have partners and dependents.
4. Censoring - customers who left within the last month – the event is stored in the column called Churn and the time to it is on column tenure.
5. Churn is the predictor variable, if the value "Yes" indicates the customer left the company.

2. Data Analysis

In this section, we will clean the database, in addition we will make an analysis of the variables to understand the customer's behavior.

2.1 Data Wrangling

Several steps were taken to prepare and clean the data for the subsequent Data Analysis and Machine Learning tasks, which are described below:

1. There are 11 missing values in the TotalCharges column, which account for only 0.16% of the total number of observations as shown in the figure bellow. So I remove those 11 rows with missing values.

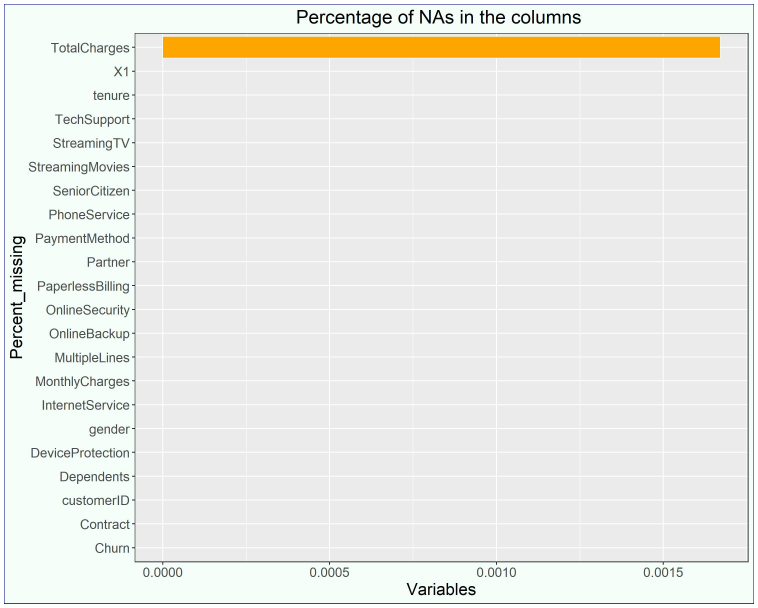

Figure 1: *Percentage of NAs in the columns in dataset*

2. We will not need the the first column and customerID variables for graphs or modeling, so they can be removed.

3. The SeniorCitizen variable is coded '0/1' rather than yes/no. We can recode this to ease our interpretation of later graphs and models.

4. The MultipleLines variable is dependent on the PhoneService variable, where a 'no' for the latter variable automatically means a 'no' for the former variable. We can again further ease our graphics and modeling by recoding the 'No phone service' response to 'No' for the MultipleLines variable.

5. Similiarly, the OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies variables are all dependent on the OnlineService variable. We will recode the responses from 'No internet service' to 'No' for these variables.

The following table illustrates for each feature value its standard deviation, the number of customers that churned vs the ones that didn´t churned and the churn percentage.

Table 1: *Calculated statistics for the churn dataset*

	variable	number	Churn_percentage	Standard_Deviation	Num_NoChurn	Num_Churn
Female	gender	2932	26.98	0.4439220	2141	791
Male	gender	3044	26.15	0.4395228	2248	796
No	SeniorCitizen	5010	23.65	0.4249918	3825	1185
Yes	SeniorCitizen	966	41.61	0.4931742	564	402
Yes1	Partner	2896	19.89	0.3992378	2320	576
No1	Partner	3080	32.82	0.4696514	2069	1011
Yes2	Dependents	1781	15.78	0.3646337	1500	281
No2	Dependents	4195	31.13	0.4630897	2889	1306
Yes3	PhoneService	5388	26.69	0.4423751	3950	1438
No3	PhoneService	588	25.34	0.4353291	439	149
No4	MultipleLines	3431	25.18	0.4341226	2567	864
Yes4	MultipleLines	2545	28.41	0.4510668	1822	723
No5	InternetService	1285	7.70	0.2667632	1186	99
DSL	InternetService	2064	19.23	0.3942381	1667	397
Fiber optic	InternetService	2627	41.53	0.4928680	1536	1091
No6	OnlineSecurity	4267	31.43	0.4642798	2926	1341
Yes5	OnlineSecurity	1709	14.39	0.3511355	1463	246
No7	OnlineBackup	3889	29.24	0.4549067	2752	1137
Yes6	OnlineBackup	2087	21.56	0.4113507	1637	450
No8	DeviceProtection	3924	28.80	0.4528754	2794	1130
Yes7	DeviceProtection	2052	22.27	0.4161663	1595	457
No9	TechSupport	4244	31.13	0.4630646	2923	1321
Yes8	TechSupport	1732	15.36	0.3606495	1466	266
No10	StreamingTV	3673	24.45	0.4298410	2775	898
Yes9	StreamingTV	2303	29.92	0.4579961	1614	689
No11	StreamingMovies	3638	24.49	0.4300958	2747	891
Yes10	StreamingMovies	2338	29.77	0.4573404	1642	696
One year	Contract	1275	11.61	0.3204446	1127	148
Two year	Contract	1432	2.86	0.1668262	1391	41
Month-to-month	Contract	3269	42.77	0.4948140	1871	1398
Yes11	PaperlessBilling	3525	33.50	0.4720698	2344	1181
No12	PaperlessBilling	2451	16.56	0.3718390	2045	406
Electronic check	PaymentMethod	2006	44.97	0.4975826	1104	902
Mailed check	PaymentMethod	1362	19.31	0.3948745	1099	263
Bank transfer (automatic)	PaymentMethod	1306	17.15	0.3771037	1082	224
Credit card (automatic)	PaymentMethod	1302	15.21	0.3592303	1104	198

2.2 Target variable

CHURN column tell us about the number of Customers who left within the last month. From the figure bellow, we can see: $\sim 27\%$ of customers in this dataset have churned:

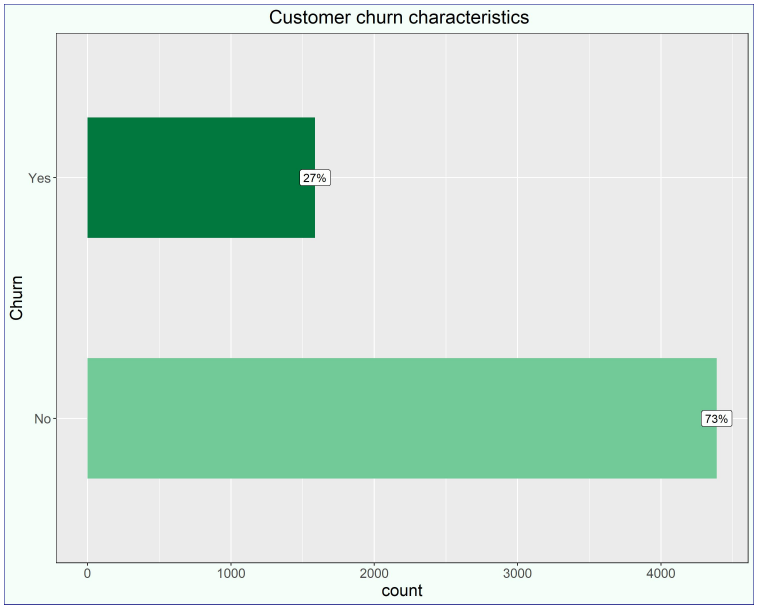

Figure 2: *Customer churn Distribution in dataset*

2.3 Continuous features

There are three numerical columns: tenure, monthly charges and total charges. let's check for distributions for them:

2.3.1 Monthly Charges distribution

We note that the number of existing customers whose monthly fees are under $25 is very high. Otherwise, the distributions are similar between those who churned and those who did not.

