Executive Summary

This report summarises the findings of the analysis of Customer Churn project taken up within our Telecommunications firm. The project aims to predict the possibility of a customer to churn by using methods of Data Analysis and Machine Learning with sound accuracy and justifies its result by showing the expected cost-benefit from following their recommendations.

The advanced data analysis from this report can be further used as a driver for another project to profile a typical customer likely to churn and thus our marketing strategies can be modified accordingly.

All processing and modelling is done by using R language (ver 3.5.1) and run on RStudio (ver 1.1.463). Several CRAN libraries and packages were used in the project and are detailed in Appendix B.

Also, the R code uses ANN which is processing exhaustive. If a user attempts to run the code, it is recommended that they use a multi-core GPU enabled machine to do so, and also employ the GPU during the processing, to avoid high processing time.

Table of Contents

Abbreviations

Abbreviations	Descriptions
ML	Machine Learning
TPR	True Positive Rate
FPR	False Positive Rate
LR	Logistic Regression
MCC	Mathew's Correlation Coefficient
TP	Subscriber expected to churn and correctly classified
TN	Subscriber classified as churn but was actually loyal
FP	Subscriber classified as loyal but was actually churn
FN	Subscriber expected as loyal and correctly classified

Introduction

Customer Churn (or attrition) is the loss of customer or clients by a typical service provider entity, like a bank, ISP, telecommunication company etc. This incurs cost to the company heavily as the cost of one customer stopping their service is multi-faced and could be considered as not only the amount of money lost as payment from the customer but also as the money required to be spent in enrolling a customer to our services. Further, among almost all related marketing researches customer acquisition is much more costly than customer retention. (Chen and Hitt, 2002) (Amaresan, 2018).

Customer Lifetime Value (CLV) here referenced as the Total Charges is another very important factor for any company when deciding its long-term goals. It is the net profit from the entire relationship with a customer that a company can expect to have. It is purely a marketing concept, but is extensively used in defining the marketing strategies and budget allocations not just on quarterly basis but on longer durations ranging from a few months to several years. The CLV prediction model can be in the form of a simple formula defining net value or a complex data model which bases its results on numerous factors of the customer and the customer-client relationship. In our report the CLV known as the Total Charges is considered to be the money charged to the customer over the period of one year.

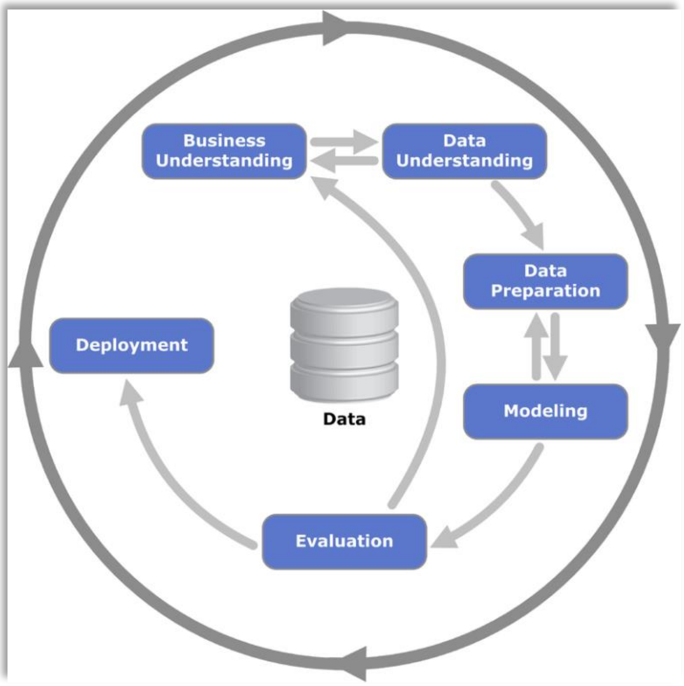


Figure 1: *The CRISP model for Data Mining (Jensen, K, Source: www.wikimedia.com)*

Several Machine Learning models were run, like Logistic Regression, Decision Tree, Random Forest and ANN (using Keras) and their results were compared and analysed. The best model was then chosen and a cost-benefit analysis was then done using the results to find the overall Uplift, the Gained Revenue and RoI that can be expected by considering the recommendations from this project.

The project follows the CRISP-DM approach (see Figure 1) and thus had several feedback loops where the data was modified post evaluating the model and then model redesigned. This however is not an evident part of this report which follows a streamlined approach to things.

Dataset (Data Understanding)

The used dataset for around 7000 customers is taken from a natural dataset and exhibits the properties of the real dataset in spirit. The telecommunications company's dataset shows details regarding gender, seniority, family structure, internet friendliness, phone connections, bills, payment methods and frequencies for the customers etc.

The dataset has some (11) missing values for *Total Charges*, although these are for rows of customers who have recently signed up for our services (within one month), and thus haven't made a payment so far.

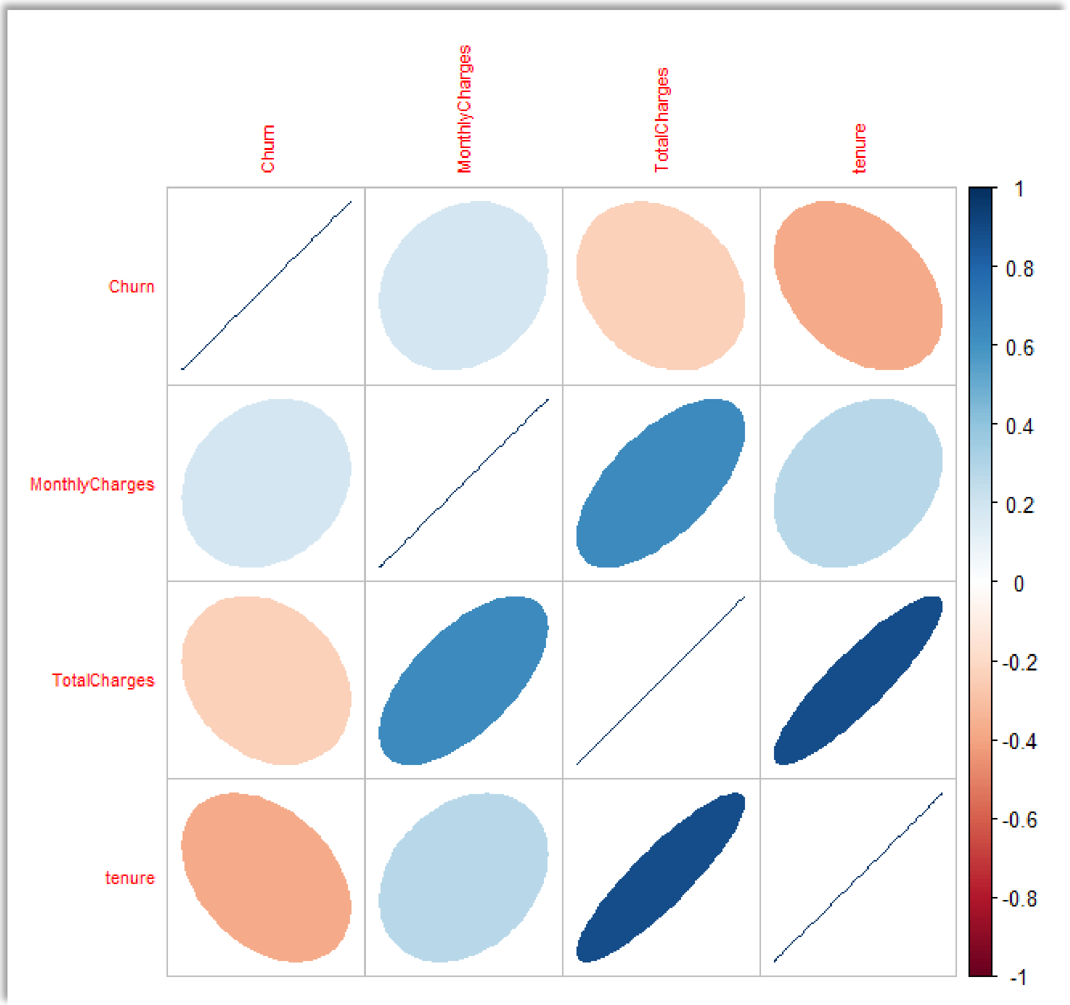


Figure 2 : *Correlation between Tenure, Charges and Churn*

Also, we have the Unique Identifier for the customers available with us as part of the dataset, known as *CustomerID*. However, it is considered that this has no relation with them Churning.

The final column is a flag telling us if the customer Churned or not.

We calculated some statistics (percentage, variance, number of customers that churned vs customers that did not churn, ...) that can be seen on Appendix C: Calculated statistics for the churn dataset. These statistics give us some relevant insights as: the percentage of customers that churn when they pay by electronic check is 45% , the ones that have a month to month contract have a churn rate of (43%), senior citizens also have a high probability of churning (42%).