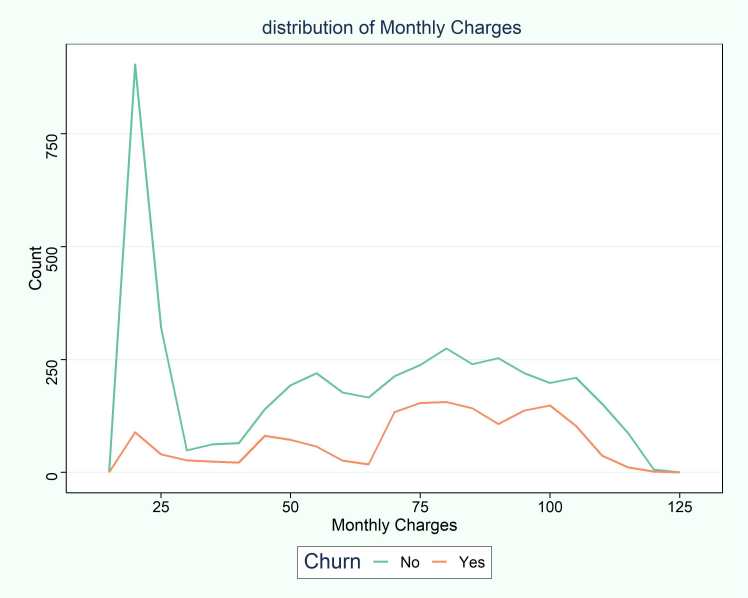

Figure 3: *Distribution of Monthly Charges in telecom dataset*

2.3.2 Total Charges distribution

The Total Charges distribution is positive skew for all customers no matter whether they churned or not.

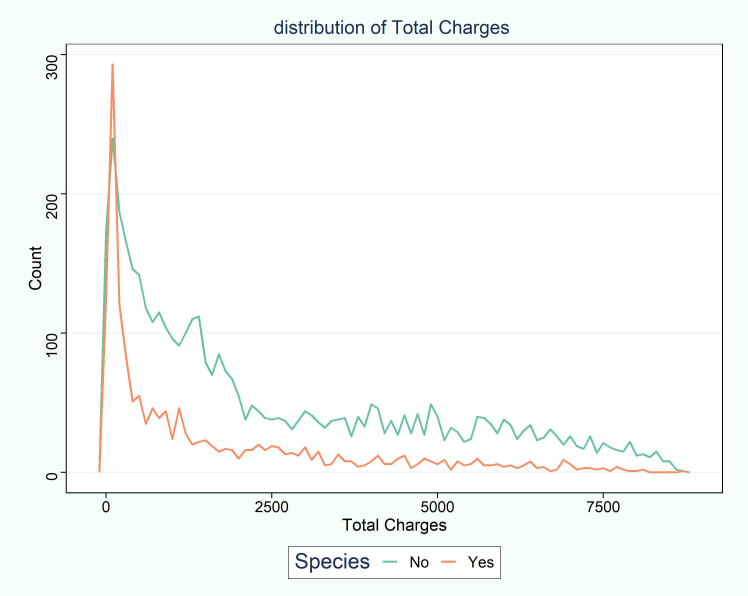

Figure 4: *Distribution of Total Charges in telecom dataset*

2.3.3 Tenure distribution

The distributions for tenure are very different between customers. The customers who churned are more likely to cancel the service in the first months.

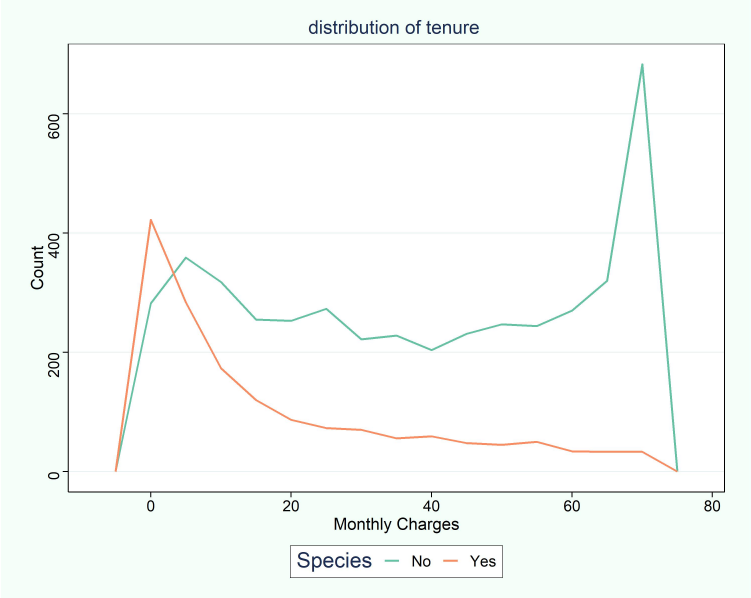

Figure 5: *Distribution of tenure in telecom dataset*

2.4 Categorical features

This dataset has 16 categorical features:

- Six binary features (Yes/No)
- Nine features with three unique values each (categories)
- One feature with four unique values

2.4.1 Gender and Senior Citizen

Figure below show the distribution of target variable "churn" across gender and senior citizen:

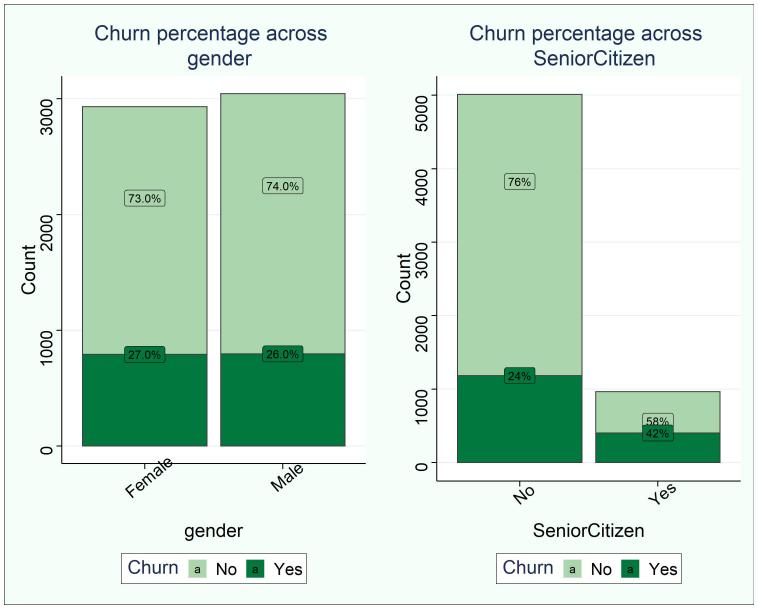

Figure 6: *Distribution of churn percentage across gender and senior citizen*

From figure above, we can conclude:

- Gender is not an indicative of churn.
- Senior Citizens have a much higher churn rate 42% against 24% for non-senior customers.

2.4.2 Partner and dependents

Figure below show the distribution of target variable "churn" across Partner and dependents:

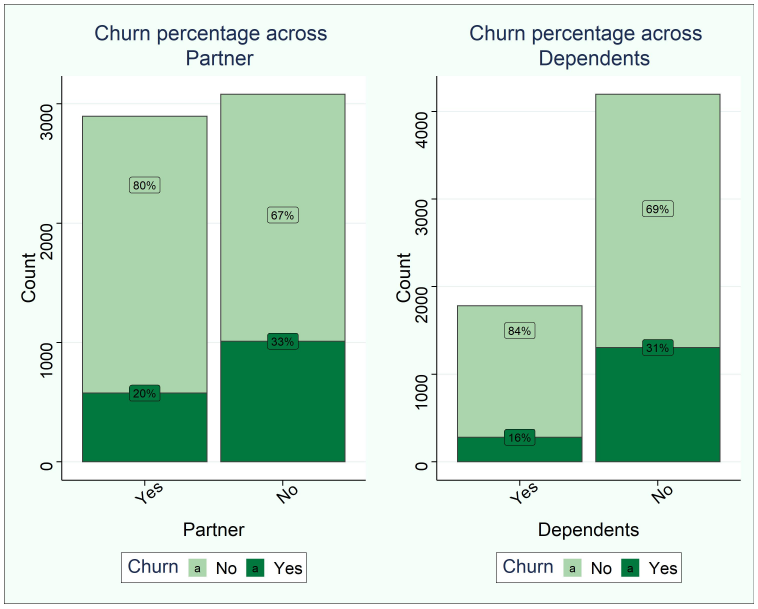

Figure 7: *Distribution of churn percentage across Partner and Dependents*

From figure above, we can conclude:

- Customers that doesn't have partners are more likely to churn.
- Customers without dependents are also more likely to churn.

2.4.3 Phone and Internet services

There are only two main services: phone and internet, figure below show the distribution of target variable "churn" across Phone and Internet services:

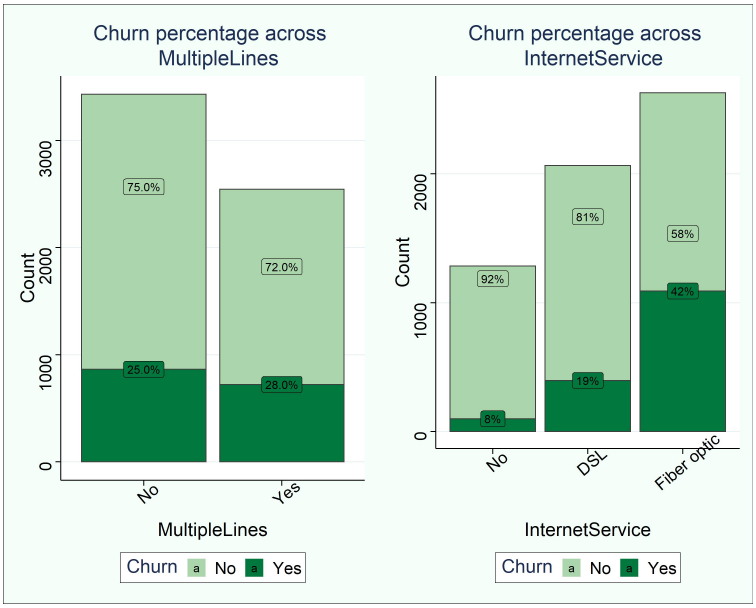

Figure 8: *Distribution of churn percentage across PhoneService and InternetService*

From above, we can conclude:

- Few customers doesn't have phone service.
- Customers with multiple lines have a slightly higher churn rate.
- Clients without internet have a very low churn rate.
- Customers with fiber are more probable to churn than those with DSL connection.

2.4.4 Other services

There are six additional services for customers with internet (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies).

Figures below show the distribution of target variable "churn" across these services:

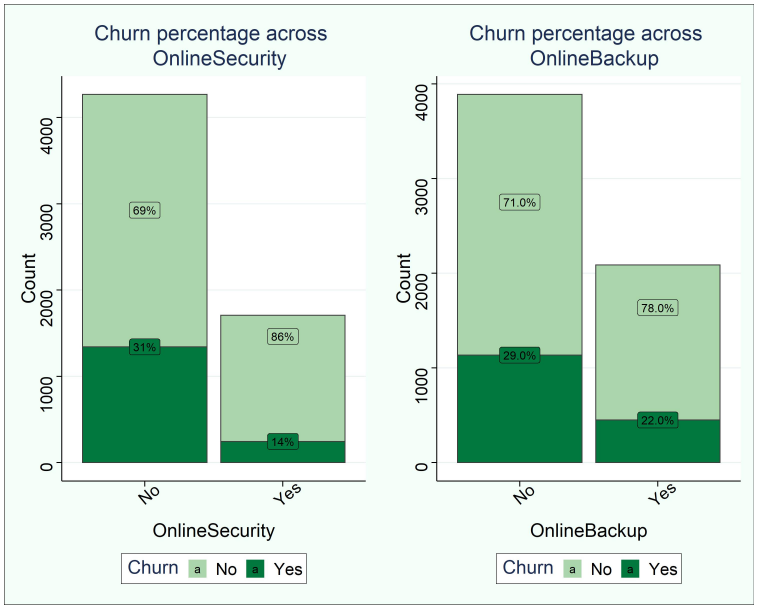

Figure 9: *Distribution of churn percentage across OnlineSecurity and OnlineBackup*

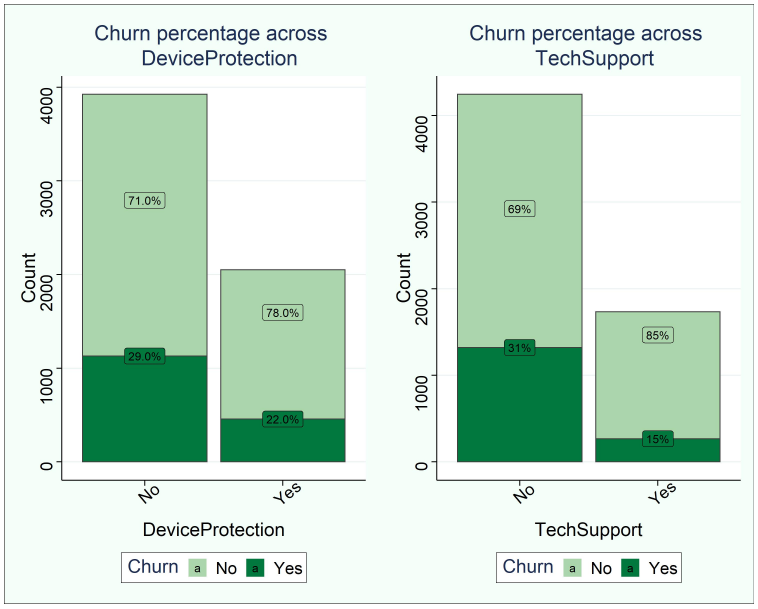

Figure 10: *Distribution of churn percentage across DeviceProtection and TechSupport*

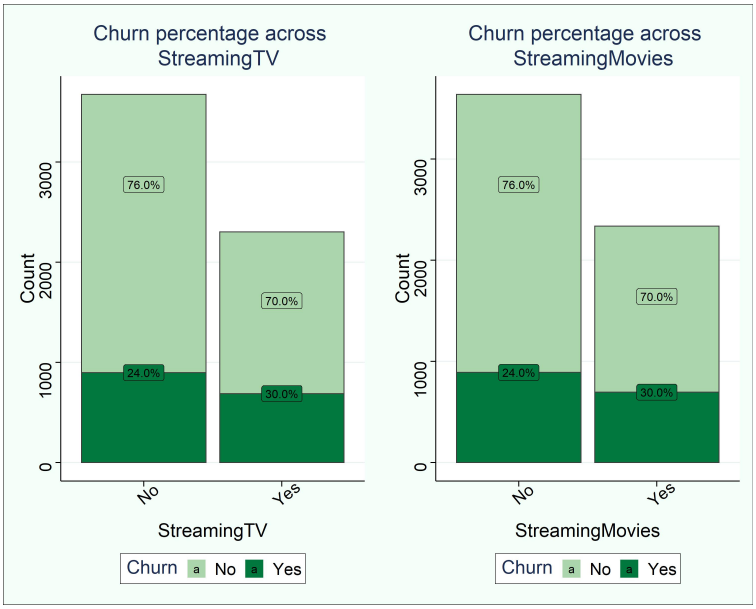

Figure 11: *Distribution of churn percentage across StreamingTV and StreamingMovies*

From figures above, we can conclude:

• The customers who subscribe the service of DeviceProtection, OnlineBackup, OnlineSecurity and TechSupport have lower churn rate compared to the customers who don't.

• Streaming service is not predictive for churn.

2.4.5 Contract and Contract

Figure below show the distribution of target variable "churn" across Contract and Contract:

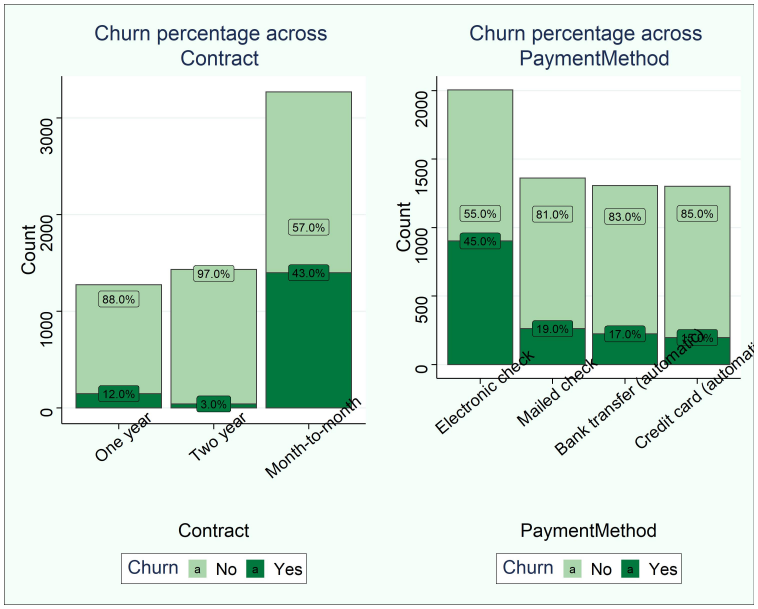

Figure 12: *Distribution of churn percentage across Contract and PaymentMethod*

From figure above, we can conclude:

- The preferred payment method is Electronic check with around 45% of customers. This method also has a very high churn rate.
- Short term contracts have higher churn rates

3. Modeling and Evaluation approach

Post preparing and visualization the data we can now move on to the step of Data Modelling, where we will implemented several models and compared them with each other on the basis of various technical performance measures: accuracy, sensitifity, specificity, MCC, ROC, AUC...