We also analysed the correlation for the categorical and numeric fields. The correlation for the numerical fields is illustrated on Figure 2. This correlation table shows that the correlation between the fields *tenure* and *total charges* is very close to 1, the correlation between *monthly charges* and *total charges* is also very high, therefore it is recommended to combine these fields into one to avoid multicollinearity issues when applying the logistic regression model. The *TotalChargesUpdated* is the combined field we will use for our analysis:

$$TotalChargesUpdated = \frac{TotalCharges + MonthlyCharges \times tenure}{2}$$

The correlation for the categorical fields is shown on **Figure 3**, from this correlation matrix we can see a high level of correlation between the fields *InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.* Theres is also a correlation close to one between the fields *MultipleLines* and *PhoneService* indicating that these two field should be combined into one. Another interesting factor from this correlation matrix is to note that the fields *Contract, PaperlessBilling, and OnlineSecurity* are high correlated with the target field Churn, indicating that these fields will be relevant for our models.

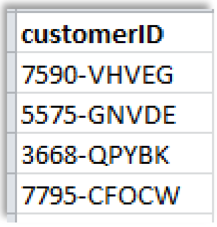
Technical Perspective: Data Preparation

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

Feature Selection (Unique Identifier Handling)

The Unique Identifier *CustomerID* was removed from the dataset as one of the assumptions of the project that this was a random ID provided to a customer and thus would have no relation with the tendency to Churn.

This could be slightly wrong, as we see that the *Customer ID* is alphanumeric and is probably encoded according to some details (maybe region, age, chosen plan, start date or another), and the tendency to churn could depend on one (or more) of these factors. For example, if a lot of people signed up when a new exciting scheme was out, they might churn if the benefits are revoked. However, no such information is provided so CustomerID allocation is considered to be random in nature and hence removed from the analysis totally.



customerID
7590-VHVEG
5575-GNVDE
3668-QPYBK
7795-CFOCW

Figure.3: *CustomerID format (which could be some kind of code)*

Handling Missing Values

Customers with tenure = 0 are removed from the analysis as they lead on to missing values in the data. Also, there are only a total of 11 such customers so removing them does not considerably hamper our dataset size.

Ordinal Discretisation (or Feature Creation for Tenure Group)

Post removing zero tenure customers, the high variability of the tenure is taken care of. The monthly stack of lifecycle is changed to yearly basis and the customers are thus categorised on the basis of tenure = number of years (1 or 2 or 3 or 4 or 5 years).

Also, another relation which was found to be quite significant was the relation between (i) Churn & Total Charges, and (ii) Churn & (Tenure*Monthly Charges). The charges in both these cases are quite similar from their business perspective and so a new field called TotalChargesUpdated is created by taking the mean of both these values.

$$TotalChargesUpdated = (TotalCharges + MonthlyCharges * tenure)/2$$

Encode Categorical (for SeniorCitizen)

The SeniorCitizen field has its values as 0/1, while all the other binary ones are as Yes/No. So to have a similar structure this field is updated to the same style.

Data Aggregation

For fields such as *OnlineBackup, DeviceProtection, TechSupport, StreamingTV,* and *StreamingMovies,* the data was changed to a simple Yes/No values so as to abstract the information in a concise manner.

Ordinal Transformation

The correlation between Churn and log(TotalCharges) was found to be higher (-0.24) than the one between Churn and TotalCharges (-0.20). So, in order to benefit from this in our ML models, the field TotalCharges was updated as the log(TotalCharges).

Normalisation

This log(TotalCharges) field is normalised when employed in the Keras model for the ANN, as it is one of the pre-requisites of the Neural Network algorithm. It is scaled, centred and normalised to enhance the speed and accuracy of the ANN.

Splitting

The training and test data set is chosen by random sampling using Bernoulli distribution (Sean X. Chen and Jun S. Liu, 1997) which is widely used for probability distribution of random variables.

The data is split in the ratio of 70:30, which although does not have a strong theoretical backing to it, but is widely employed in the industry and considered by many as a suitable fit for splitting the data into testing and training subsets.

Technical Perspective: Modelling

Post preparing the data we can now move on to the step of Data Modelling, where we have implemented several models and compared them with each other on the basis of various technical performance measures, inter alia accuracy, precision (good), precision (bad), MCC, ROC and AUC.

The methods used for prediction of Customer Churn are:

(i) Logistic Regression

(ii) Decision Tree

(iii) Random Forest

(iv) Multi-Layer Neural Network (using Keras Library)

Logistic Regression

Logistic Regression was the primary approach followed for the dependency of a customer to churn on the other factors. The Spearman Correlation matrix was generated for the finding out features which had an impact on the value of Churn.

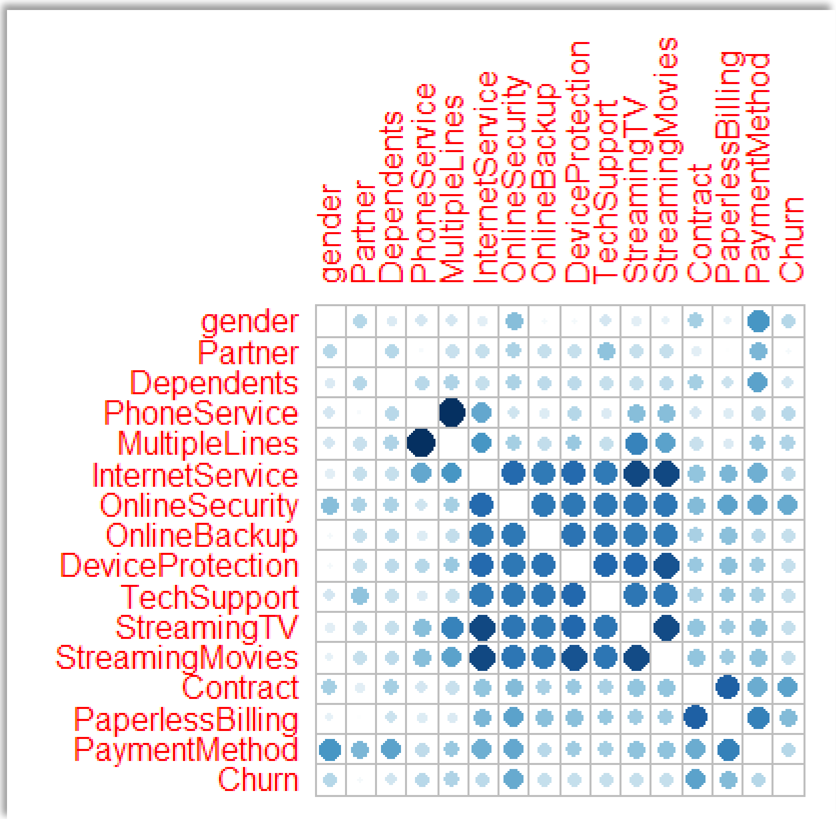
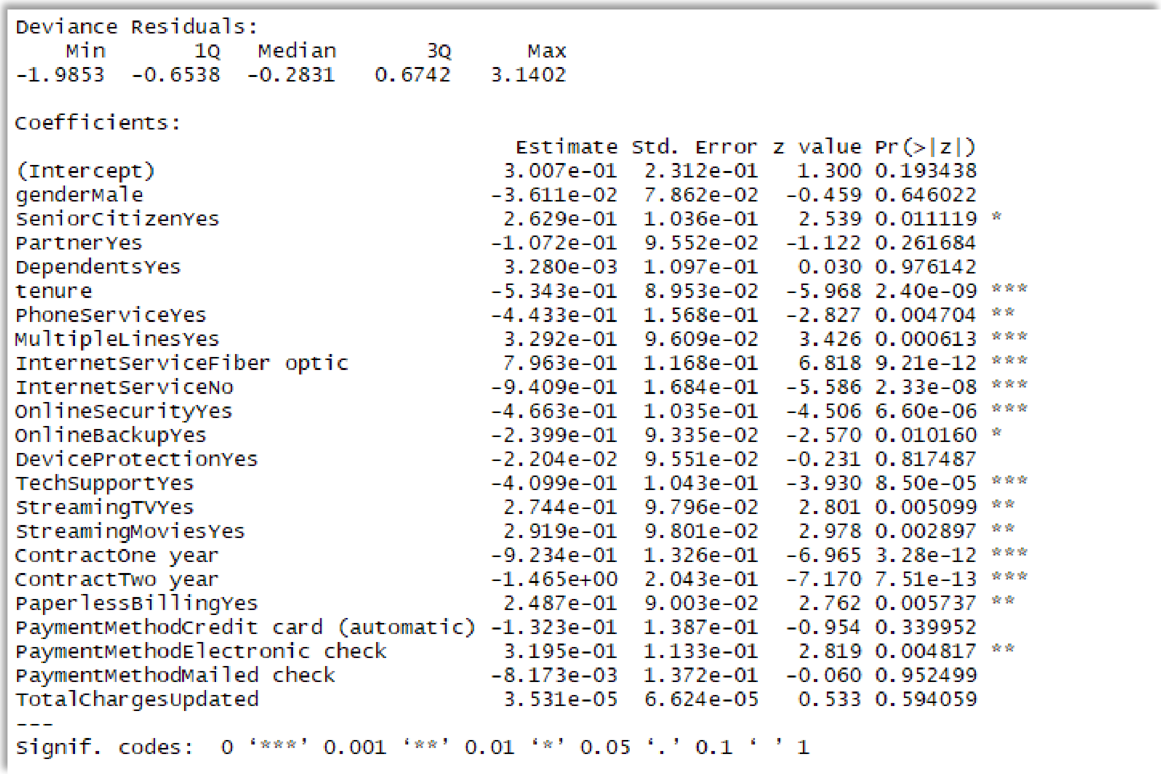


Figure 3 *Spearman correlation for the categorical fields*

The major reason as to why Logistic Regression can be considered to be a suitable approach for this project is that the target variable is Categorical in nature, which makes this approach more suitable.

The results from the LR model by employing all features to predict Churn can be seen below:



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9853  -0.6538  -0.2831   0.6742   3.1402

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          3.007e-01  2.312e-01   1.300 0.193438
genderMale                          -3.611e-02  7.862e-02  -0.459 0.646022
SeniorCitizenYes                     2.629e-01  1.036e-01   2.539 0.011119 *
PartnerYes                          -1.072e-01  9.552e-02  -1.122 0.261684
DependentsYes                        3.280e-03  1.097e-01   0.030 0.976142
tenure                              -5.343e-01  8.953e-02  -5.968 2.40e-09 ***
PhoneServiceYes                     -4.433e-01  1.568e-01  -2.827 0.004704 **
MultipleLinesYes                     3.292e-01  9.609e-02   3.426 0.000613 ***
InternetServiceFiber optic           7.963e-01  1.168e-01   6.818 9.21e-12 ***
InternetServiceNo                   -9.409e-01  1.684e-01  -5.586 2.33e-08 ***
OnlineSecurityYes                   -4.663e-01  1.035e-01  -4.506 6.60e-06 ***
OnlineBackupYes                     -2.399e-01  9.335e-02  -2.570 0.010160 *
DeviceProtectionYes                 -2.204e-02  9.551e-02  -0.231 0.817487
TechSupportYes                      -4.099e-01  1.043e-01  -3.930 8.50e-05 ***
StreamingTVYes                       2.744e-01  9.796e-02   2.801 0.005099 **
StreamingMoviesYes                   2.919e-01  9.801e-02   2.978 0.002897 **
ContractOne year                    -9.234e-01  1.326e-01  -6.965 3.28e-12 ***
ContractTwo year                    -1.465e+00  2.043e-01  -7.170 7.51e-13 ***
PaperlessBillingYes                  2.487e-01  9.003e-02   2.762 0.005737 **
PaymentMethodCredit card (automatic) -1.323e-01  1.387e-01  -0.954 0.339952
PaymentMethodElectronic check        3.195e-01  1.133e-01   2.819 0.004817 **
PaymentMethodMailed check           -8.173e-03  1.372e-01  -0.060 0.952499
TotalChargesUpdated                  3.531e-05  6.624e-05   0.533 0.594059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure.5: *Output of Logistic Regression Model*

The features which were found to have the most effect on Churn and were also statistically significant were:

A new model can also be made by using a select group of features for better results.

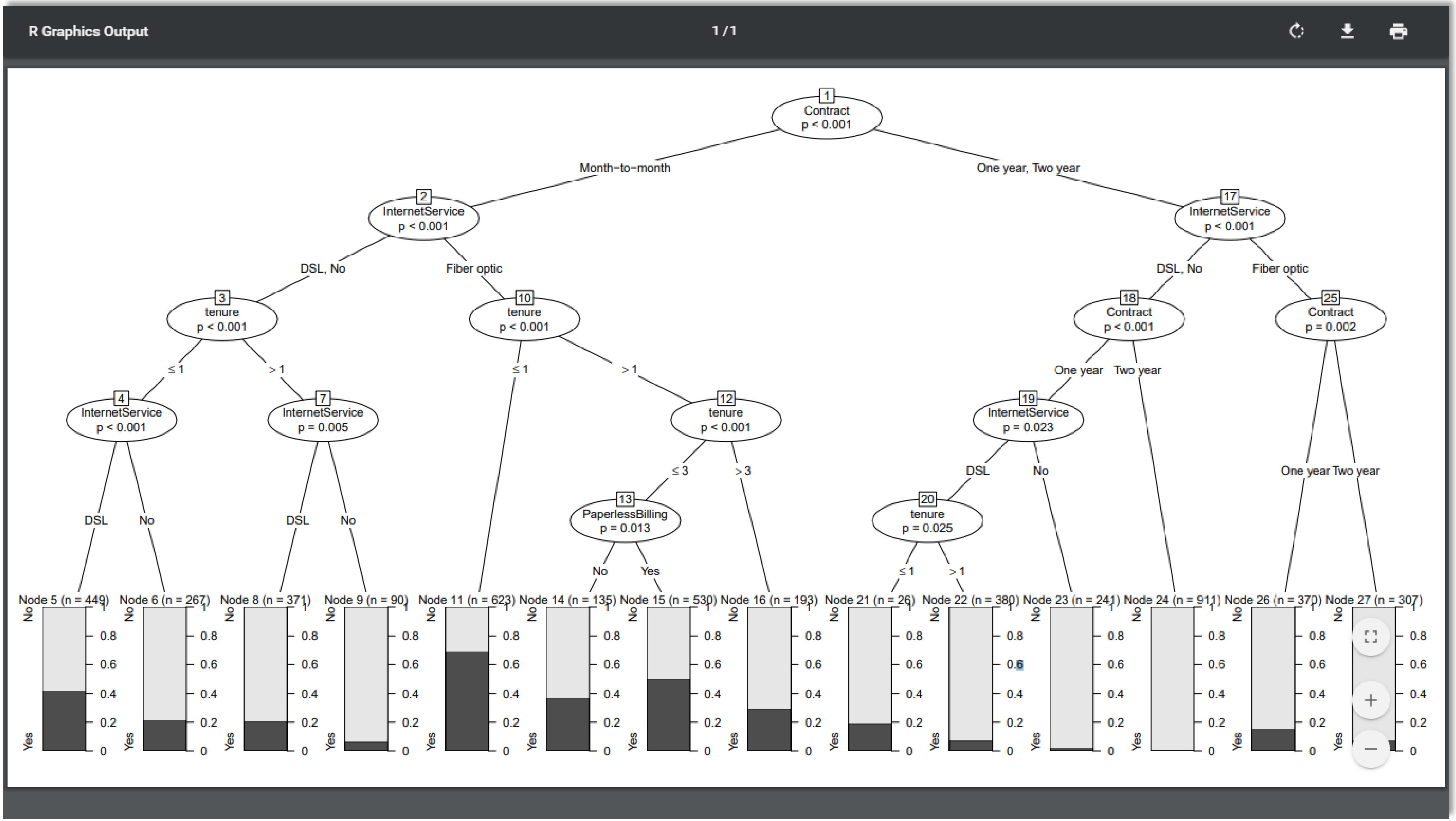
Decision Tree

Decision Tree is a strong approach for Supervised Learning and is particularly good at dealing with target results which are categorical. Another positive thing about Decision Trees is that they are particularly easy to interpret and can be converted to business rules with much finesse.

After few repetitive trials a decision tree using select four features was found to have a significantly high performance and was also not very wide. These features are:

Contract || Tenure || Paperless Billing || Internet Service

Figure.6: *Decision Tree (on following page)*


R Graphics Output
1 / 1
1 Contract p < 0.001
Month-to-month
One year, Two year
2 InternetService p < 0.001
DSL, No
Fiber optic
3 tenure p < 0.001
≤ 1
> 1
4 InternetService p < 0.001
DSL
No
7 InternetService p = 0.005
DSL
No
10 tenure p < 0.001
≤ 1
> 1
12 tenure p < 0.001
≤ 3
> 3
13 PaperlessBilling p = 0.013
No
Yes
17 InternetService p < 0.001
DSL, No
Fiber optic
18 Contract p < 0.001
One year
Two year
19 InternetService p = 0.023
DSL
No
20 tenure p = 0.025
≤ 1
> 1
25 Contract p = 0.002
One year
Two year
Node 5 (n = 449)
Node 6 (n = 267)
Node 8 (n = 371)
Node 9 (n = 90)
Node 11 (n = 623)
Node 14 (n = 135)
Node 15 (n = 530)
Node 16 (n = 193)
Node 21 (n = 26)
Node 22 (n = 380)
Node 23 (n = 241)
Node 24 (n = 911)
Node 26 (n = 370)
Node 27 (n = 307)
No
Yes
1
0.8
0.6
0.4
0.2
0

The performance of the Decision Tree model was found to be good, but better results were found with other models. This is why the Decision Tree model was left as is and not pruned further as this would have only decreased the performance of the model even more. The more direct business rules which could have been formulated hence, would not have been accurate enough and thus not given sufficiently high results in the real world, where every bad decision add to revenue lost or extra cost incurred.