The methods used for prediction of Customer Churn are:

- Logistic Regression.
- k-Nearest Neighbours (KNN).
- Decision Tree.
- Random Forest.

3.1 Important definitions:

- **Accuracy:** The proportion of cases that were correctly predicted in the test set.
- **Sensitivity:** Also known as the true positive rate (TPR) or recall, is the proportion of actual positive outcomes correctly identified as such.
- **Specificity:** Also known as the true negative rate (TNR), is the proportion of actual negative outcomes that are correctly identified as such.
- **FPR:** False positive rate, percentage of misclassified observations in the positive class. Also called false alarm rate or fall-out.
- **PPV:** The positive predictive value tells you how often a positive test represents a true positive.
- **NPV:** The Negative predictive value represents the proportion of individuals with negative test results who are correctly identified or diagnosed.
- **The Matthews correlation coefficient MCC:** The MCC is in essence a correlation coefficient between the observed and predicted binary classifications, it returns avalue between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.
- **ROC curve (A receiver operating characteristic curve):** is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve plots sensitivity (TPR) versus 1 - specificity or the false positive rate (FPR).
- **AUC (Area under the curve):** Computing the area under the curve is one way to summarize it in a single value. This metric is so common that if data scientists say "area under the curve" or "AUC", you can generally assume they mean an ROC curve unless otherwise specified.

• **Confusion Matrix:** A confusion matrix tabulates each combination of prediction and actual value, it determines the results by combining the referenced and predicted outputs. It consists of 4 quadrants which for our ML project can be described as:

- `TP = A subscriber was expected to churn and was correctly classified.`
- `FP = A subscriber was classified as churn but was actually loyal.`
- `FN = A subscriber was classified as loyal but was actually churn.`
- `TN = A subscriber was expected as loyal and was correctly classified.`

3.2 Model Evaluation:

The four ML models employed for the prediction of Churn will be evaluated on the basis of their performance in predicting the tendency to Churn from a Technical perspective. For this purpose the models are made to run on the 20% unseen data which was split during the Data Preparation stage previously.

The performance measures which are looked at are the models accuracy, sensitifity, specificity, its precision in predicting both Good and Bad cases,MCC, ROC and AUC characteristics.

Also, the confusion matrix is generated for all models to find out the performance measures such as TPR, TNR, FNR etc.

3.3 Logistic Regression Model

This section constructs a logistic regression model. The reason logistic regression is used instead of linear regression is that class is a binary variable. Therefore, it is appropriate for a model to predict the probability that the churn of a customer is yes, for example.

The general form of a logistic regression model is

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \mathbf{x}_i^T\beta \tag{1}$$

where $\hat{\pi}_i$ is the estimated probability that observation i is positive, $\mathbf{x}_i$ is the i^{th} vector in the design matrix and β is the vector of coefficients. In this case, the first element of $\mathbf{x}_i$ is 1 to activate the intercept in β, the second element of $\mathbf{x}_i$ is the age of observation i, and the rest of the elements are 1-0 dummy variables.

Let's fit the model using the base general linear modeling function in R, glm.

```
##
## Call:  glm(formula = as.numeric(Churn == "Yes") ~ ., family = "binomial",
##     data = train_data)
##
## Coefficients:
##                              (Intercept)                                genderMale
##                                -0.6976107                                -0.0052865
##                          SeniorCitizenYes                                 PartnerNo
##                                 0.1648333                                 0.0030045
##                              DependentsNo                                    tenure
##                                 0.0400921                                -0.0583788
##                           PhoneServiceNo                     MultipleLinesYes
##                                 0.0753423                                 0.4281794
##                       InternetServiceDSL             InternetServiceFiber optic
##                                 1.5384449                                 3.0967391
##                        OnlineSecurityYes                         OnlineBackupYes
##                                -0.3068482                                -0.0770639
##                      DeviceProtectionYes                          TechSupportYes
##                                 0.0697983                                -0.2185936
##                          StreamingTVYes                      StreamingMoviesYes
##                                 0.5017056                                 0.5467980
##                       ContractTwo year                 ContractMonth-to-month
##                                -0.6682189                                 0.6729219
##                       PaperlessBillingNo              PaymentMethodMailed check
##                                -0.2760212                                -0.4006068
## PaymentMethodBank transfer (automatic)    PaymentMethodCredit card (automatic)
##                                -0.3695346                                -0.4894320
##                           MonthlyCharges                               TotalCharges
```

```
##                                -0.0316326                    0.0003242
##
## Degrees of Freedom: 4779 Total (i.e. Null);  4756 Residual
## Null Deviance:       5532
## Residual Deviance: 3971  AIC: 4019
```

From the result The Churn is found to be relatively more dependent and having a statistically significant dependence on factors like: tenure, tech-savvy features, contract and the payment methods.

One choice that has to be made when constructing a logistic regression model is what cutoff to use. The cutoff p is such that $\hat{\pi}_i > p \Rightarrow$ observation i is churn as yes. A typical choice is 0.5, however the context of this problem gives reason to consider a value lower than 0.5. Not specifying the customer who want to leave the company is more expensive than incorrectly categorizing a customer don't want to leave.

This report aims to find the optimal probalility cutoff which will give maximum accuracy, sensitivity and specificity.

Figure below indicating that a cutoff value 0.313 is the optimal choice, where the three curves for accuracy, specificty and sensitivity meet.

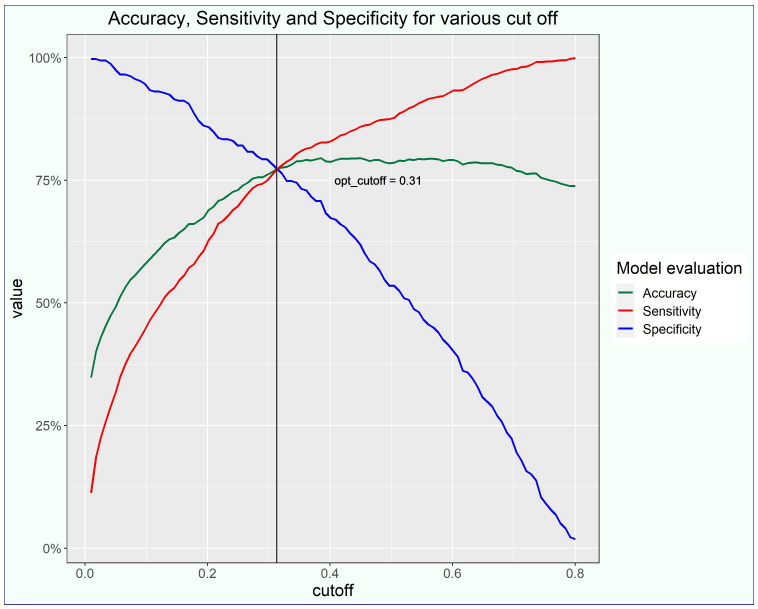

Figure 13: *Accuracy, Sensitivity and Specificity for various cutoffs in Logistic Regression Model*