Random Forest

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. 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.

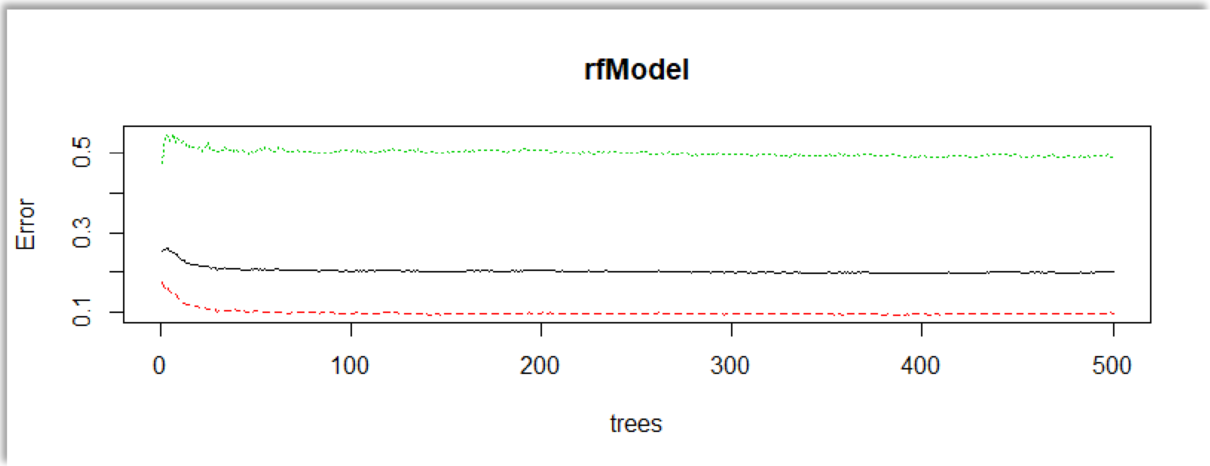


Figure.7: *rfModel with decreasing OOB error as Number of Trees increases*

There were trials done to enhance the accuracy of the random forest by tuning and then fitting the model. However, not significant enhancement of accuracy (>2%) and other performance measures was seen by the process. Thus, the model is left in its raw form as there are other better predictors available anyway.

Multi-Layer Neural Network (using Keras Library)

A multi-layer, multi-perceptron deep neural network is also employed to predict the tendency to Churn. Neural Network is used because it can give more precise and accurate results and predict with better accuracies. Out of the several NN packages available, the open-source library, Keras, originally written for Python is implemented (using its image for R).

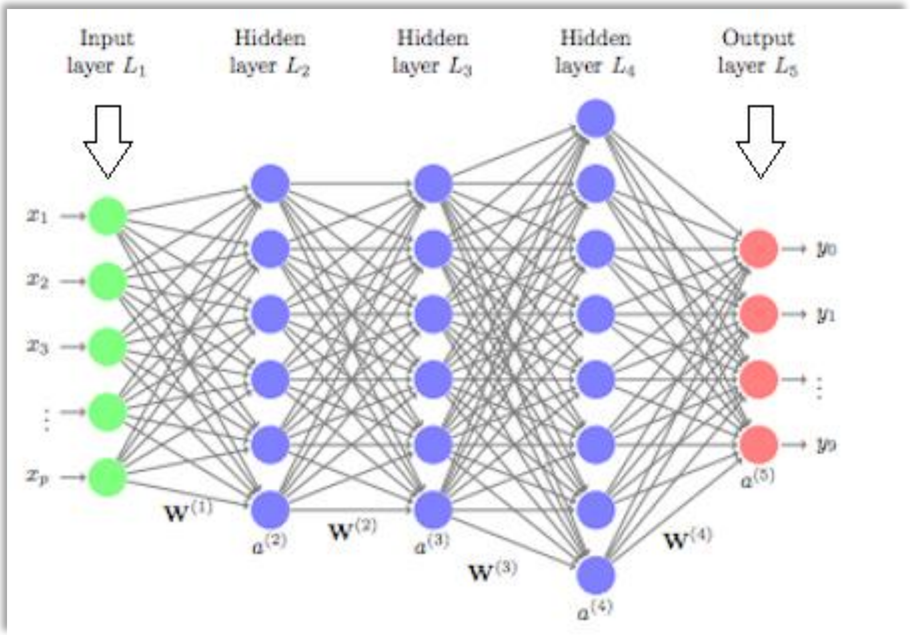


Figure.8: *Architecture of a Deep Neural Network*

The ANN works well with discreet values, all the selected attributes are discreet except "TotalChargesUpdated". The {"Yes","No"} categorical fields were converted to dummy variables, enhancing the speed of the ANN algorithm which is often a major challenge while using Deep Neural Networks.

The configuration parameters used to run the Keras model are indicated in the following table:

Parameter	**Method**
Initial weights	Selection of parameters. Kernel_initializer='uniform'
Activation function	Non-Linear (sigmoid)
Avoid local minima	Creating of a dropout layer to avoid overfitting (weight =0.1)

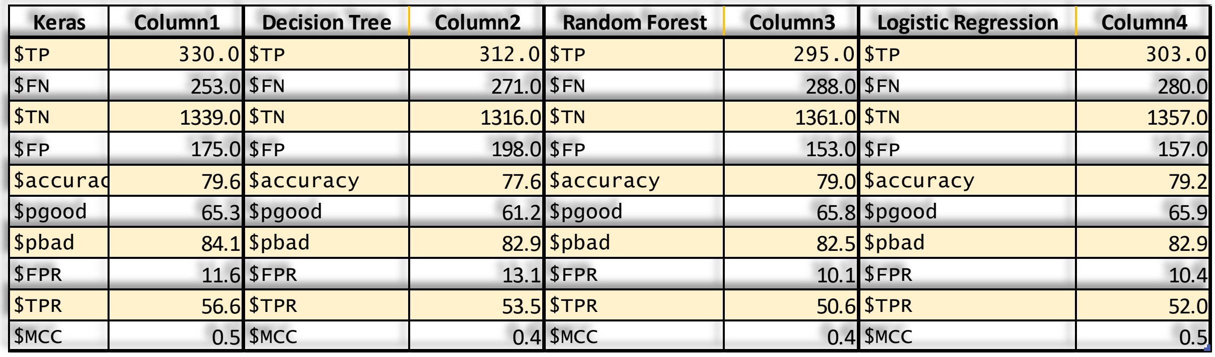
Hidden Units	2 with 16 Neurons each.
Epochs	20

Table.1: *Parameters of the Keras Model*

Technical Perspective: 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 30% unseen data which was split during the Data Preparation stage previously.

The performance measures which are looked at are the models accuracy, its precision in predicting both Good and Bad cases, the Matthew's Correlation Coefficient, ROC and AUC characteristics. Also, the confusion matrix is generated for all models to find out the performance measures such as TPR, FNR etc. These and the number of TP, FP, TN and FN from the predicted class help in the Cost Benefit Analysis which is done to evaluate the project from a Business Perspective.



Keras	Column1	Decision Tree	Column2	Random Forest	Column3	Logistic Regression	Column4
$TP	330.0	$TP	312.0	$TP	295.0	$TP	303.0
$FN	253.0	$FN	271.0	$FN	288.0	$FN	280.0
$TN	1339.0	$TN	1316.0	$TN	1361.0	$TN	1357.0
$FP	175.0	$FP	198.0	$FP	153.0	$FP	157.0
$accurac	79.6	$accuracy	77.6	$accuracy	79.0	$accuracy	79.2
$pgood	65.3	$pgood	61.2	$pgood	65.8	$pgood	65.9
$pbad	84.1	$pbad	82.9	$pbad	82.5	$pbad	82.9
$FPR	11.6	$FPR	13.1	$FPR	10.1	$FPR	10.4
$TPR	56.6	$TPR	53.5	$TPR	50.6	$TPR	52.0
$MCC	0.5	$MCC	0.4	$MCC	0.4	$MCC	0.5

Table.2 *Performance Metrics of the Four Models*

(back to Business Perspective: Evaluation)

Logistic Regression

The result of the Logistic Regression model, which was run on all the variables, is shown below in Table 3. They further establish the idea from the visualisations earlier in the project as 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.*

This model is highly accurate and can predict with almost 80% accuracy. It's precision to classify the bad cases is also very high ~83% and has a MCC value of 0.5, which is at par with the Neural Network model.

Also, is shown the ROC curve which plots the Sensitivity against the Specificity and gives us the threshold value for the model. The Area Under the Curve (AUC) = 70.8% is quite high.

Field Relevance for the Logistic Regression model was calculated using the *VarImp* function in caret package of R and its results are shown below:

Field	Field relevance [%]
Contract = 'Two year'	100%
Contract = 'One year'	97%
InternetService= 'Fiber optic'	95%
tenure	83%
InternetService='No'	78%
OnlineSecurity='Yes'	63%
TechSupport='Yes'	55%
MultipleLines='Yes'	48%
StreamingMovies='Yes'	41%
PhoneService='Yes'	39%
PaymentMethod='Electronic check'	39%
StreamingTV='Yes'	39%
PaperlessBilling='Yes'	38%
OnlineBackup='Yes'	36%
SeniorCitizen='Yes'	35%
Partner='Yes'	15%
PaymentMethod='Credit card' (automatic)	13%

Table.3: *Relevance of fields by LR*

This information helps in tweaking the models and making final inferences.

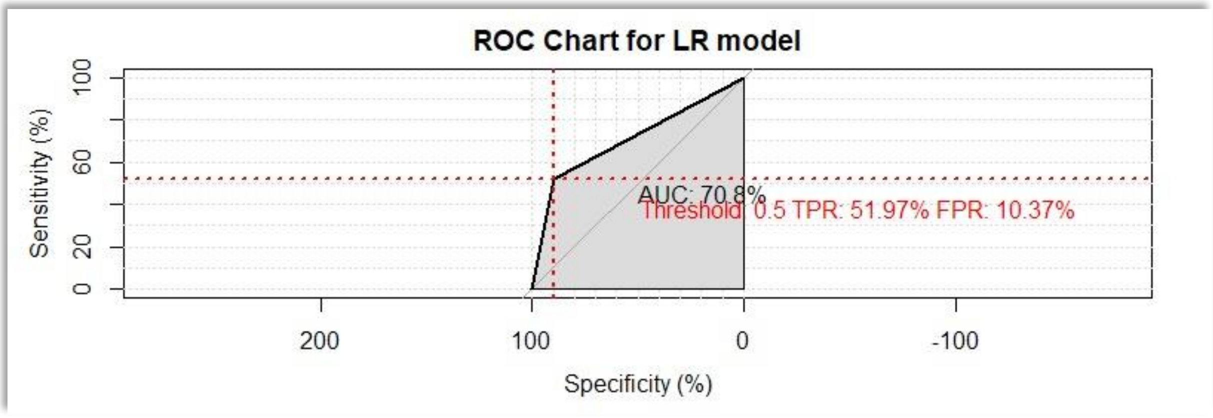


Figure.9: *ROC and AUC of the LR Model*

Decision Tree

The Decision Tree model based on a select few features (to keep it under control) is one which giving us good performance overall. Its results are comparable with the other models and it also helps in deriving rules which are easy to understand by the business. However, since the other models are giving us better results already, this is left as is.

The ROC and AUC of this model is shown below in Figure 9.

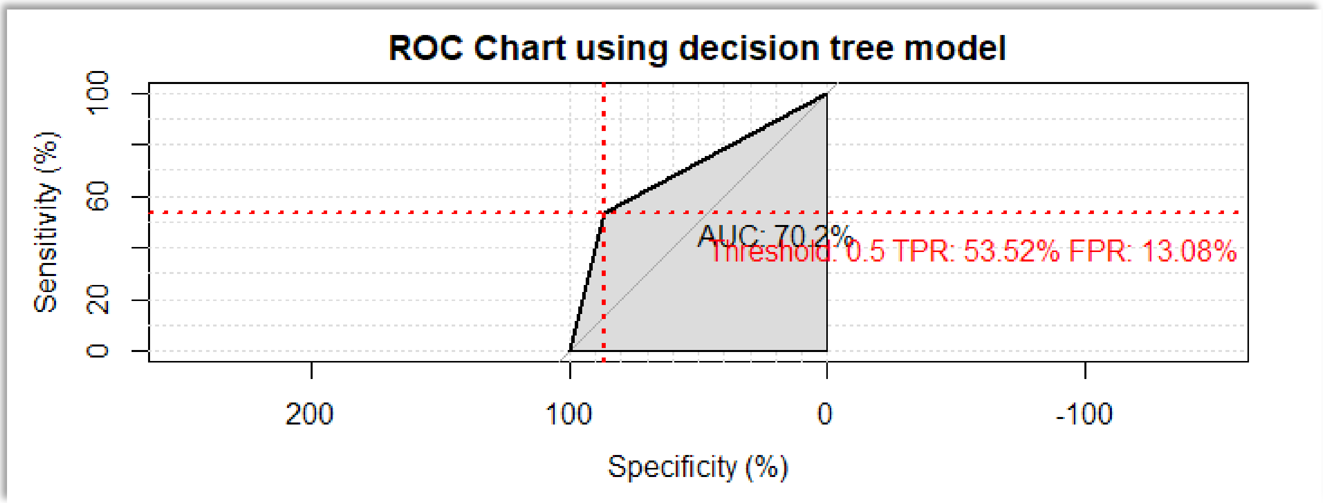


Figure.10: *ROC and AUC of Decision Tree*

Random Forest

Since good results were obtained from the Decision Tree, the Random Forest model was the next choice of implementation. And as would be expected the model performs either equivalent to or significantly better than the Decision Tree model employed before. The ROC of RandomForest is however not as impressive and its AUC is almost the same.

Artificial Neural Network

The ANN model is the seemingly the best one amongst our four models and performs better in all aspects ranging from accuracy & precision to ROC & AUC, albeit only slightly. But small enhancements in every metric lead to a better performing model overall.

The dropout layer created during the model creation is also very significant in enhancing the accuracy when testing against the test data, as it can substantially avoid over-fitting.

The model's results were found to be quite similar after the traversal of some 20 epochs, so the final model used was set to run for 20 with validation_split = 0.30