We can see below the summary of the confusion matrix when the optimal cutoff of 0.313 is used.

```
## $positive
## [1] "No"
##
```

```
## $table
##           Reference
## Prediction  No Yes
##        No  678  72
##        Yes 200 246
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   7.725753e-01   4.837039e-01   7.477468e-01   7.960556e-01   7.341137e-01
## AccuracyPValue  McnemarPValue
##   1.256899e-03   1.355284e-14
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7722096            0.7735849            0.9040000
##       Neg Pred Value            Precision               Recall
##            0.5515695            0.9040000            0.7722096
##                   F1           Prevalence       Detection Rate
##            0.8329238            0.7341137            0.5668896
## Detection Prevalence    Balanced Accuracy
##            0.6270903            0.7728972
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

The following table summarizes the performance Metrics for Logistic Regression model.

Table 2: *Performance Metrics for Logistic Regression Model*

TP	FP	TN	FN	Accuracy	TPR	TNR	NPV	PPV	FPR	MCC
678	**72**	**246**	**200**	**77.25753**	**77.22096**	**77.35849**	**90.4**	**55.15695**	**22.64151**	**0.4986455**

Also, figure below is shown the ROC curve which plots the Sensitivity against the Specificity and gives us the threshold value for the model and the Area Under the Curve (AUC).

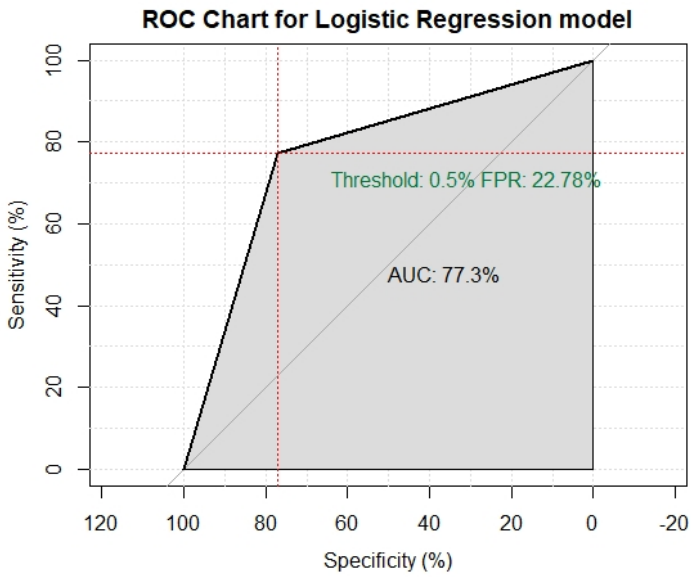

Figure 14: *ROC Chart for Logistic Regression model*

3.4 k-Nearest Neighbours Model

The second approach is to construct a k-nearest neighbours model. A typical distance metric used in kNN is the euclidean distance, however it is not suitable for this data set. Since the most features are binary it does not make sense for them to be some geographical distance apart. Instead, a common similarity measure used for binary variables is the Jaccard distance, which is used in this report. The neighbr package allows for relatively easy implementation for kNN using the Jaccard distance.

However, before the model is constructed, an optimal K parameter is chosen (the number of nearest neighbours to include in the majority of the voting process). TRAIN function using 20 bootstrap samples with replacement and 10-fold cross validation is used to select the optimal K parameter. The results are shown in Figure below, highlighting the optimal value of K = 23.

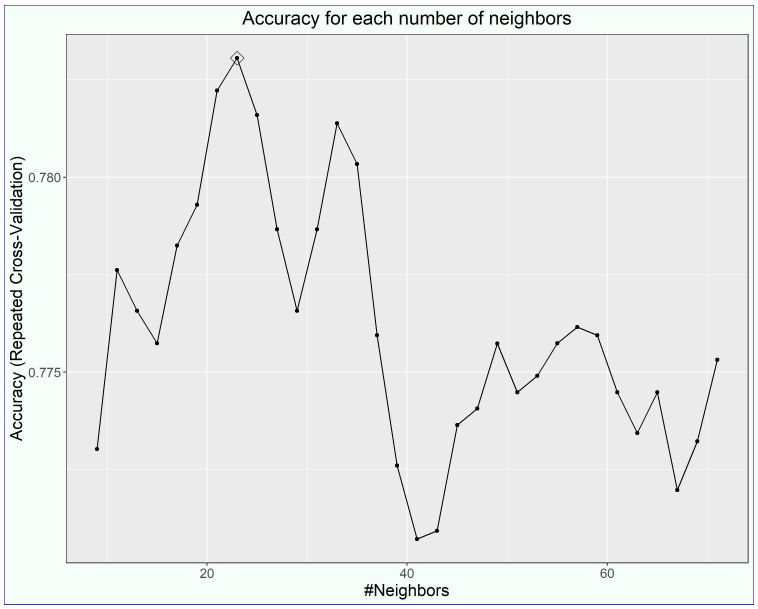

Figure 15: *Accuracy for each number of neighbors selected predictors*

Again, we aims to find the optimal probability cutoffs, which will give maximum accuracy, sensitivity and specificity. Figure below indicating that a cutoff value 0.26 is the optimal choice , where the three curves for accuracy, specificity and sensitivity meet.

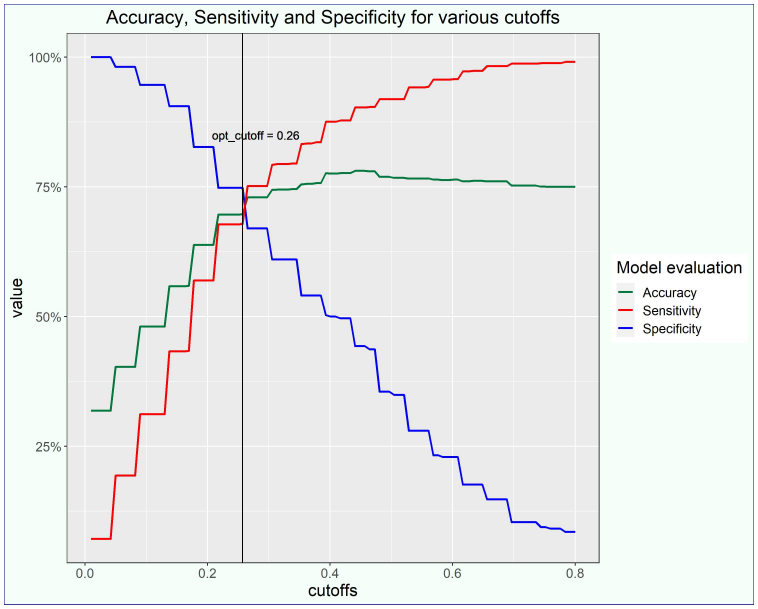

Figure 16: *Accuracy, Sensitivity and Specificity for various cutoffs in KNN Model*

We can see below the summary of the confusion matrix when the optimal cutoff of 0.257 is used.

```
## $positive
## [1] "No"
##
## $table
##           Reference
## Prediction  No Yes
##        No  596  80
##        Yes 282 238
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   6.973244e-01   3.552733e-01   6.704072e-01   7.232676e-01   7.341137e-01
## AccuracyPValue  McnemarPValue
##   9.979934e-01   4.360925e-26
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.6788155            0.7484277            0.8816568
##       Neg Pred Value            Precision               Recall
##            0.4576923            0.8816568            0.6788155
##                   F1            Prevalence       Detection Rate
##            0.7670528            0.7341137            0.4983278
## Detection Prevalence    Balanced Accuracy
```

```
##             0.5652174             0.7136216
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

The following table summarizes the performance Metrics for Logistic Regression model.

Table 3: *Performance Metrics for KNN Model*

TP	FP	TN	FN	Accuracy	TPR	TNR	NPV	PPV	FPR	MCC
596	**80**	**238**	**282**	**69.73244**	**67.88155**	**74.84277**	**88.16568**	**45.76923**	**25.15723**	**0.3807684**

Also, figure below is shown the ROC curve which plots the Sensitivity against the Specificity and gives us the threshold value for the model and the Area Under the Curve (AUC).