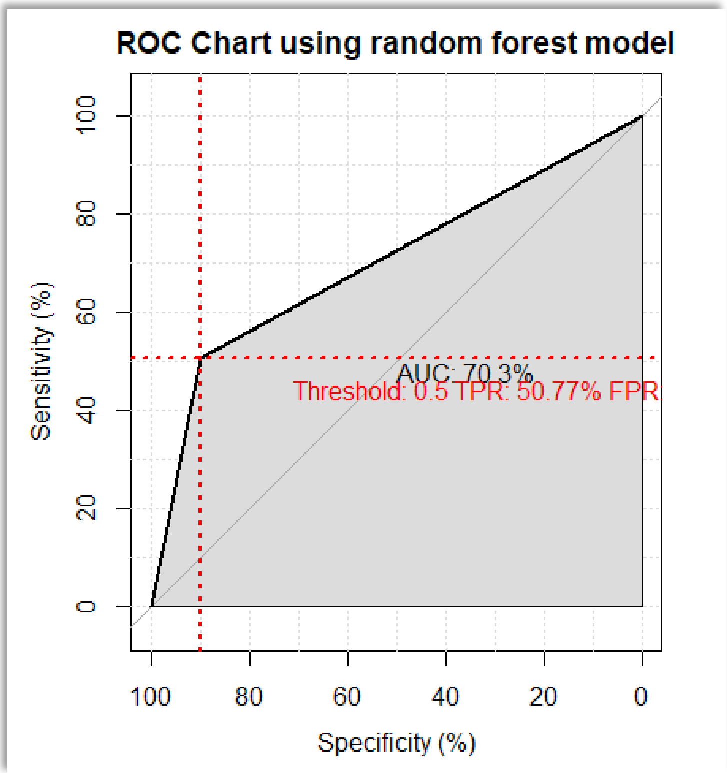


Figure.11: *ROC and AUC of Random Forest Model*

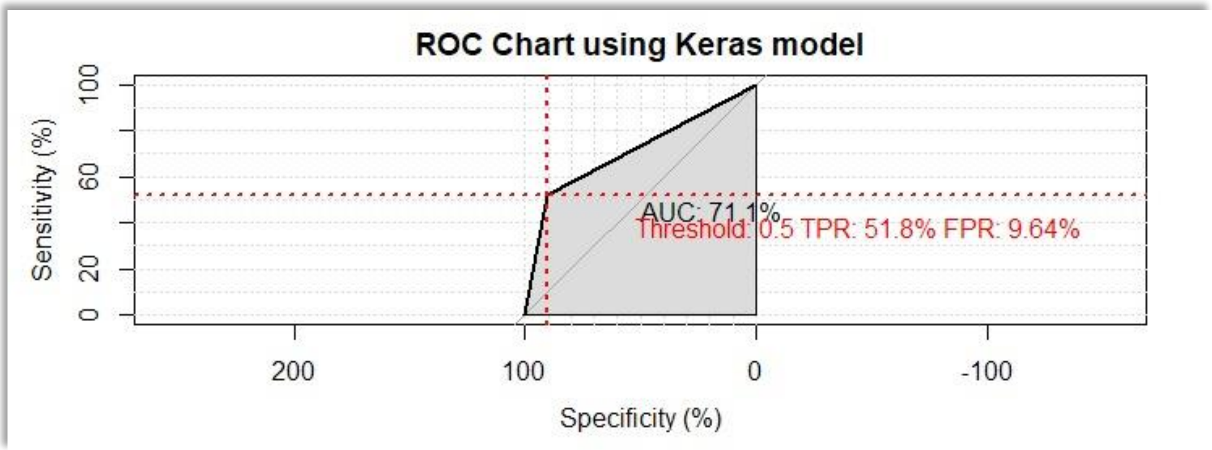


Figure.12: *ROC and AUC of ANN Model (using Keras)*

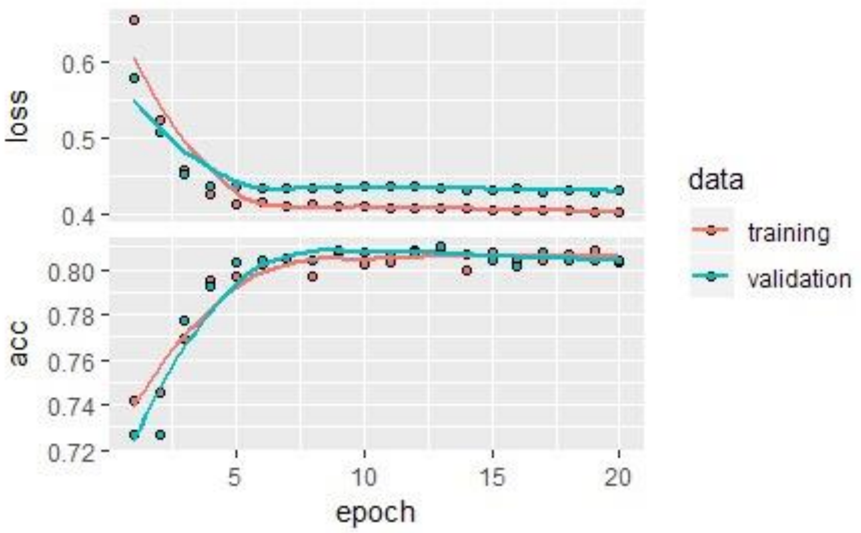


Figure.13: *The Change in loss and accuracy over time (or number of epochs)*

Business Perspective: Churn

A balanced and smart business is one which knows where the customer attrition is coming from and it can then modify its approaches of customer enticement and marketing accordingly so that the important customers do not leave their services. Telecom sector of the world is one of the most competitive of all where several local and international players bite at each other to gain the most customers and to try and port people from the other companies. There is always a finite customer base and all companies are trying to become the leader by having the most customers.

Thus the idea of churn analysis and finding out the primary drivers is an important part of decision making in the telecom sector. The factors can be identified either by surveys or by analysing the aspects of the persons who churn; this also would help in profiling a customer likely to churn.

Another major factor why telecom companies try and minimise the customer churn is due the fact that the cost to gain a new customer is several (~5-6) times higher than the cost to retain one by offering enticements. In our analysis the cost to win-over a new customer from a competitor is considered as $750, and it is assumed that for retaining a customer who is likely to churn we have to give them benefits to the tone of 10% of their Total Charges. One more assumption of this project, to limit the scope of analysis, is to consider CLV (Customer Lifetime Value) or Total Charges as their one year charges, instead of the real lifetime value. This allows in making the analysis and final decision making easier.

Business Perspective: Methodology

The project was done by following the CRISP-DM (Cross-Industry Standard Process for Data Mining) process model, and moving forward/backward was done several times upon finding a newer insight in the data or to try another approach which included introduction of newer variables, modifying data-split, number of trees in random forest, number of neurons etc.

A major focus was laid on visualisations and data exploration in the project, as this gives several insights and helps in quickening the machine learning process by making it more direct from the very start itself.

Shown below, is the comparison of the dependency of Churn of both TotalCharges and log(TotalCharges), which shows a more flattened 'Yes' dependence when the feature is used in its logarithmic form.

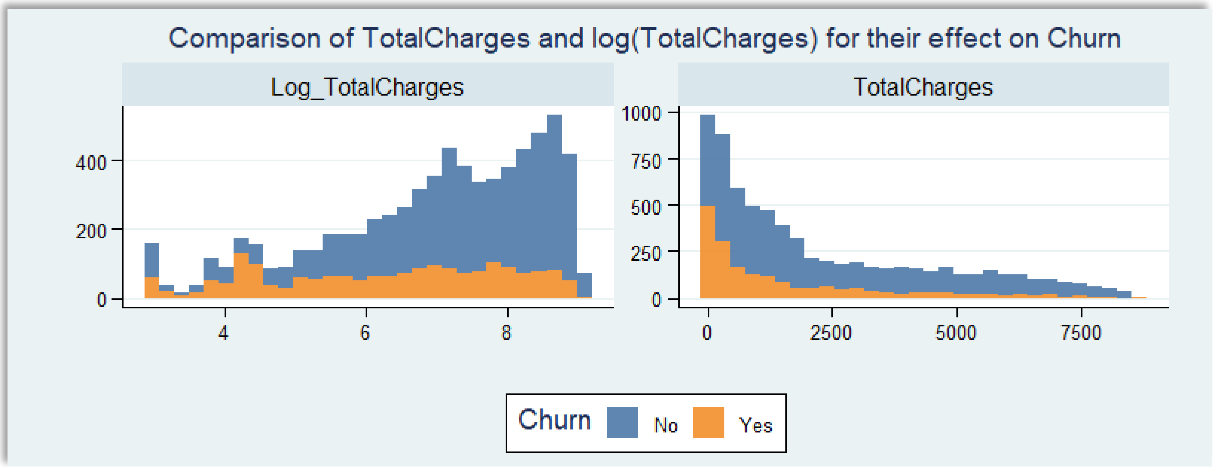


Figure.14: *Comparison of Total Charges and log(Total Charges) on Churn*

Churn is also visualised as against the Charges and Tenure of the customers and it can be seen that if a customer has a higher charge overall across the year they are less likely to stray from our services. And that people often start to use our service and being dissatisfied leave them. Thus there is a very high churn tendency in the low tenure range. Also, there is significantly high churn number of people who are slightly above the minimum MonthlyCharge and also at the maximum MonthlyCharge. This can be attributed to people who expect low bills and left when they received a slightly higher one and to the people who got a very high bill ($100) and decided to move to low cost providers. Our lowest billed customers are the most loyal, as can be seen from the first chart below with a blue spike of 900+.

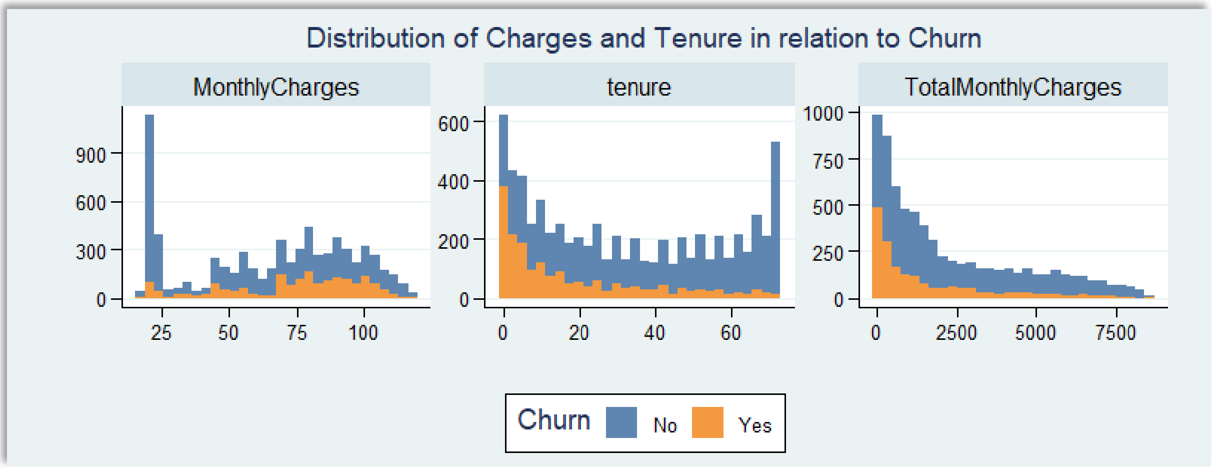
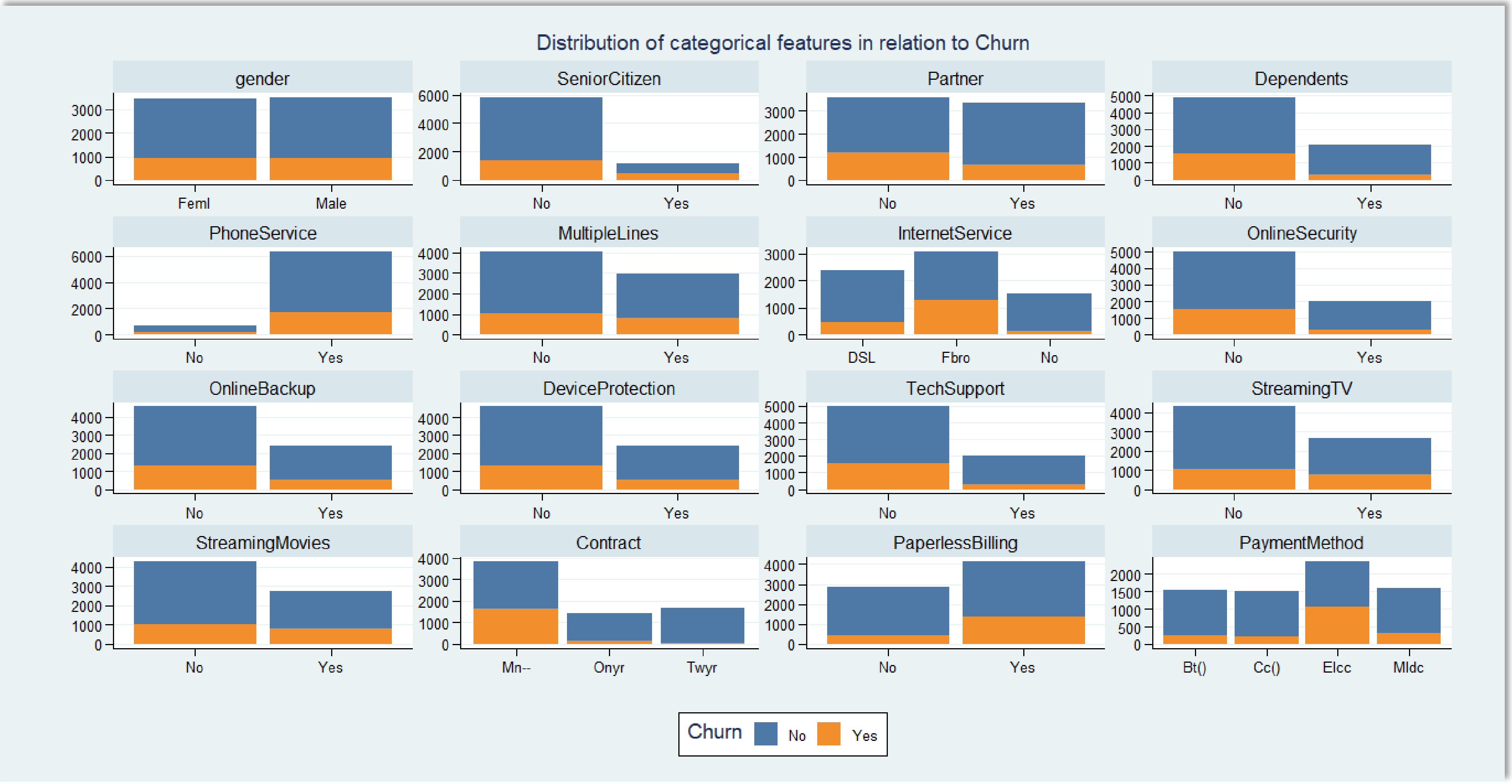


Figure.15: *Churn against Tenure and Charges*

Similar visual analysis is also done with other categorical fields as shown on the next page. Some insights from this are that:

- Gender is not a factor for a person to leave
- Senior citizens are more loyal to us
- People with dependents are also more stable and less likely to churn
- Tech-savvy people are probably always looking out for better options and are often the ones to churn. This idea is further established by the chart which shows that the number of people having Tech Support is lower, the tech savvy people would search for things themselves on the internet and try to resolve while the those that are not would call in.
- People using Fiber Optic service are known to Churn a lot, so if our company offers these services we should critically look at them as they might require some overhauling.
- If there is someone taking a contract with us, they can be considered as satisfied with our services and there are minimal chances for them to churn.
- People doing an Electronic transfer payment of their bills, probably keep track of expenses and are quick to jump to a low cost provider, on the other hand automatic payment persons are more lenient in their expenses.

Figure.16: *Churn against Various Categorical Fields (on following page)*


Distribution of categorical features in relation to Churn
gender
3000
2000
1000
0
Feml
Male
SeniorCitizen
6000
4000
2000
0
No
Yes
Partner
3000
2000
1000
0
No
Yes
Dependents
5000
4000
3000
2000
1000
0
No
Yes
PhoneService
6000
4000
2000
0
No
Yes
MultipleLines
4000
3000
2000
1000
0
No
Yes
InternetService
3000
2000
1000
0
DSL
Fbro
No
OnlineSecurity
5000
4000
3000
2000
1000
0
No
Yes
OnlineBackup
4000
3000
2000
1000
0
No
Yes
DeviceProtection
4000
3000
2000
1000
0
No
Yes
TechSupport
5000
4000
3000
2000
1000
0
No
Yes
StreamingTV
4000
3000
2000
1000
0
No
Yes
StreamingMovies
4000
3000
2000
1000
0
No
Yes
Contract
4000
3000
2000
1000
0
Mn--
Onyr
Twyr
PaperlessBilling
4000
3000
2000
1000
0
No
Yes
PaymentMethod
2000
1500
1000
500
0
Bt()
Cc()
Elcc
Mldc
Churn
No
Yes

Business Perspective: Data Preparation and Modelling

The Data Preparation and Modelling is dealt with in detail in the Technical Perspective part of the report.

Business Perspective: Critical Evaluation

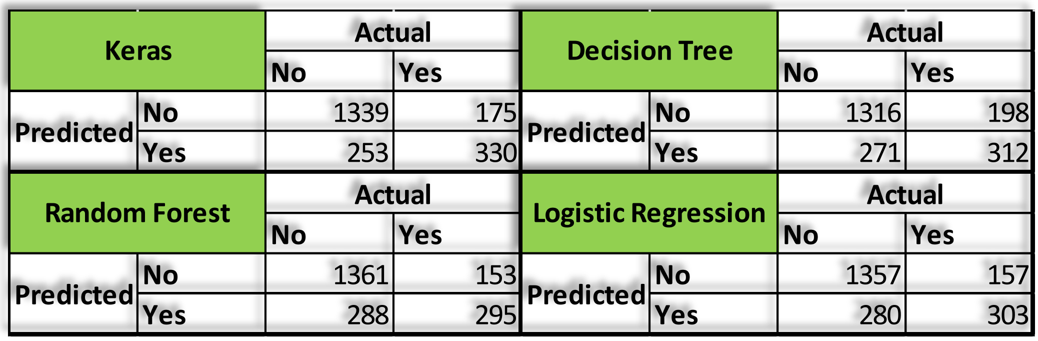
The Evaluation of the model from the Business Perspective is the Cost-Benefit Analysis, which speaks about the Uplift, Additional Profit and the Return on Investment by applying the outcome of this Machine Learning project.

The model finally chosen was the one giving us the best cost benefit across all the four machine learning models employed.

Confusion Matrix

A confusion matrix 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



Keras		Actual		Decision Tree		Actual	
		No	Yes			No	Yes
Predicted	No	1339	175	Predicted	No	1316	198
	Yes	253	330		Yes	271	312
Random Forest		Actual		Logistic Regression		Actual	
		No	Yes			No	Yes
Predicted	No	1361	153	Predicted	No	1357	157
	Yes	288	295		Yes	280	303

Figure.17: *Confusion Matrix for the Four Models*

From the above comparison we can say that the comparison on the basis of Confusion Matrix would be rendered inconclusive, as the figures are almost similar to each other and thus the performance measures are not that far apart (shown in Technical Evaluation). Thus to choose a model in this scenario we will do a cost benefit analysis of the four models and deploy the one giving us the best Return on Investment.

Cost Benefit Analysis

While the probabilities can be estimated from data, the cost and benefits cannot. They generally depend on external information provided via analysis of the consequences of decision in the context of the specific business problem. In our Churn analysis, we try to find out how much is it worth for us to retain a customer. The value depends on the billed amount, the amount spent for enticement, and also the metrics of the model. The value of costs is quite important in this analysis and we will take it as the TotalCharges_Updated which was calculated earlier during the analysis. By marketing surveys we know that typically 10% of it can be considered as the *retainment* cost. Also, the cost to poach a new customer is $750.

By employing our model we are able to change the perspective of a person lying in the TP quadrant who was going to Churn and retain them by providing an enticing scheme like offering them a discount on their bills or some free movie tickets, the value of which will be around 10% of their yearly value as stated before.

The Cost Benefit Analysis of the models represents the overall benefit achieved (RoI) by employing the findings from the project.

$$RoI = (GainfromInvestment/Invesment) * 100$$

The **Gain from Investment** is calculated as the **difference between Revenue from the Retained Subscribers and Lost Subscribers**.