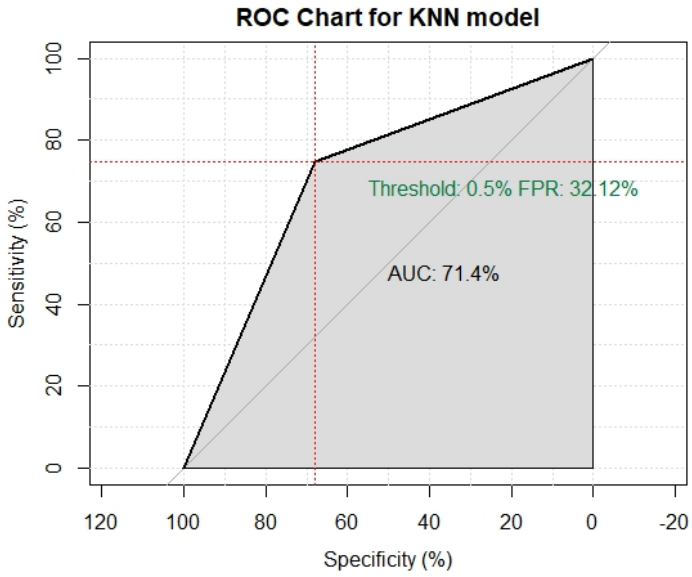

Figure 17: *ROC Chart for KNN model*

3.5 Decision Tree Model

Decision tree analysis is a classification method that uses tree-like models of decisions and their possible outcomes. This method is one of the most commonly used tools in machine learning analysis. We will use the rpart library in order to use recursive partitioning methods for decision trees. This exploratory method will identify the most important variables related to churn in a hierarchical format. Again, before the model is constructed, an optimal complexity parameter is chosen (the factor by which the models performance needs to improve by to warrant another split). TRAIN function using 20 bootstrap samples with replacement and 10-fold cross validation is used to select the optimal complexity parameter. The results are shown in Figure below, highlighting the optimal value of cp = 0.05.

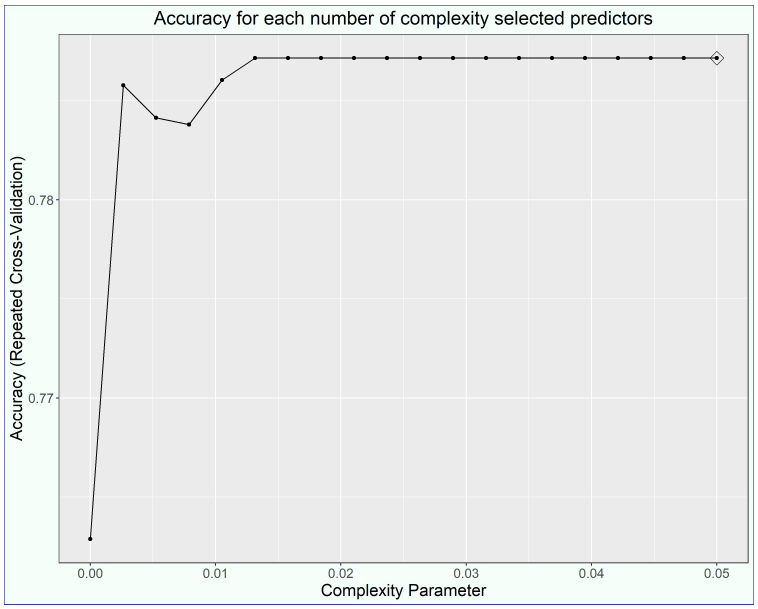

Figure 18: *Accuracy for each number of complexity selected predictors*

Following cross-validation, the train data set is used to construct a decision tree model using cp = 0.05, figure below show the plot of decision tree model.

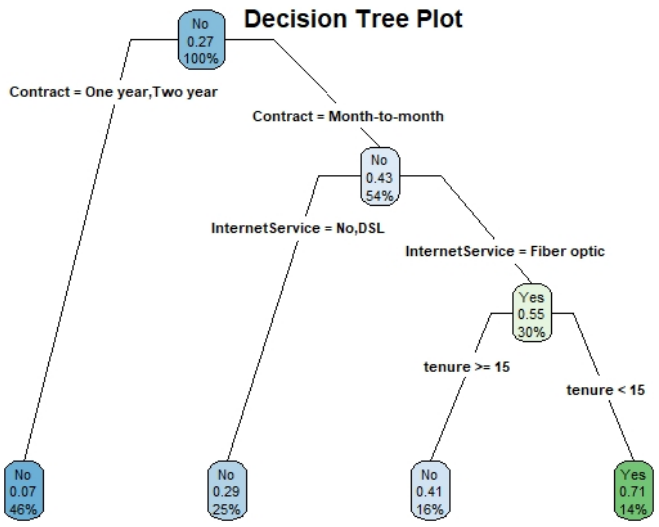

Figure 19: *Decision Tree Plot*

Again, we aims to find the optimal probability cutoffs, which will give maximum accuracy, sensitivity and specificity. Figure below indicating that a cutoff value 0.297 is the optimal choice , where the three curves for accuracy, specificity and sensitivity meet.

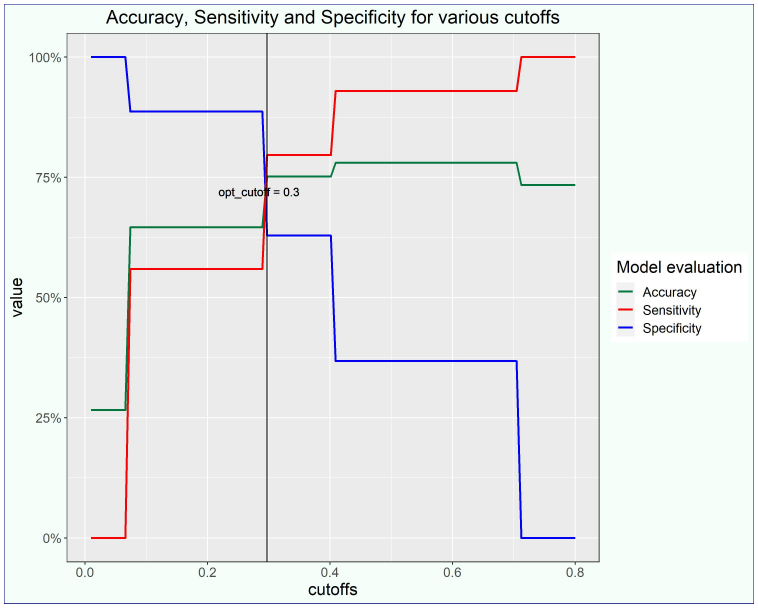

Figure 20: *Accuracy, Sensitivity and Specificity for various cutoffs in Decision Tree Model*

We can see below the summary of the confusion matrix when the optimal cutoff of 0.297 is used.

```
## $positive
```

```
## [1] "No"
##
## $table
##           Reference
## Prediction  No Yes
##        No  699 118
##        Yes 179 200
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   0.7516722408   0.4005548730   0.7261679728   0.7759325902   0.7341137124
## AccuracyPValue  McnemarPValue
##   0.0891512823   0.0004985149
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7961276            0.6289308            0.8555692
##       Neg Pred Value            Precision               Recall
##            0.5277045            0.8555692            0.7961276
##                   F1           Prevalence       Detection Rate
##            0.8247788            0.7341137            0.5844482
## Detection Prevalence    Balanced Accuracy
##            0.6831104            0.7125292
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

The following table summarizes the performance Metrics for Logistic Regression model.

Table 4: *Performance Metrics for Decision Tree Model*

TP	FP	TN	FN	Accuracy	TPR	TNR	NPV	PPV	FPR	MCC
699	**118**	**200**	**179**	**75.16722**	**79.61276**	**62.89308**	**85.55692**	**52.77045**	**37.10692**	**0.4036257**

Also, figure below is shown the ROC curve which plots the Sensitivity against the Specificity and gives us the threshold value for the model and the Area Under the Curve (AUC).

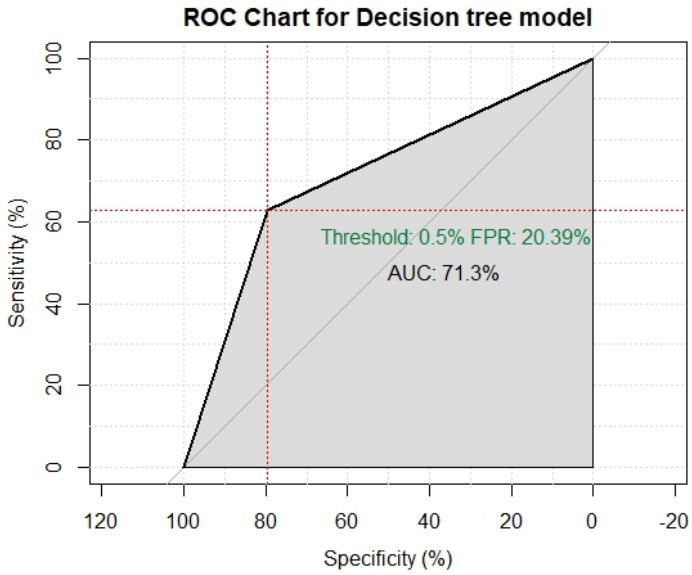

Figure 21: *ROC Chart for Decision Tree model*

3.6 The Random Forest Model

The Random Forest model is processing intensive but is known to be better at giving results than a Decision Tree. In this a number of trees are formed by selecting features at random and making trees. Then out of these the best trees are selected and a new tree is made by using the stronger features (or a group of weak features). In our modelling we use a simple technique to form a Random Forest using the package called RandomForest.

Again, before the model is constructed, an optimal mtry parameter is chosen (Number of variables available for splitting at each tree node). TRAIN function using 20 bootstrap samples with replacement and 10-fold cross validation is used to select the optimal mtry parameter. The results are shown in Figure below, highlighting the optimal value of mtry = 3.

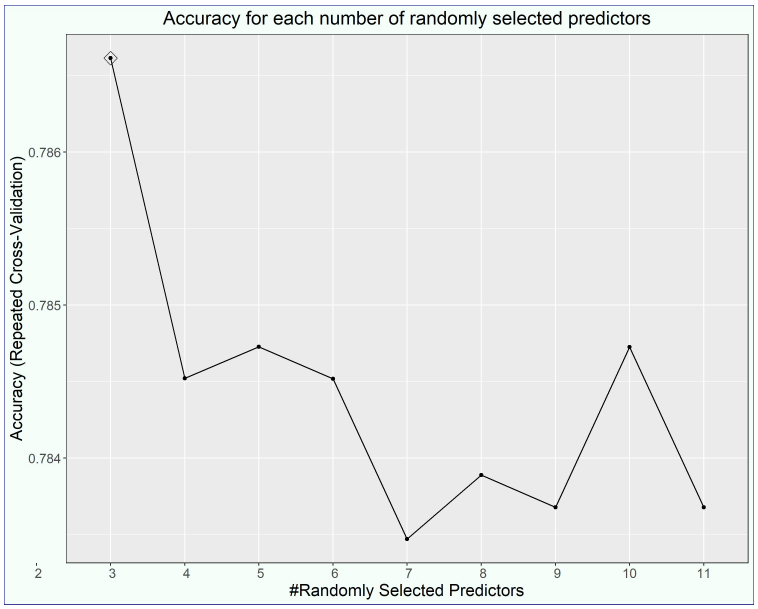

Figure 22: *Accuracy for each number of randomly selected predictors*

Following cross-validation, the train data set is used to construct a random forest model using mtry = 3. Upon plotting the model, we can see that the Out-of-Bag error rate is pretty consistent after about 150 trees and thus we fix the number of trees to the same.

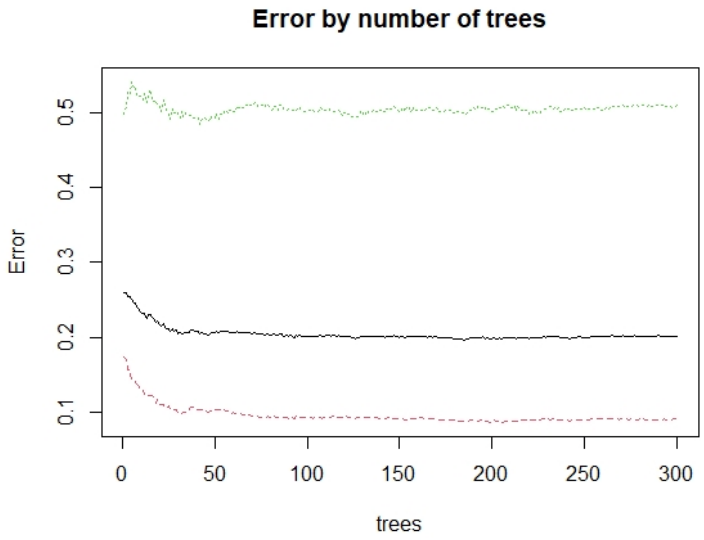

Figure 23: *Random Forest Model with decreasing OOB error as Number of Trees increases*

The importances are the mean decrease in impurity for each feature across all trees, useing the Random Forest classifier the figure below show the inportances of variables. We can see that there are some interaction variables that replaced original columns in importance.

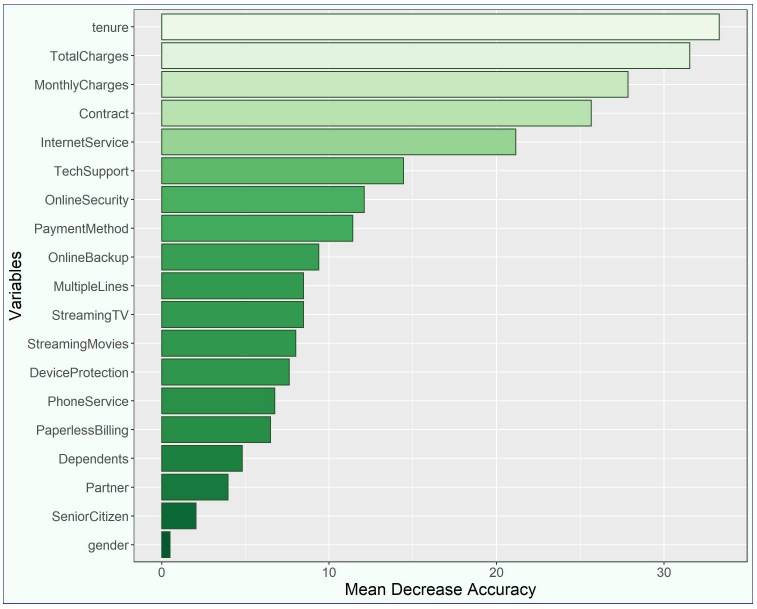

Figure 24: *The importance Variables Plot*

Again, we aims to find the optimal probability cutoffs, which will give maximum accuracy, sensitivity and specificity. Figure below indicating that a cutoff value 0.265 is the optimal choice , where the three curves for accuracy, specificity and sensitivity meet.

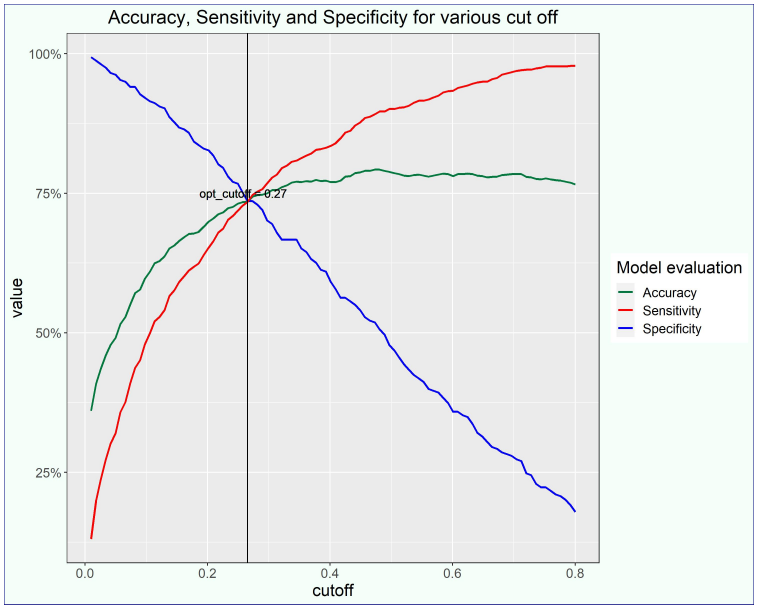

Figure 25: *Accuracy, Sensitivity and Specificity for various cutoffs in Random Forest model*

We can see below the summary of the confusion matrix when the optimal cutoff of 0.265 is used.

```
## $positive
## [1] "No"
##
## $table
##           Reference
## Prediction  No Yes
##        No  645  84
##        Yes 233 234
##
## $overall
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##   7.349498e-01   4.093119e-01   7.089603e-01   7.597786e-01   7.341137e-01
## AccuracyPValue  McnemarPValue
##   4.889838e-01   9.370021e-17
##
## $byClass
##          Sensitivity          Specificity       Pos Pred Value
##            0.7346241            0.7358491            0.8847737
##       Neg Pred Value            Precision               Recall
##            0.5010707            0.8847737            0.7346241
##                   F1           Prevalence       Detection Rate
##            0.8027380            0.7341137            0.5392977
## Detection Prevalence    Balanced Accuracy
```

```
##             0.6095318             0.7352366
##
## $mode
## [1] "sens_spec"
##
## $dots
## list()
##
## attr(,"class")
## [1] "confusionMatrix"
```

The following table summarizes the performance Metrics for Random Forest model.

Table 5: *Performance Metrics for Random Forest Model*

TP	FP	TN	FN	Accuracy	TPR	TNR	NPV	PPV	FPR	MCC
234	**233**	**645**	**84**	**73.49498**	**73.58491**	**73.46241**	**50.10707**	**88.47737**	**26.53759**	**0.4260627**

Also, figure below is shown the ROC curve which plots the Sensitivity against the Specificity and gives us the threshold value for the model and the Area Under the Curve (AUC).

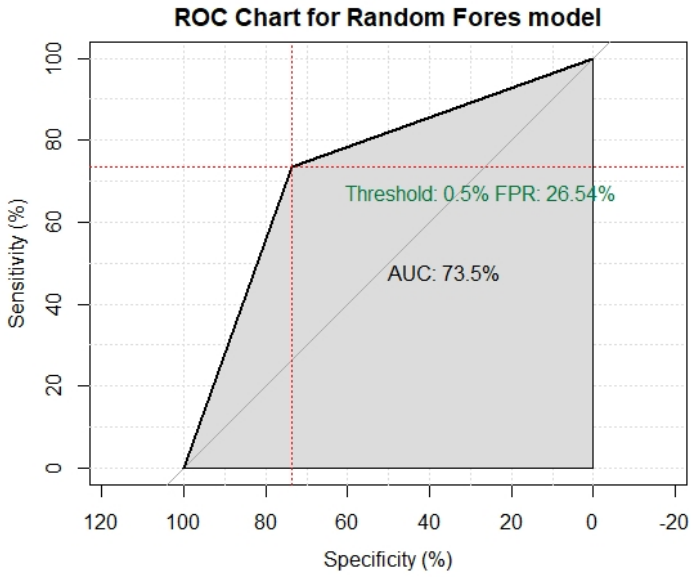

Figure 26: *ROC Chart for Random Forest model*

4. Results

In typical data science projects, it is usually to use the case that the ensemble achieves the best results. However, in this report The four ML models (Logistic Regression, k-Nearest Neighbours, Decision Tree and Random Forest) employed for the prediction of Churn evaluated on the basis of their performance in predicting the tendency to Churn from a Technical perspective.

The performance measures which used to evaluate the models are accuracy, sensitifity, specificity, its precision in predicting both Good and Bad cases,MCC, ROC and AUC characteristics.

The following table summarizes the performance Metrics for The four ML models.

Table 6: *Performance Metrics of the Four Models*

Models	TP	FP	TN	FN	Accuracy	TPR	TNR	NPV	PPV	FPR	MCC
GLM	**678**	**72**	**246**	**200**	**77.25753**	**77.22096**	**77.35849**	**90.40000**	**55.15695**	**22.64151**	**0.4986455**
KNN	**596**	**80**	**238**	**282**	**69.73244**	**67.88155**	**74.84277**	**88.16568**	**45.76923**	**25.15723**	**0.3807684**
DT	**699**	**118**	**200**	**179**	**75.16722**	**79.61276**	**62.89308**	**85.55692**	**52.77045**	**37.10692**	**0.4036257**
RF	**234**	**233**	**645**	**84**	**73.49498**	**73.58491**	**73.46241**	**50.10707**	**88.47737**	**26.53759**	**0.4260627**

Also, figure below shown the ROC curve for The four ML models which plots the Sensitivity against the Specificity and gives us the threshold value for the model and the Area Under the Curve (AUC).

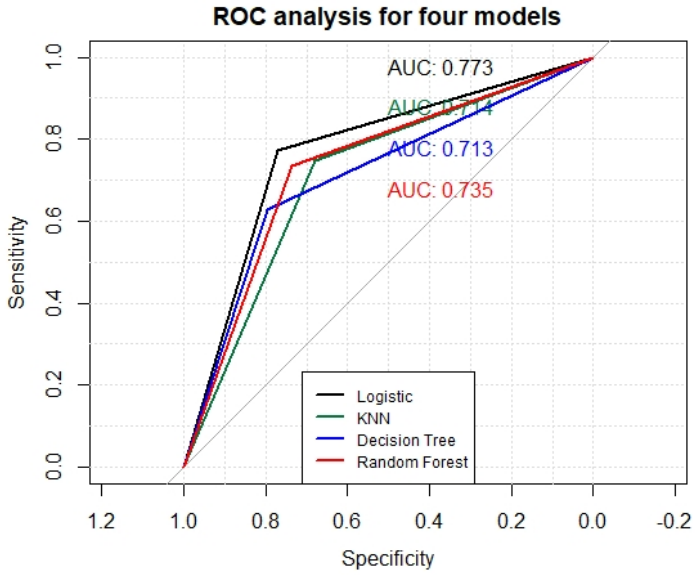

Figure 27: *ROC Chart forthe four ML models*

From above, we can conclude that Logistic Regression model was highly accurate and had the best performance Metrics. It can predict almost 77.25% accuracy, 77.22% sensitivity,

77.35% specificity, and has a MCC value of ~ 0.5, which is at par with the Neural Network model.

Also, the ROC curve which plots the Sensitivity against the Specificity gave The Area Under the Curve (AUC) = 77.3% which was high.

5. Conclusion

The main goal of this project is to explore survival analysis techniques and their influence on the customer churn rate in order to propose an action plan to improve customer retention using several techniques of Machine Learning.

We analyse the dataset and focuse on the behavior of telecom customers who are more likely to leave the platform, and then we use several techniques of Machine Learning to build a model focused on maximizing the true predictions of customers that will stay with the company.

The machine-learning models which used to predict whether the customer is at risk of churn are: Logistic Regression, k-Nearest Neighbours, Decision Tree and Random Forest.

We use various technical performance measures to compare the models with each other on the basis of : accuracy, sensitifity, specificity, MCC, ROC, AUC.

The Logistic Regression can be seen as the best model and had The the best performance Metrics and can be used as the model to build strategy guarantee stay the customers.

The telecum data set has 5,986 observations. The models would be much more reliable if it was trained and tested on a larger data set.

The data is only sampled from one telecommunications company. A significant improvement on this report would be if the data set was sampled from various telecom across the world. Thus, the final model would be useful on a global scale.

Many thanks are due to Rafael Irizarry, the course instructor of HarvardX's Professional Certificate in Data Science, and to the teaching staff who were always at hand to answer questions and queries raised by students.

This edX series has been thoroughly enjoyable and valuable. Irizarry delivered engaging lectures and provided a range of useful coding examples throughout the series.

6. Reference

[1] Kaggle Machine learning competitions *Telecom users dataset* link

[2] "Introduction to Data Science - Data Analysis and Prediction Algorithms with R", Dr. Rafael A. Irizarry link

[3] "R Markdown: The Definitive Guide", Yihui Xie, J. J. Allaire, Garrett Grolemund, 2019-06-03 link

[4] Teknomo, Kardi *Distance for Binary Variables* https://people.revoledu.com/kardi/tutorial/Similarity/BinaryVariables.html (date last accessed - 27/11/2020)

[5] Bolotov, D. (2020) *Package 'neighbr'. Classification, Regression, Clustering with K Nearest Neighbors*

[6] Therneau, T., Atkinson, B. (2019) *Package 'rpart'. Recursive Partitioning and Regression Trees*

[7] Kuhn, M., ... (2020) *Package 'caret'. Classification and Regression Training*

[8] Breiman, L., Cutler, A. (2018) *Package 'randomForest'. Breiman and Cutler's Random Forests for Classification and Regression*