The **Investment** done by taking suggestions from the model is the sum of:

(i) money spent of preventing the customers from churning,

(ii) money spent to entice customers who were loyal in the first place, and

(iii) acquisition money spent to regain the number of customers churned.

$$TotalChargesUpdated = 2283$$

$$AcquisitionCost = 750$$

		Keras	Decision Tree	Random Forest	Logistic Regression
Net Revenue	TP*0.9*TotalCharges	678051.0	630792.9	606136.5	622574.1
Cost of Wrong Enticement	FP*0.1*TotalCharges	39952.5	60956.1	34929.9	35843.1
Lost Revenue	FN*TotalCharges	577599.0	383544.0	657504.0	639240.0
Cost to Replace	FN*AcquisitionCost	189750.0	126000.0	216000.0	210000.0
Uplift		448348.5	443836.8	355206.6	376731.0
P	TP+FN	583.0	475.0	583.0	583.0
NoModel	P*750	437250.0	356250.0	437250.0	437250.0
Retainment Investment	TP*0.1*TotalCharges	75339.0	70088.1	67348.5	69174.9
Total Investment		305041.5	257044.2	318278.4	315018.0
Model Gained Rev		488301.0	504792.9	390136.5	412574.1
%ROI		**160.08%**	**196.38%**	**123.58%**	**131.97%**

Table.4: *Cost Benefit Analysis of the Four Models*

From the above table we can see that out of the Four Models the Decision Tree model gives us the best ROI, almost close to 200% (i.e. double the investment).

This however, can be considered due to the fact that it was run on only a select few features which are known to have a high effect on Churn, its results although are very good they may not be such that can be implemented across the entire customer base.

Neural Network also gives us a great RoI and can be considered as an applicable scenario, but it must be kept in mind that it runs as a black box and so any domain expertise or assumption cannot be made to reflect in its results.

The Random Forest which gave us slightly better performance results than the Decision Tree can now be seen to be the worst amongst the best, giving us around 23% in enhanced revenue.

The Logistic Regression is also one that is run on all variables and the results (be it statistically significant or not) are giving us approximately a 33% enhanced revenue when compared to the investment. Its results are quite important as they assisted in modelling the other techniques and helped in enhancing the results from the Decision Tree to an extent that it now gives returns almost twice the money invested (working on a select type of customers though).

Deployment: Conclusion & Recommendations

The Machine Learning models from this project are analysed on the basis of their technical specifications and also by the business benefit they provide. Although they are quite similar in various performance measures the slightly better accuracy and strength to predict by a model (ANN) when multiplied with the amount of bill leads to a significant increase in the RoI%. The Decision Tree can be seen as the best model in terms of RoI and can be used as the model to target marketing and strategy building on the customer base it relates to so that value can be extracted from it.

It is suggested that the management decide as according to the short term and long term plans of the company to focus on the results of these models then try and use directed approaches to segments. All in all, the project can be considered a success as each employed technique gives a significant benefit and helps in enhancing revenue for the company.

References

Amaresan, S. (2018). What Is Customer Churn? [Definition]. [online] Blog.hubspot.com. Available at: https://blog.hubspot.com/service/what-is-customer-churn [Accessed 6 Apr. 2019].

Chen, P. and Hitt, L. (2002). Measuring Switching Costs and the Determinants of Customer Retention in Internet-Enabled Businesses: A Study of the Online Brokerage Industry. Information Systems Research, [online] 13(3), pp.255-274. Available at: https://pubsonline.informs.org/doi/abs/10.1287/isre.13.3.255.78 [Accessed 17 Apr. 2019].

Ryman-Tubb, N. F. (2019), Pre-processing functions in R Code. The Surrey Business School, University of Surrey, MSc Machine Learning & Visualisation Module.

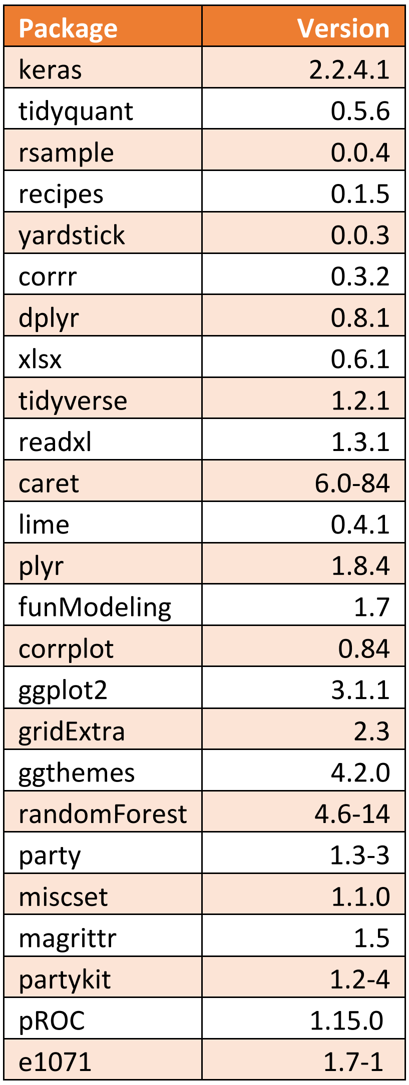
Appendices

Appendix A: Data Dictionary and Summary

Fields	Comments	type	Number of unique values
customerID	customerID	factor	7001
gender	gender (female, male)	factor	2
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)	integer	2
Partner	Whether the customer has a partner or not (Yes, No)	factor	2
Dependents	Dependents (Whether the customer has dependents or not (Yes, No))	factor	2
tenure	Number of months the customer has stayed with the company	integer	73
PhoneService	Whether the customer has a phone service or not (Yes, No)	factor	2
MultipleLines	Whether the customer has multiple lines r not (Yes, No, No phone service	factor	3
InternetService	Customers internet service provider (DSL, Fiber optic, No)	factor	3
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)	factor	3
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)	factor	3
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)	factor	3
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)	factor	3
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)	factor	3
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)	factor	3
Contract	The contract term of the customer (Month-to-month, One year, Two year)	factor	3
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No))	factor	2
PaymentMethod	The customers payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)))	factor	4
MonthlyCharges	The amount charged to the customer monthly - numeric)	numeric	1582
TotalCharges	The total amount charged to the customer - numeric)	numeric	6490
Churn	Whether the customer churned or not (Yes or No))	factor	2

Appendix B: Package Details

Back to Executive Summary



Package	Version
keras	2.2.4.1
tidyquant	0.5.6
rsample	0.0.4
recipes	0.1.5
yardstick	0.0.3
corrr	0.3.2
dplyr	0.8.1
xlsx	0.6.1
tidyverse	1.2.1
readxl	1.3.1
caret	6.0-84
lime	0.4.1
plyr	1.8.4
funModeling	1.7
corrplot	0.84
ggplot2	3.1.1
gridExtra	2.3
ggthemes	4.2.0
randomForest	4.6-14
party	1.3-3
miscset	1.1.0
magrittr	1.5
partykit	1.2-4
pROC	1.15.0
e1071	1.7-1

Appendix C: Calculated statistics for the churn dataset

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.

variable	level	num	Churn [%]	Standard Deviatin	Num No Churn	Num Churn
gender	Female	3460	27%	0,44	2526	934
gender	Male	3530	26%	0,44	2608	922
SeniorCitizen	No	5858	24%	0,42	4473	1385
SeniorCitizen	Yes	1132	42%	0,49	661	471
Partner	No	3618	33%	0,47	2422	1196
Partner	Yes	3372	20%	0,40	2712	660
Dependents	No	4905	31%	0,46	3373	1532
Dependents	Yes	2085	16%	0,36	1761	324
PhoneService	No	673	25%	0,43	505	168
PhoneService	Yes	6317	27%	0,44	4629	1688
MultipleLines	No	4038	25%	0,43	3024	1014
MultipleLines	Yes	2952	29%	0,45	2110	842
InternetService	DSL	2404	19%	0,39	1947	457
InternetService	Fiber optic	3078	42%	0,49	1792	1286
InternetService	No	1508	7%	0,26	1395	113
OnlineSecurity	No	4985	31%	0,46	3422	1563
OnlineSecurity	Yes	2005	15%	0,35	1712	293
OnlineBackup	No	4580	29%	0,46	3239	1341
OnlineBackup	Yes	2410	21%	0,41	1895	515
DeviceProtection	No	4587	29%	0,45	3270	1317
DeviceProtection	Yes	2403	22%	0,42	1864	539
TechSupport	No	4963	31%	0,46	3413	1550
TechSupport	Yes	2027	15%	0,36	1721	306
StreamingTV	No	4303	24%	0,43	3253	1050
StreamingTV	Yes	2687	30%	0,46	1881	806
StreamingMovies	No	4278	24%	0,43	3231	1047
StreamingMovies	Yes	2712	30%	0,46	1903	809
Contract	Month-to-month	3852	43%	0,49	2207	1645
Contract	One year	1459	11%	0,32	1295	164
Contract	Two year	1679	3%	0,17	1632	47
PaperlessBilling	No	2844	16%	0,37	2379	465
PaperlessBilling	Yes	4146	34%	0,47	2755	1391
PaymentMethod	Bank transfer (automatic)	1532	17%	0,37	1278	254
PaymentMethod	Credit card (automatic)	1514	15%	0,36	1283	231
PaymentMethod	Electronic check	2351	45%	0,50	1287	1064
PaymentMethod	Mailed check	1593	19%	0,39	1286	307

Appendix D: Correlation matrix

This appendix shows the correlation matrix for the categorical fields: