**FACULTY OF ARTS AND** 

**SOCIAL SCIENCES**

**COURSEWORK COVERSHEET**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coursework Details** | | | |
| **Module Name and Code**  **(please check front of module handbook)** | Machine Learning and Visualisations | | |
| **Coursework Title** |  | | |
| **Date of Submission** | 20.06.2019 | **Word Count** | 3434 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Students Details** | | | |
| **Students URN (7 digit number on Uni card)** | 6345112 | **Student Names** | Vartan Zahorodnykov |
| **Programme** | Business Analytics MSc - 2018/9 | | |

|  |
| --- |
| **Students Declaration**  *To be agreed by all students* |
| Please refer to the **University of Surrey Regulations for the Conduct of Examinations and Other Forms of Assessments** and your departmental **Student Programme Handbook** for more information on Academic Misconduct and Plagiarism.  ***Declaration:***  *I confirm that the submitted work is my own work and that I have clearly identified and fully acknowledged all material that is entitled to be attributed to others (whether published or unpublished) using a referencing system. I agree that the University may submit my work to means of checking this, such as the plagiarism detection service Turnitin® UK. I confirm that I understand that assessed work that has been shown to have been plagiarised will be penalised.*  By completing and submitting this form, we confirm that:   * We have read and fully understand the University’s Regulations and guidance on Academic Misconduct and Plagiarism * This submission is our own work * All quotes and sources have been fully and properly attributed and referenced * This work has not been previously submitted, in full or in part, for the purpose of assessment at this or any other institution * No effort has been made to subvert plagiarism detection processes of the University * This submission may be transferred to and stored in the Turnitin Plagiarism Detection database for the purpose of plagiarism detection now and in the future * We understand that all required work must be received within the published deadline * We understand that work received after the published deadline will be penalised in line with University Regulations * We understand that any request for mitigating circumstances must be made formally, using the appropriate form and including evidence; the application and associated evidence must be received by the stipulated date |

*BY SUBMITTING THIS FORM, YOU AGREE TO THE STUDENT DECLARATION IN FULL*

**Machine Learning and Visualisations**

**2018-2019**

**Special Thanks** to

Prof Nick Ryman-Tubb

**Content:**

[**1.0 Executive Summary**](#Executive_Summary_2)…………………………………………………………………….

[**2.0 Introduction**](#Introduction_2)…………………………………………………………………………….

[**3.0 Business Understanding**](#Business_Understanding_2)……………….……………………………………………...

[**4.0 Data Understanding**](#Data_Understanding_2)…….……………………………………………………………...

[**5.0 Data Preprocessing**](#Data_Preprocessing_2)…………………….……………………………………………….

5.1 *Dimensionality Reduction*……………………………………………………………………......

*5.2 Feature* *Construction*……………………………………………………………………………

*5.3 Correlation Analysis and Feature Constructio*n….……………………………………………..

[**6.0 Data Modelling and Technical Evaluation**](#Data_Modelling_2)………………………………………… *6.1 Data Randomisation and subset split*………………..………………………………………………

*5.2 LDAmodel*……………………………………………………………………………………

*6.3 Decision Tree*…………………………………………………………………………............

[**7.0 Modelling Business Evaluation**](#Modelling_Business_Evaluation)…………………........................................................

*7.1 Modelling Business Application*………………………………………….................................

*7.2 “What-if” and Sensatitivity Business Evaluation*……………………………………………….

[**8.0 Recommendations**](#Recommendations_2)…………………………………………………………………......

[**9.0 References**](#References_2)……………………………………………………………………….............

[**10.0 Appendix**](#Appendix_2)…………………………………………………………………………..........

**1.0 Executive Summary**

This report will present an analysis on the Telco Dataset csv file, which contains various socio-demographic, product choice behaviour, purchasing, and payment information supplied from the internally based company database. The reports aims to give a clear and concise overview of the typical profile of individual Telco subscriber, determine which type of subscriber is likely to churn and the potential chances of retention. Financial incentives, overall unlift value, investments required and ROI of retained potential Churn customers will be considered as well. Several machine learning techniques will be deployed in order to predict accurately the churn behaviour within the dataset and provide comprehensive recommendations to various stakeholders at different levels within the organisation.

**2.0 Introduction**

The reports is set to follow the conventional CRISP-DM (Cross Industry Standard Process Database) structure. The reader will be introduced to the business problem and understanding of the problem and then shown the structure of the dataset and dataset visualisation insights. Following that the reader will be able to follow [Data Pre-processing](#Data_Preprocessing_2), Modelling and Technical evaluation, which forms the core technical part of the report. In addition, the reader will be presented with the Modelling and Business evaluation section, which intends to give practical recommendations and forms the core business part of this report.

To see more information the user can make use of [Appendix](#Appendix_2), which contains other useful information, relevant visualisations, tables etc.

**3.0 Business understanding**

Contemporary telecommunications industry has matured and saturated. As a result, it becomes really hard for companies to retain the current client base, existing market share and remain competitive. Acquiring new customer becomes a real challenge since the cost of acquiring a customer can often outweigh by subscribers’ continuous retention policies. It is more feasible therefore for the telecommunications business to focus on customer retention and loyalty built through customer satisfaction and incentives.

Presented customer Telco dataset presents 21 fields about 7001 different customers. The data understanding, modelling and deployment sections of this report will ultimately aim to describe the core context and insights behind Telco dataset. This is done to correctly predict a number of subscribers who actually churn (True Positive number) and not remain loyal (True Negative number). Wrong predictions in terms of the number of subscribers actually stay loyal despite having been classified as churn (False Positive number) as well as customers who do churn despite having been classified as loyal (False Negative number) will be presented too. TP, TN, FP, FN number will then be used to be able to successfully calculate various business matrics such as:

1-net revenue from subscribers the business can gain by spending 10% on retention

2-wrongly enticed customers

3-lost revenue based on missed subscribers who could have been enticed successfully

4-cost of acquiring a new customer to retain the same total customer base

5- subtotals of total proportional metrics are calculated

6-total cost for business if no incentive model is applied

7-total opportunity benefit if the inceptive model is applied as opposed to having no model at all (step 5-step 6)

8-gain on Investment also known as ROI.

The results from the various business metrics are then used to better understand the current business problem, technical part feasibility of this report as well directly applied to [Recommendations](#Recommendations_2) section.

**4.0 Data Understanding**

The Telco database is delimited and presented in the txt format. It contains 7001 rows (observations) within 21 fields. 5 of them describe socio-demographic variables (attribute number 2-6), 9 describe other product or service use (7-15), 2 contract duration and mail information, (16-17), 3-information charges (18-20). The 21st first is determined as a key target field (i.e. Churn) and will be used an output field throughout this report.

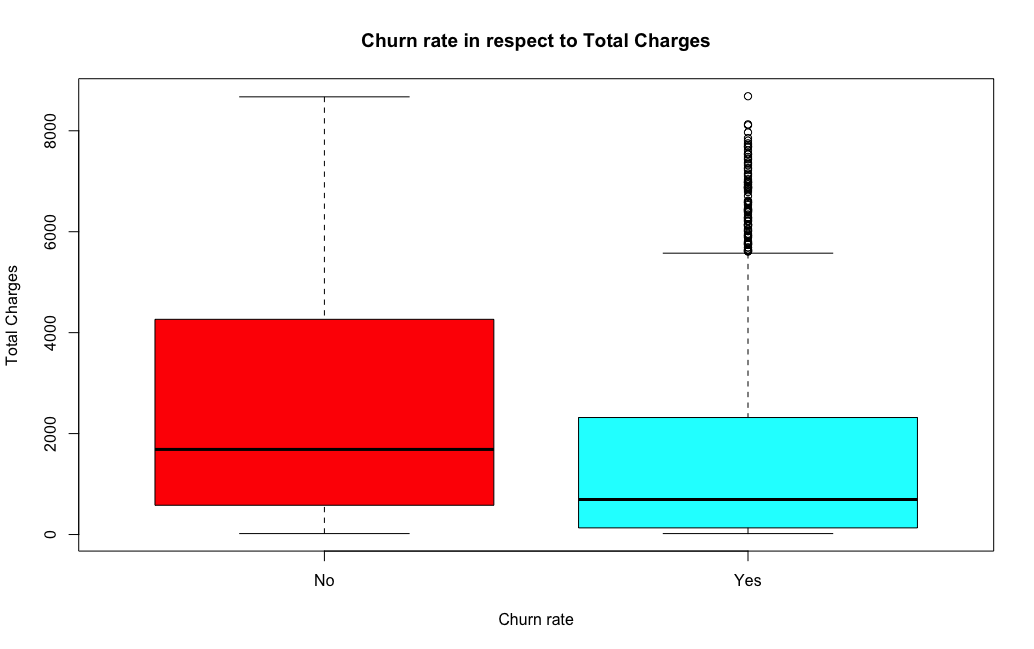
Upon examining type of attributes it is discovered that all attribute belong to the symbolic type with the exception of Tenure, Monthly and Total charges (which are numeric). According to the project domain knowledge tenure is measured in months, and charges in USD, which inevitably led R to further classify these three attributes as ordinal. A relation between these three ordinal attributes and Churn rate has been visualised through the use of barplot R code, which describes a median value of the given attribute both within Churn and Non-churn rate category.

The following patterns and relationships were discovered:

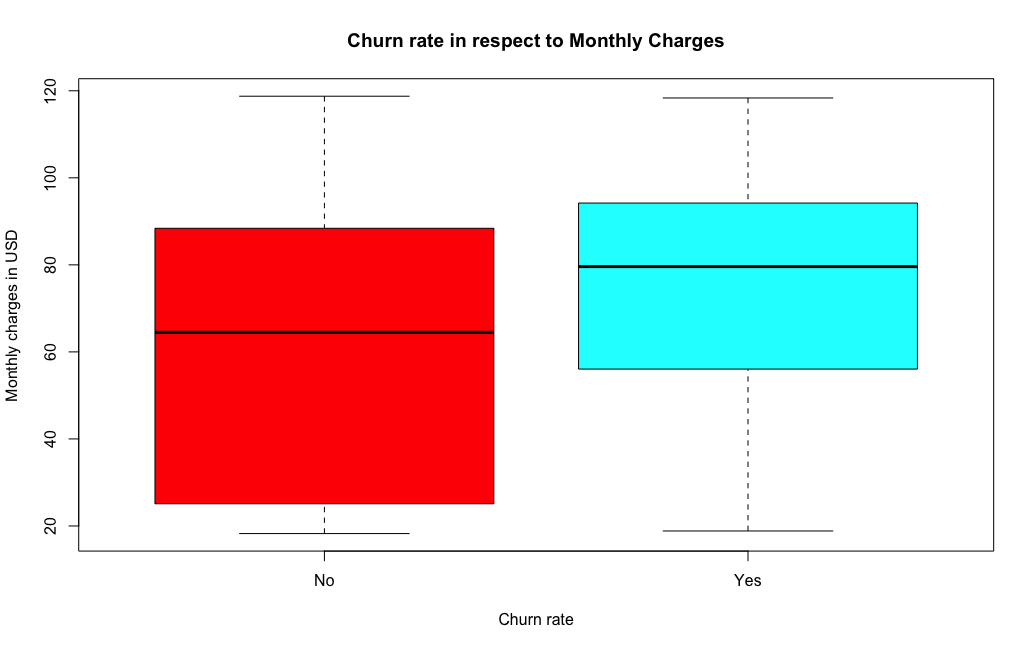
1) Customers usually leave the service after 10 months of service use



2) Customers who churn pay quite little (overall median of around 900 USD)



3) Customers who churn have high Monthly charges (median of around 80 USD)

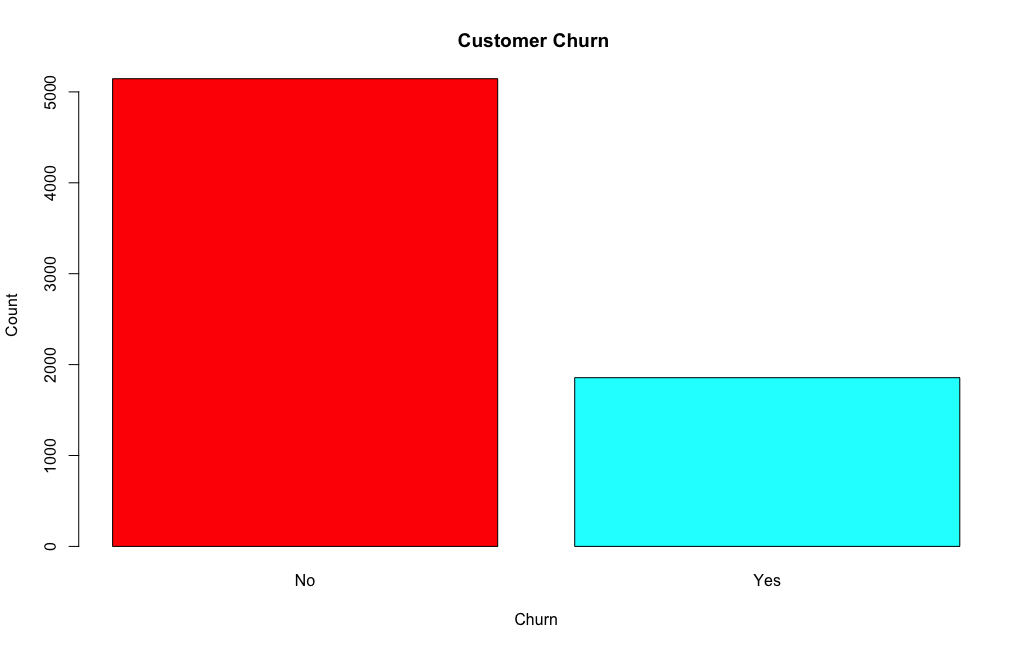


Although the first field of the dataset (Customer ID) contains unique code ID description of each customer the data itself presents no insight to the reader and analysis and hence being removed (as per first steps of [4.0 Data pre-processing](#Data_Preprocessing_2) section).

To understand each of the remaining 17 attributes within the dataset and their relation to output field Churn each field has been visualised through the bar plot diagram (see [Appendix-Barplot Graphs](#Barplot_Graphs_2) section). Some diagrams are also included in this section to stress the significant relationship between visualised attribute and Churn rate.

The following patterns and relationships were discovered:

1) The dataset is relatively unbalanced with more than 5140 customers being a non-churn customer against 1860 Churn customers

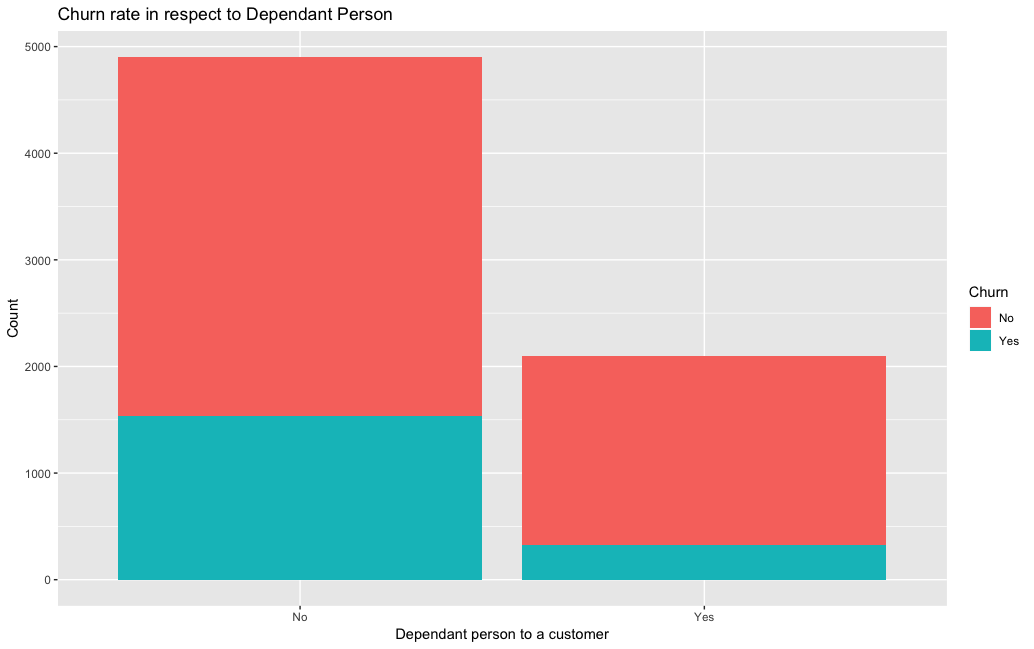


2) Both males and females have the same likelihood to churn

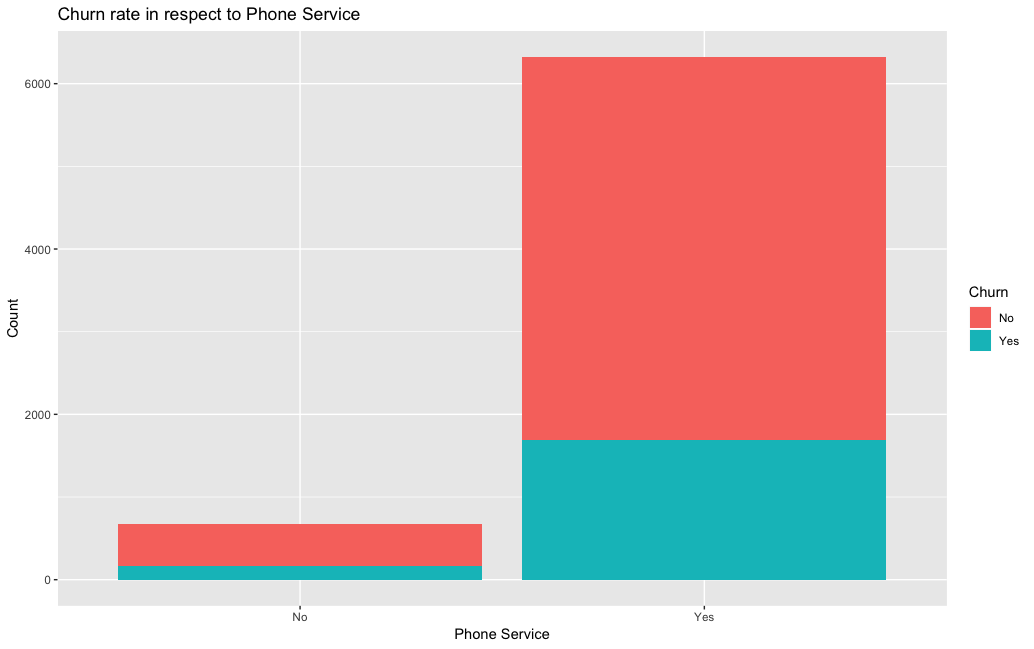
3) Greater number of customers who are not Senior Citizens are likely to churn (around 200)

4) Customers involved in the relationship are less likely to churn (around 500 only)

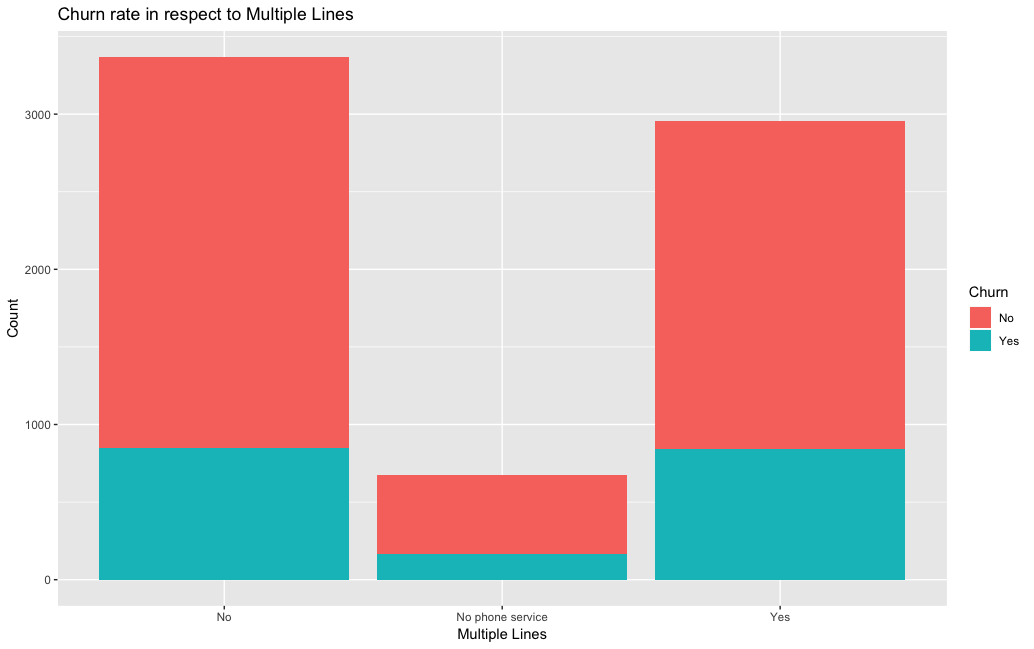
5) Customers with Dependent are less likely to churn (around 150 only)

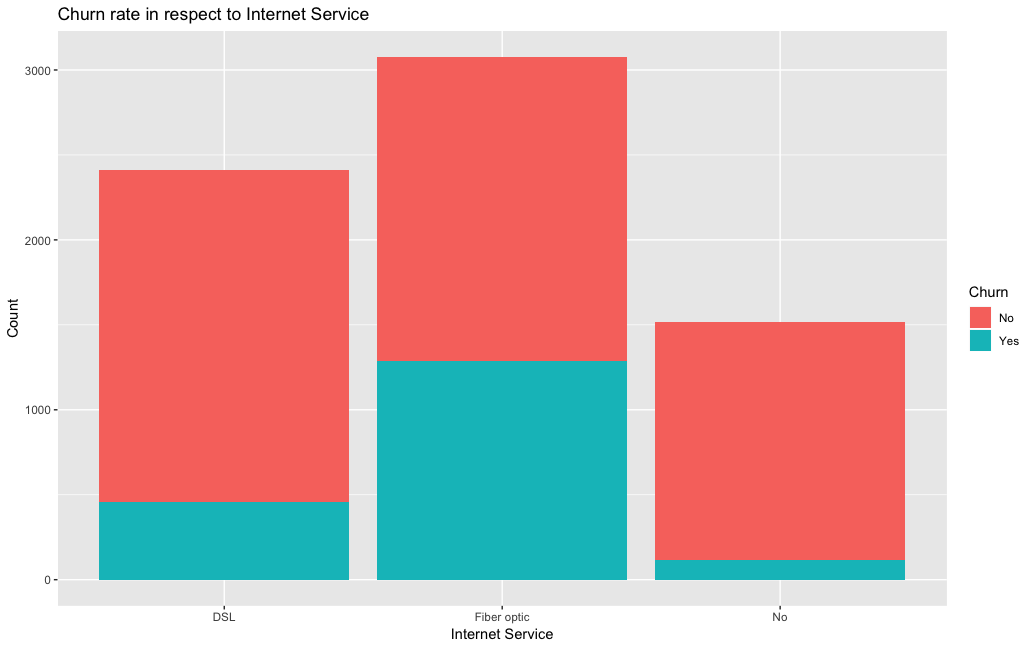


6) As 90% of customers have phone service Churn rate is much higher in this category (around 1800)

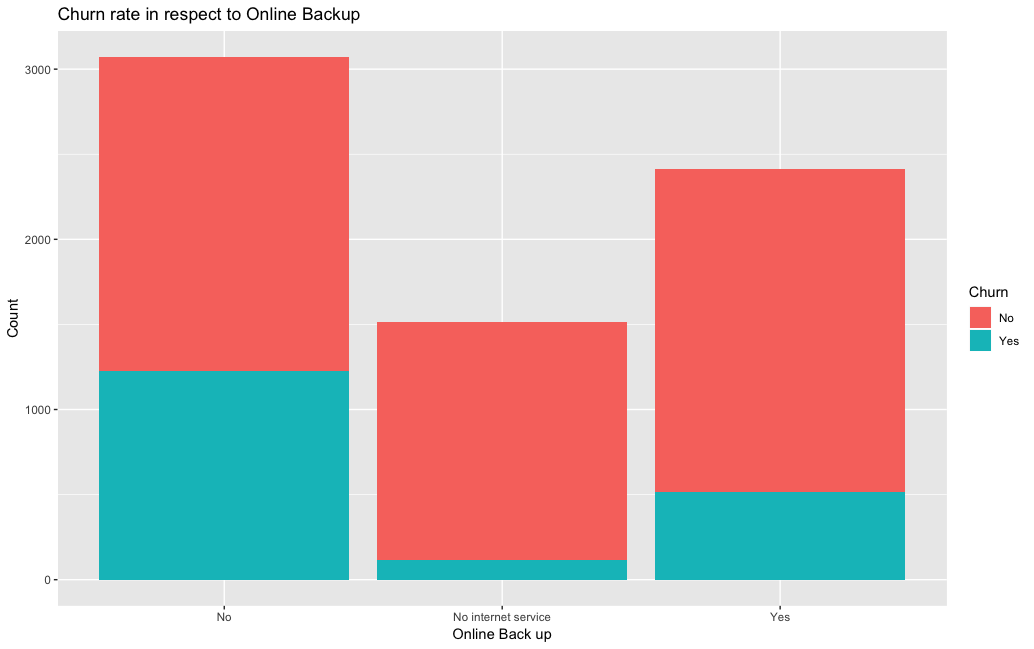


7) Churn rate is relatively the same regardless of the Number of Lines customers have (around 900)

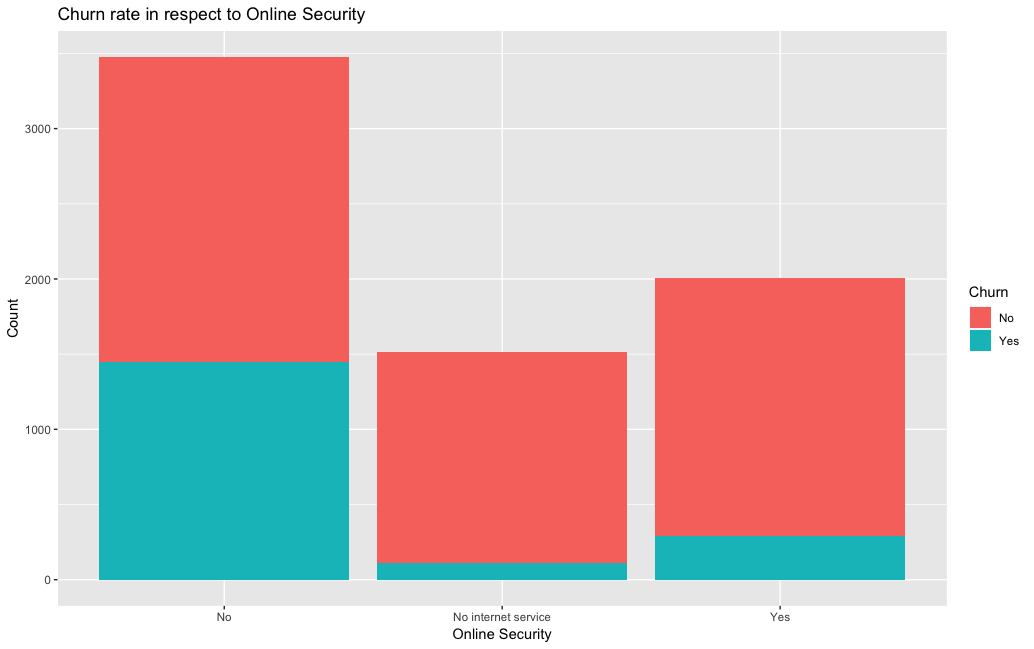


8) Fibre Optic Customers are more likely to churn (around 1300) 

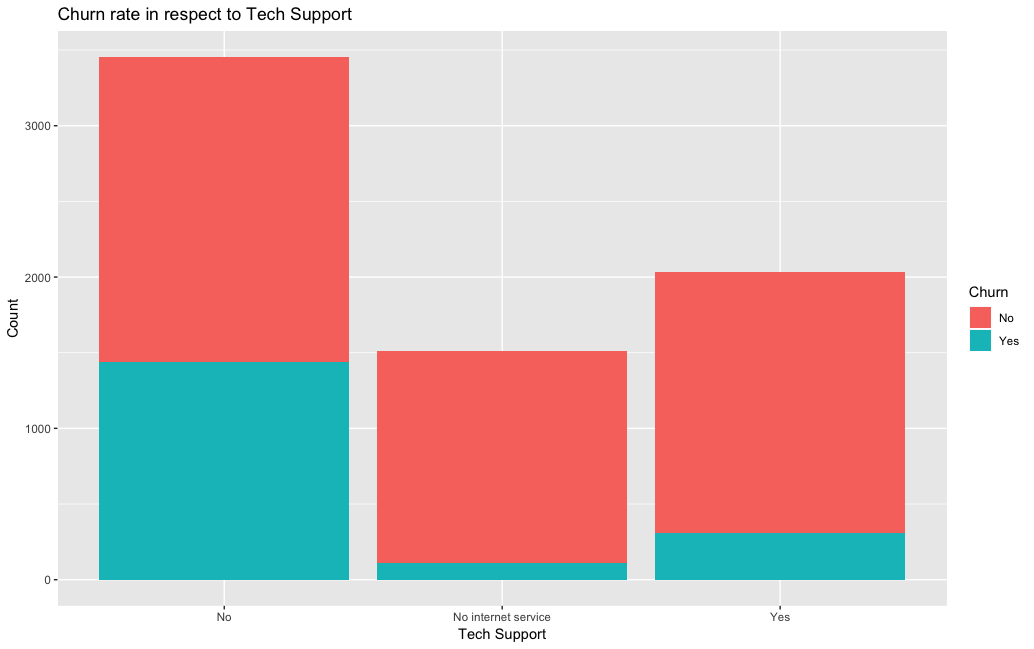
9) Customers without online backup are more likely to churn (around 1100)



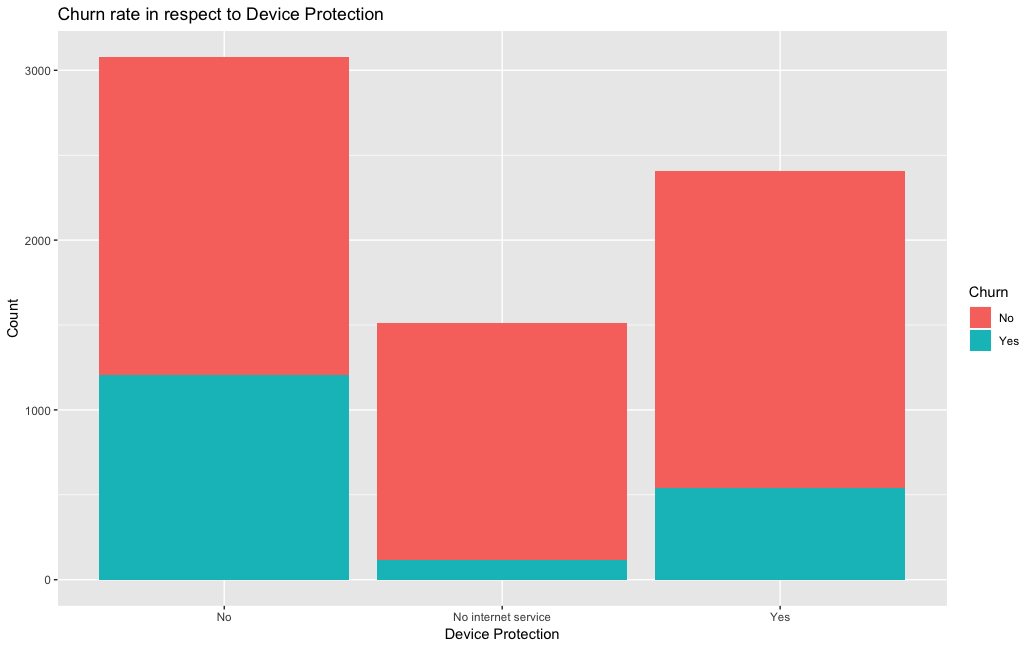
10) Customers without security are more likely to churn (around 1500)



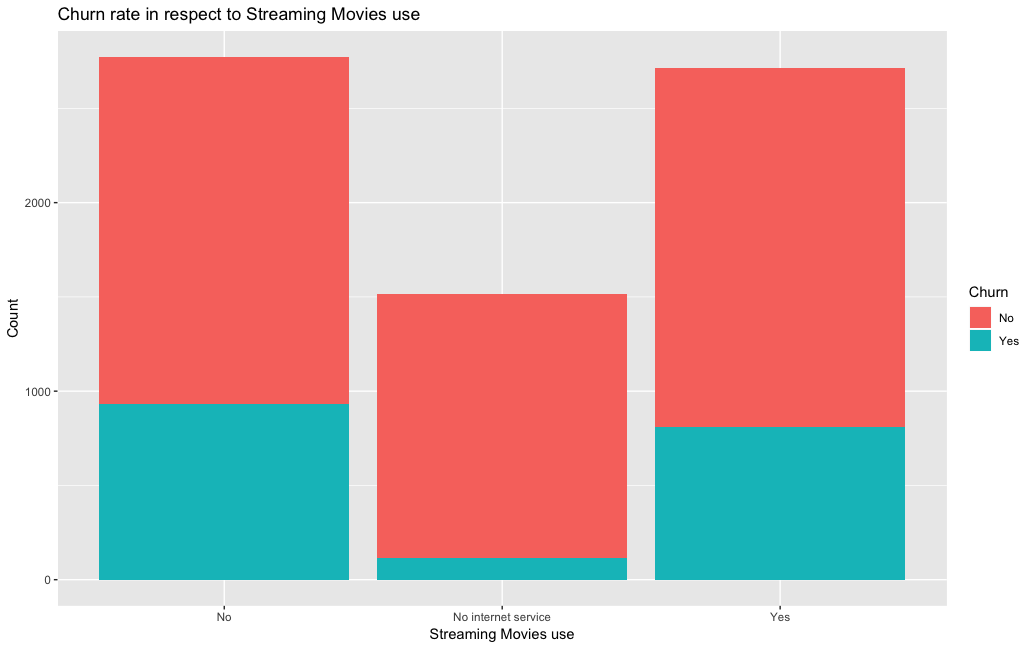
11) Customers without tech support are more likely to churn (around 1500)



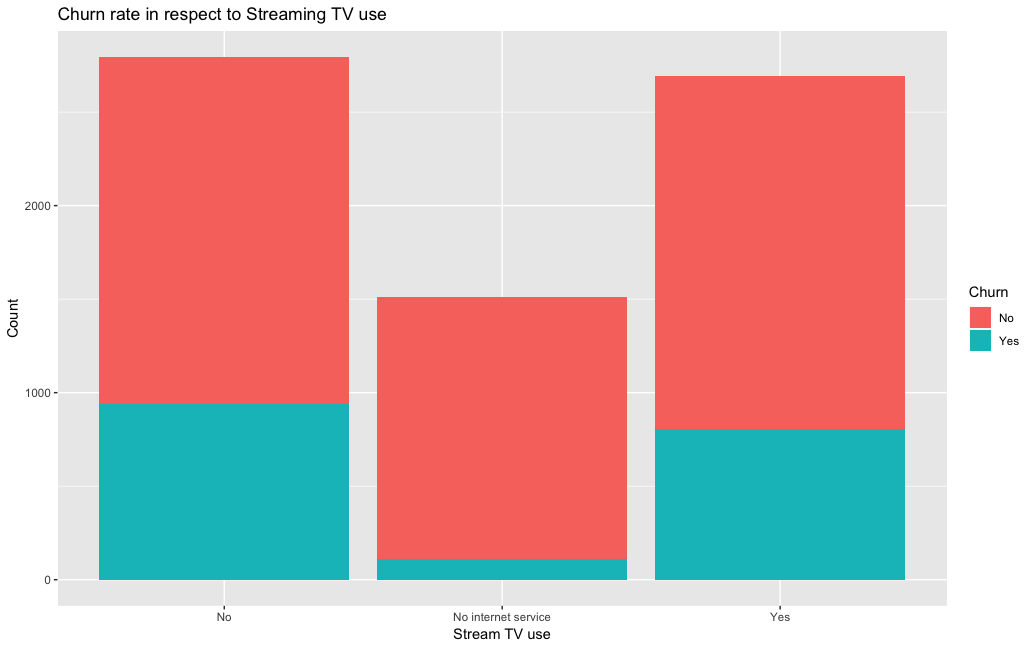
12) Customers without the device protection are more likely to churn (around 1100)



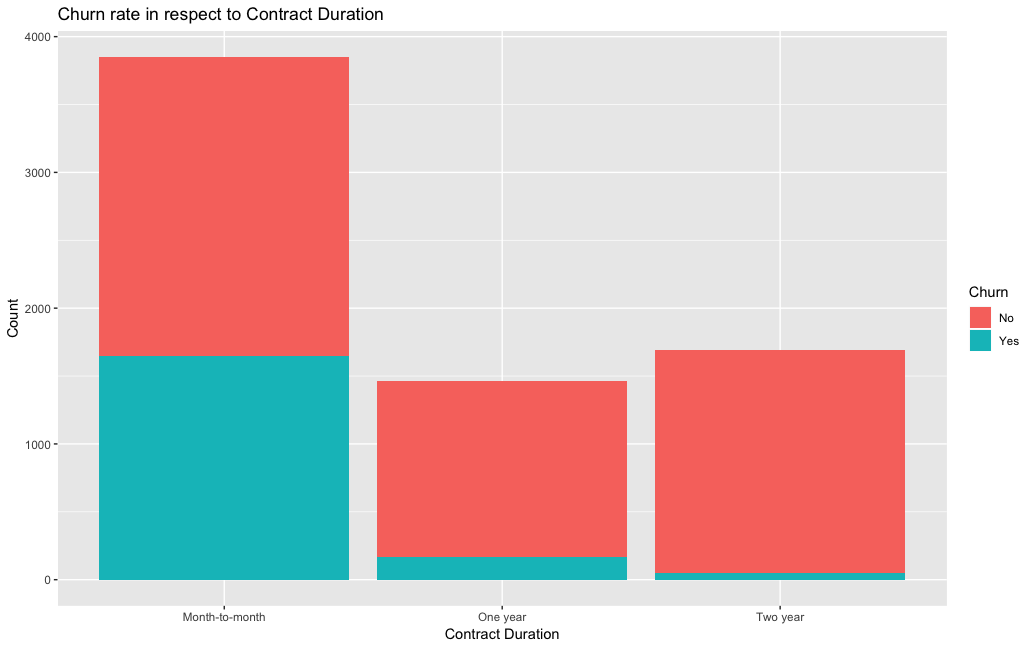
13) Customers who don't stream movies are more likely to churn (almost 1000)



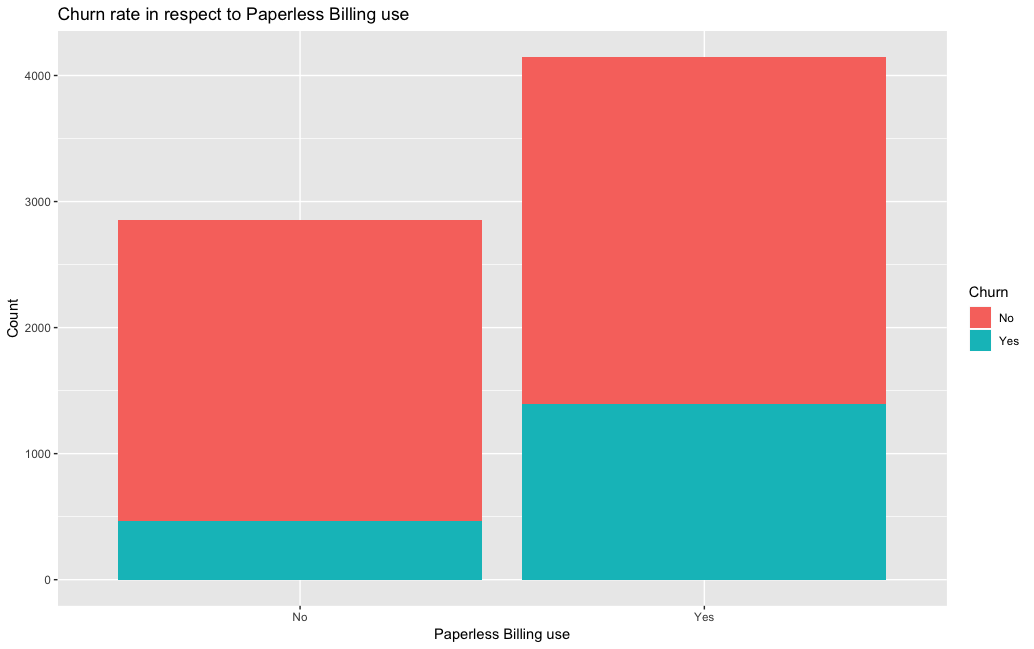
14) Customers who don’t stream TV are as likely to churn as customers who do (almost 1000 vs 900)

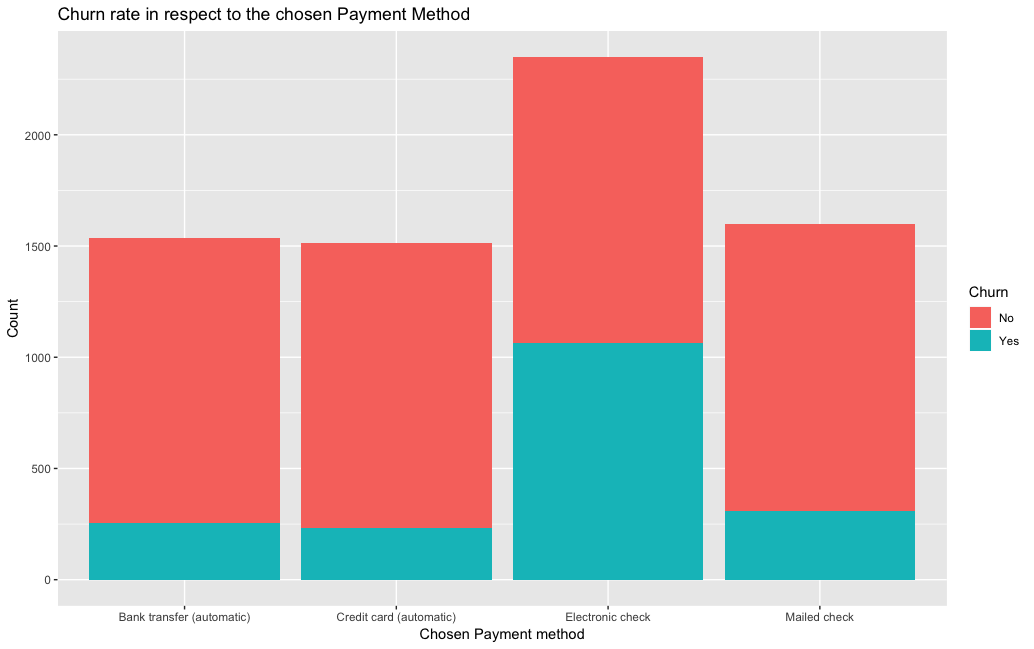


15) Customers who have short monthly contracts are much more likely to churn (around 1700)



16) Customers who use paperless billing use are much more likely to churn (around 1400)



17) Customers who use electronic check as a Payment method are much more likely to churn (around 1050). \

To conclude customers who are connected to the internet, use electronic checks without certain type of services (i.e. OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV and StreamingMovies) showed a visible greater tendency to churn.

**5.0 Data Preprocessing**

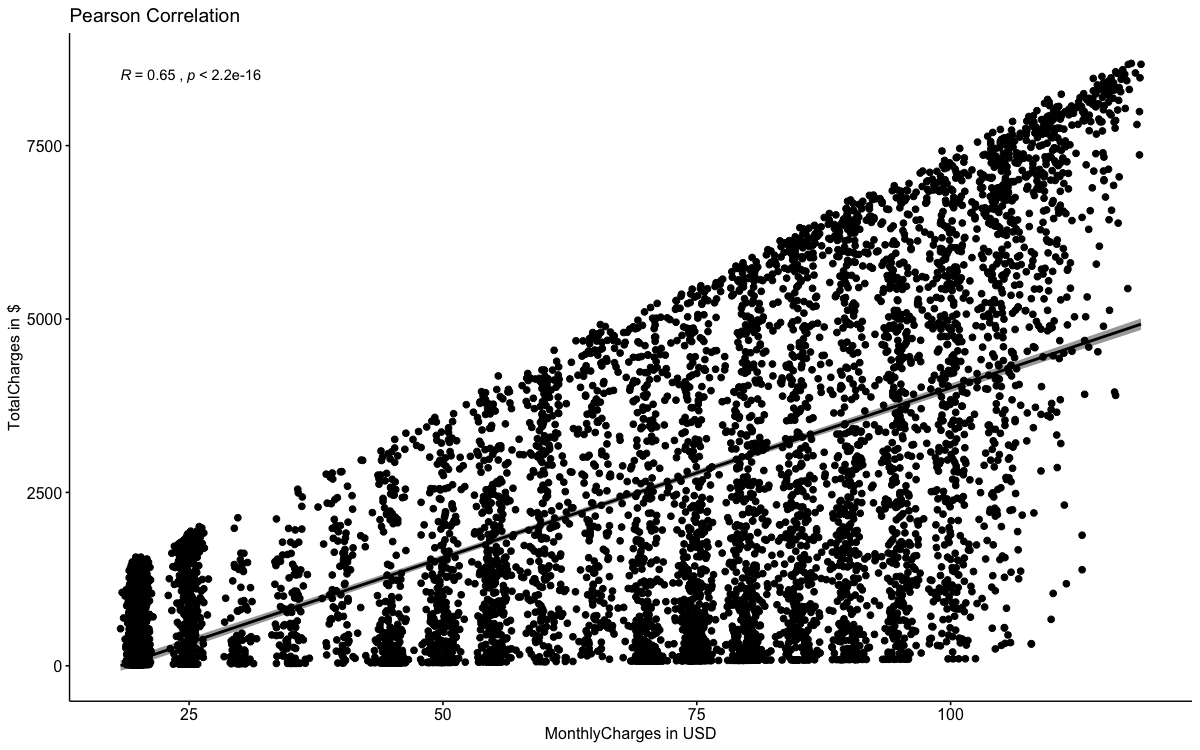
As the training dataset contains 21 different attributes with some similar descriptions it is necessary to undertake data cleaning, feature selection and construction steps. This defines the most and least important attributes and prepares data for further modelling and analysis.

*5.1 Dimensionality reduction*

As mentioned [3.0 Data Understanding](#Data_Understanding_2) section attribute 1 (CustomerID) is removed as redundant.

According to the domain knowledge it has been assumed that Total Charges could be calculated as Monthly Charges\*Tenure. To verify the dependency between the Total and Monthly Charges a Pearson Correlation ggscatterplot is applied.

*Total Charges and Monthly Charges Pearson Correlation plot*



As the graph above confirms high correlation between the two Total charges are being removed to achieve further dataset dimensionality reduction.

*5.2 Feature Construction*

The Telco1 dataset is then normalised to categorical values of 0 and 1 (0 as no and 1 as yes). This creates 20 new variables, which describe various subcategories within a few attributes (e.g. Internet service attribute contains DSL, Fibre Opec and no Internet service-all separated into 3 new attributes with 0 to 1 values now).

Monthly Charges and Total charges are scaled in separate line of code and then combined to the Telco\_All\_Scaled dataset. Due to the nature of ordinal type of these attributes they are z-scaled first and then scaled to 0 and 1 as a continuous (ordinal values). For the simplicity the values of these values are subsequently rounded to one decimal place.

*5.3 Correlation analysis and Feature Selection*

As a final step of data pre-processing an attribute correlation analysis is applied to determine which field correlate to each other. Highly correlated fields beyond the chosen level of 0.6 are then removed as redundant (i.e. 24 fields-see Appendix-[Field Correlation plot](#Field_Correlation_Plot_2)).

**6.0 Data Modelling and Technical Evaluation**

As opposed to unsupervised machine learning approaches where no label and/or predictor attribute is being given supervised learning methods determine how exactly the remaining labelled attribute interact with the output field Churn. Specifically, Linear Discriminant Analysis and Decision Tree approaches are applied in this part of the report. A general overview of these models, their strengths, weaknesses are provided. To evaluate the performance of these models confusion matrix calculations and ROC chart are also applied.

*6.1 Data Randomisation and subset split*

Prior to the initialising the modelling process Telco\_Uncorr dataset records are being randomised to ensure the unbiasness and consistency of results. Data is then split into training and test subsets in a ratio of 70:30 to enable to modelling evaluation and classifier application.

*6.2 LDA model*

Linear Discriminant Analysis algorithm consists of statistical properties of the data, calculated for each class (Brownlee 2019). For a single input variable this is the mean and the variance of the variable for each class. For multiple variables, these are the same properties calculated over the multivariate Gaussian, namely the means and the covariance matrix (Brownlee, 2019).

Gaussian parameters are then plugged into the LDA equation to make predictions

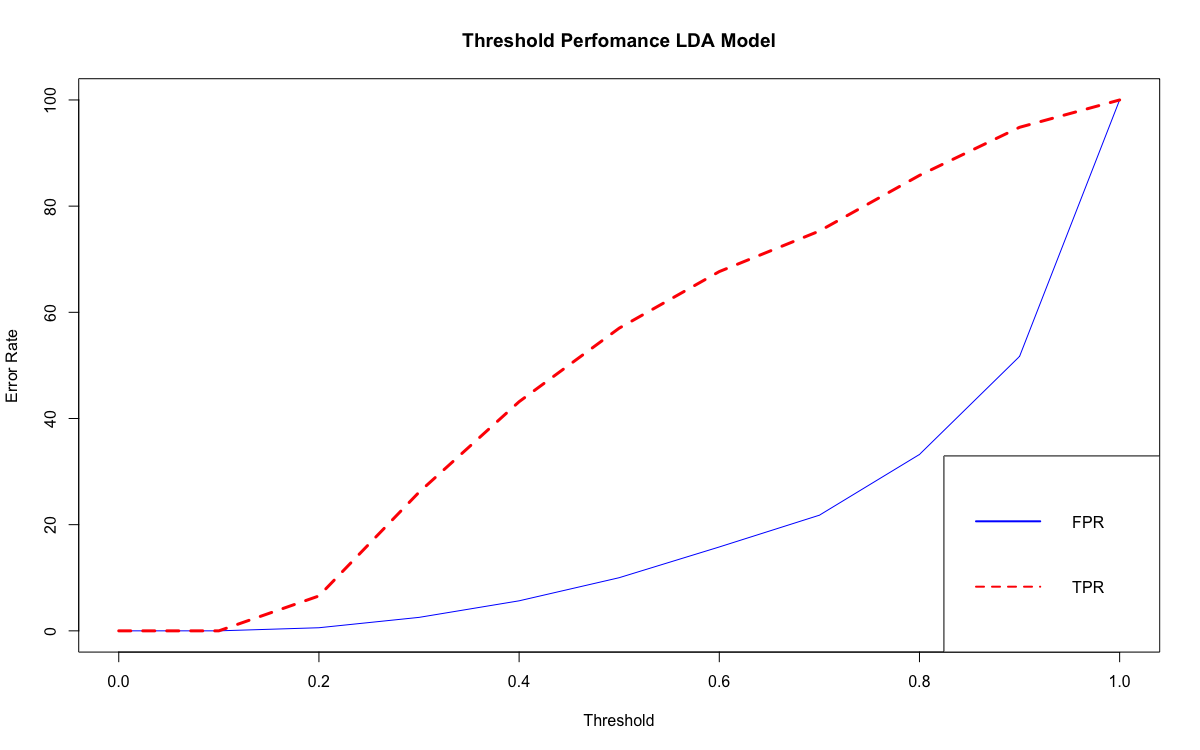
These statistical properties are estimated from your data and plug into the LDA equation to make predictions. The predictions are made by estimating the probability that a new set of input variables belong to each class. The class that gets the highest probability is the output class and a prediction is made.

LDA tries its hard to detect if the within-class covariance matrix is singular and is great to determine a distinct margin among the attributes of highly dimensional data. Highly dimensional data is likely to have a quite big significant overlap between the attributes, which can have a mix impact on the Output field. This means that the fields that have a district predictive impact on the Output field could be overlapped by other less predictive ‘clutter’ fields in terms of impact. LDA maximises the distance between various attributes effectively minimising the overall area between them (Nistrup, 2019).

On the other hand LDA cannot be applied to the data where attribute possess more discriminatory information, determined by variances rather than mean values or where attribute distributions are not based on either mean values/variances (Research.cs.tamu.edu, 2019).

The graph below describes LDA model TPR and FPR at a different level of thresholds and an error rate with a widening gap between FPR and TPR at the 0.7 level of threshold. This suggests that the level of around 0.7 presents an error compromise on various levels for FPR and TPR.

*LDA model evaluation, Threshold and error rate*



To evaluate the all the models in this report Confusion matrics approach is being used:

1) Accuracy of LDAmodel with the threshold of 0.7 has been calculated as (TP+TN)/( TP+FP) +(FN+TN)

2) Precision Good as TP/(TP+FP)

3) Precision bad as FP/(TP+FP)

4) FPR as FP/N where N (all negative values) is calculated as TN+FN

5) TPR as TP/P where P (all positive values) is calculated as TP+FP

*LDAmodel evaluation with threshold of 0.7*

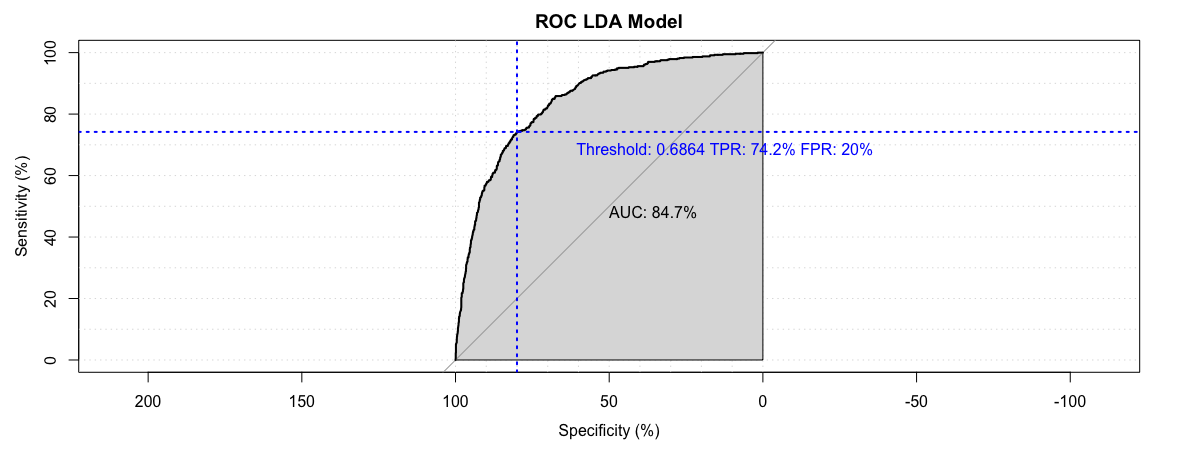
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TP** | **424** | **FN** | **TN** | **FP** | **accuracy** | **pgood** | **pbad** | **FPR** | **TPR** |
|
| **FN** | 139 | 139 | 1203 | 335 | 77.43 | 55.86 | 89.64 | 21.8 | 75.3 |
|

The ROC curve (Receiver Optimising Characteristic) is also produced to describe a ratio between True positive rate (TPR, sensitivity) and True Negative Rate (TNR, specificity) determined by the various threshold levels (cut-off values). Any ROC curve is meant to have a optimal level of threshold where maximum TPR and minimum TNR are achieved.

However, the optimum level may not necessarily mean the best point for the given situation. The highest TPR is likely to lead to the higher level of FPR. Similarly, the lowest FPR levels are likely to lead to lower TPR.

In addition, the optimum level of threshold (0.68) is then being applied in the original LDA model function, which produces slightly different confusion table.

*ROC LDAmodel evaluation*



*ROC model evaluation with optimum threshold of 0.68 applied*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FN** | **TN** | **FP** | **accuracy** | **pgood** | **pbad** | **FPR** | **TPR** |
|
| 1228 | 146 | 420 | 307 | 78.43 | 80 | 74.20 | 42.22 | 89.379025 |

*6.3 Decision Tree model*

To further validate results produced from the LDAmodel as primary regression and classification technique another supervised machine learning algorithm is applied. Whilst using similar to real life tree-like branches for decision classification standard decision tree involves deciding which features to choose by and, upon what measures it should split. Moreover it has a recursive algorithm nature, which means that upon finding interesting relationships between attributes the main tree can be subdivided further into the deeper decision trees should the user wish to do so (Gupta, 2017).

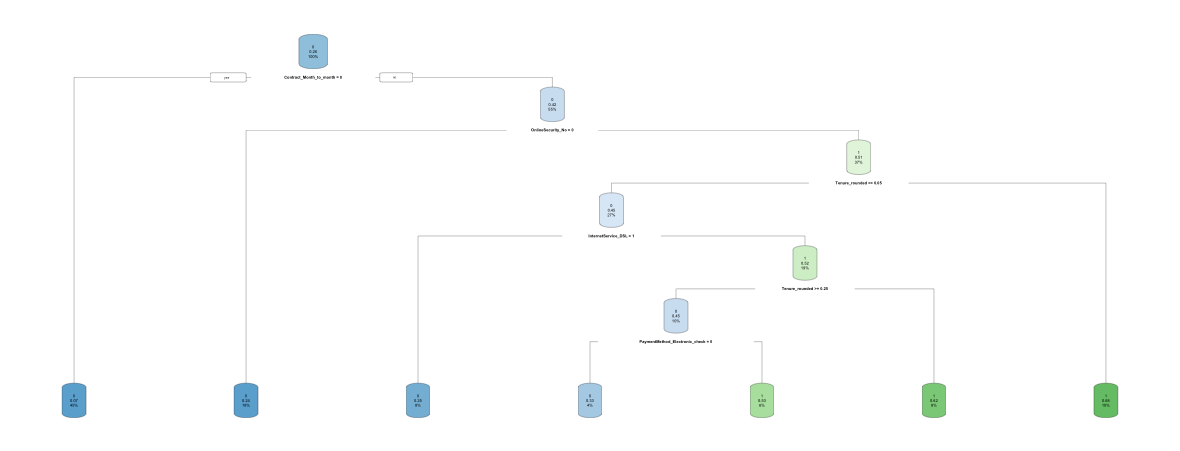
The main advantages of decision tree algorithm are its feature categorisation simplicity, plain visualisation sight, which is similar to how human brain thinks. However, it is not widely applicable for the modelling on the big sets of data, as highly dimensional decision tree becomes really big and clumsy with meaningless weaker branches. It also often loses its generalisation strengths, which means it may not be widely applicable for unseen sets of data. This also cause cause general classifier overfit on test data.

Following the application of the decision tree model on testing data the following results are being given

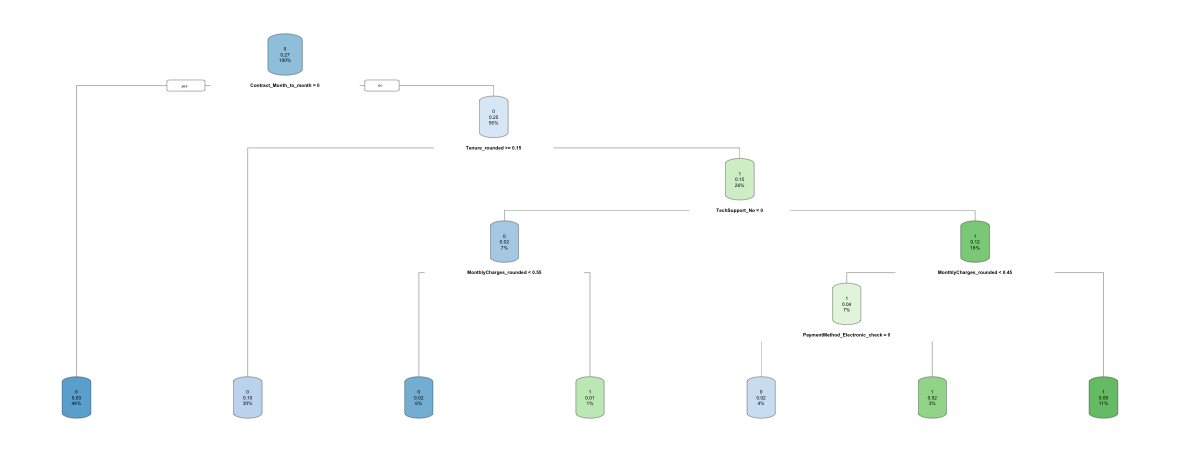
*Decision tree evaluation*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FN** | **TN** | **FP** | **accuracy** | **pgood** | **pbad** | **FPR** | **TPR** |
|
| 1405 | 268 | 298 | 130 | 81.06 | 92 | 52 | 30.38 | 83.98 |

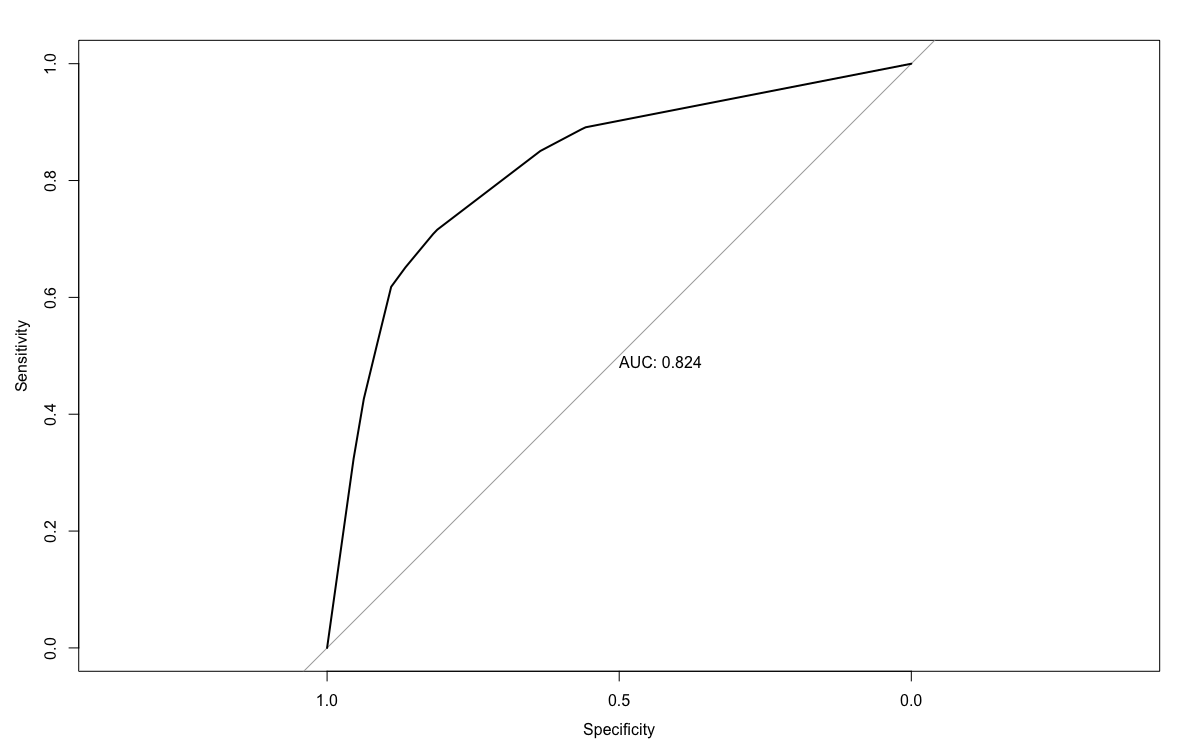
*Decision tree visualisation applied on training data*



*Decision tree visualisation applied on testing data*



*Roc curve decision tree evaluation*



**7.0 Modelling Business Evaluation**

*7.1 Modelling Business Application*

The following table incorporates the confusion metrics results of 3 model variations and highlights in bold the highest performing value within the given 9 catergories (i.e. from TP to TPR).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FN** | **TN** | **FP** | **accuracy** | **pgood** | **pbad** | **FPR** | **TPR** |
|
| LDA | 139 | 139 | **1203** | **335** | 77.43 | 55.86 | **89.64** | 21.8 | 75.3 |
| ROC LDA | 1228 | 146 | 420 | 307 | 78.43 | 80 | 74.25 | **42.22** | **89.37** |
| Decision tree | **1405** | **268** | 298 | 130 | **81.06** | **92** | 52 | 30.38 | 83.98 |

Decision tree produced by far the highest number of TP and FN responses as well as inevitably showed great results in accuracy and precision good. It can be argued that Decision tree could be the best model to predict the churn rate of subsribers in the generalised cases.

LDA produced by far the best results in TN (when the customers do not churn) and precision bad rate. It can be argued that this is the best algorithm to use for the detection of loyal customers and results of effective retention policies.

LDA based on the optimum ROC threshold level inevitably produced the best results for FPR and TPR proving the theory ROC best threshold concepts. Yet, the model gives only average results for all other 7 metrics.

Thus, the telecommunication business does effectively have a choice of two robust models. When data is collected over the long term and thus big and highly dimentional LDA model can be effectively applied which allow business to focus on its loyal client base and development of effective long run retention policies. On the other hand where data is small and has been collected in a relatively short time span the decision tree model can be used to determine problematic subcribers who are likely to leave soon and thus take a swift action.

*7.2 “What-if” and Sensatitivity Business Evaluation*

To further estimate the benefit of deleloped models LDA and Decision tree “What-if” incentive policy scenarios have been examined. The business may likely to implement ‘5%’ or ‘10%’ incentive policies for preventing customers to churn (Decision tree, short run policy) or building loyalty with existing customer base (LDA, long run policy). It has been also assumed that the cost of aquiring a new customer would be equal to 800 USD.

*1. Summary of “What-if” assumptions used*

|  |  |
| --- | --- |
| Tota average charges | USD 2,283 |
| Incentive 1, 10% | 0.1 |
| 1-Incentive, % | 0.9 |
| Subscriber replacement cost | USD 800 |
| Incentive 2, 5% | 0.05 |
| 1-Incentive 2, % | 0.95 |

The obtained in [Modelling and Technical Evaluation](#Data_Modelling_2) confusion matrix results of both Decision Tree and LDA model are being used alongside the 5% and 10% inceptive use in tables below. Camputation formulas are also given.

Table 2 and 4 & table 6 and 8 present initial uplift benefit costs for both Decision Tree and LDA model for 5 and 10% incentives applied, respectively.

5% incentive gives uplift amount of 2,206,151 USD for Decision tree and -165,307 USD cost for LDA.

10% incentive gives 2,030,931 USD for Decision Tree and -219,414 USD cost for LDA.

*2. Decision tree, short run policy with 5% incentive applied*

|  |  |  |
| --- | --- | --- |
| Short run uplift | Decision Tree |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| 5% spent on right customer group | TP\*95%\*Total Charges | USD 3,047,234 |
| wrongly enticed customers | -FP\*5%\*Total Charges | -USD 14,840 |
| Not enticed enough | -FN\*Total Charges | -USD 611,844 |
| Subscriber replacement | -FN\*800 | -USD 214,400 |
| Total uplift |  | USD 2,206,151 |
|  |  |  |

*3. Decision tree, short run policy with 5% incentive applied*

|  |  |  |
| --- | --- | --- |
| Short run uplift | Decision Tree |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| Customers retention | TP\*5%\*TotalCharges | USD 160,381 |
| wrongly enticed customers | -FP\*5%\*TotalCharges | USD 14,840 |
| Subscriber replacement | -FN\*800 | USD 238,400 |
| Invested based on predictions |  | USD 413,620 |

*4. LDA, long run policy with 5% incentive applied*

|  |  |  |
| --- | --- | --- |
| Long run uplift | LDA |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| 5% spent on right customer group | TP\*95%\*Total Charges | USD 301,470 |
| wrongly enticed customers | -FP\*5%Total Charges | -USD 38,240 |
| Not enticed enough | -FN\*Total Charges | -USD 317,337 |
| Subscriber replacement | -FN\*800 | -USD 111,200 |
| Total uplift |  | -USD 165,307 |

*5. LDA, long run policy with 5% incentive applied*

|  |  |  |
| --- | --- | --- |
| Long run uplift | LDA |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| Customers retention | TP\*5%\*TotalCharges | USD 15,867 |
| wrongly enticed customers | -FP\*5%\*TotalCharges | USD 38,240 |
| Subscriber replacement | -FN\*800 | USD 111,200 |
| Invested based on predictions |  | USD 165,307 |

The tables 7 and 9 below present estimated required investments from the deployment of both Decision trees and LDA models.

LDA requires 219,414 USD in comparison to Decision tree required investment of much higher 588,841 USD with 10% inceptive policy applied.

This contrasts to the amount of investments required for 5% incentive policies presented in tables 3 and 5 above ( 413,620 USD for Decision Tree and 165,307 USD for LDA).

*6. Decision tree, short run policy with 10% incentive applied*

|  |  |  |
| --- | --- | --- |
| Short run uplift | Decision Tree |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| 10% spent on right customer group | TP\*90%\*Total Charges | USD 2,886,854 |
| wrongly enticed customers | -FP\*10%\*Total Charges | -USD 29,679 |
| Not enticed enough | -FN\*Total Charges | -USD 611,844 |
| Subscriber replacement | -FN\*800 | -USD 214,400 |
| Total uplift |  | USD 2,030,931 |

*7. Decision tree, short run policy with 10% incentive applied*

|  |  |  |
| --- | --- | --- |
| Short run uplift | Decision Tree |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| Customers retention | TP\*10%\*Total Charges | USD 320,762 |
| wrongly enticed customers | -FP\*10%\*Total Charges | USD 29,679 |
| Subscriber replacement | -FN\*800 | USD 238,400 |
| Invested based on predictions |  | USD 588,841 |

*8. LDA, long run policy with 10% incentive applied*

|  |  |  |
| --- | --- | --- |
| Long run uplift | LDA |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| 10% spent on right customer group | TP\*90%\*Total Charges | USD 285,603 |
| wrongly enticed customers | -FP\*10%Total Charges | -USD 76,481 |
| Not enticed enough | -FN\*Total Charges | -USD 317,337 |
| Subscriber replacement | -FN\*800 | -USD 111,200 |
| Total uplift |  | -USD 219,414 |

*9. LDA, long run policy with 10% incentive applied*

|  |  |  |
| --- | --- | --- |
| Long run uplift | LDA |  |
| Cost/benefit analysis based on cases | Calculation formula | Results |
| Customers retention | TP\*10%\*Total Charges | USD 31,734 |
| wrongly enticed customers | -FP\*10%\*Total Charges | USD 76,481 |
| Subscriber replacement | -FN\*800 | USD 111,200 |
| Invested based on predictions |  | USD 219,414 |

The tables 10-11 below present ROI from deployment of both Decision trees and LDA models. LDA demonstrates a much higher ROI with 60.60% and 48.66% for 10% and 5% incentive policies, respectively.

*10. Decision tree, short run policy with both 10% and 5% incentive and its Return on Investment*

|  |  |  |  |
| --- | --- | --- | --- |
| Short run | Decision Tree |  |  |
| Cost/benefit analysis based on cases | Calculation formula | Results, 10% inceptive | Results, 5% inceptive |
| Model existence benefit | (TP+FN)\*800 | USD 1,338,400 | USD 1,338,400 |
| Additional profit made | Model existence -Total Uplift | -USD 719,242 | -USD 867,751 |
| ROI in % | Investment/Invest Gain | 20.19% | 13.51% |

*11. LDA, long run policy with both 10% and 5% incentive and its Return on Investment*

|  |  |  |  |
| --- | --- | --- | --- |
| Long Run | LDA |  |  |
| Cost/benefit analysis based on cases | Calculation formula | Results, 10% inceptive | Results, 5% inceptive |
| Model existance benefit | (TP+FN)\*800 | USD 222,400 | USD 222,400 |
| Additional profit made | Model existence -Total Uplift | USD 441,814 | USD 387,707 |
| ROI in % | Investment/Invest Gain | 60.60% | 48.66% |

# 8.0 Recommendations

As [“What-if” and Sensatitivity Business Evaluation](#What_IF_Sensativity) section demonstated that Decision tree has shorter term applications over longer run LDA model. Decision tree initial uplift benefits far outweight the LDA ones (where there are no visible uplift benefits but only costs). However, Decision tree policy requires much higher investment and inevitably produces much lower ROI.

Both models show a good robustness and could be argued to be selected as a primary algorithm models. It can therefore be concluded that the use of Decision tree and LDA can be fully justified for the shorter and longer term telecommunications predictions.

# 9.0 References

Brownlee, J. (2019). *Linear Discriminant Analysis for Machine Learning*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/ [Accessed 10 Jun. 2019].

Guru99.com. (2019). [online] Available at: https://www.guru99.com [Accessed 9 Jun. 2019].

Nistrup, P. (2019). *Linear Discriminant Analysis (LDA) 101, using R*. [online] Towards Data Science. Available at: https://towardsdatascience.com/linear-discriminant-analysis-lda-101-using-r-6a97217a55a6 [Accessed 9 Jun. 2019].

Gupta, P. (2019). *Decision Trees in Machine Learning*. [online] Towards Data Science. Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed 10 Jun. 2019].

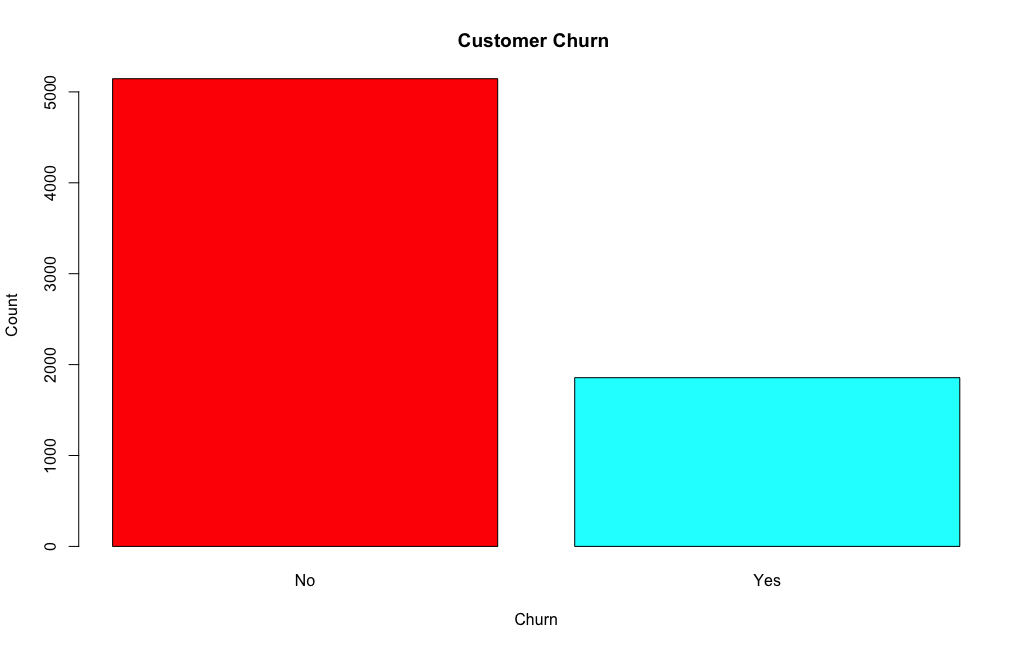
Rdocumentation.org. (2019). *coords function | R Documentation*. [online] Available at: https://www.rdocumentation.org/packages/pROC/versions/1.15.0/topics/coords [Accessed 7 Jun. 2019].

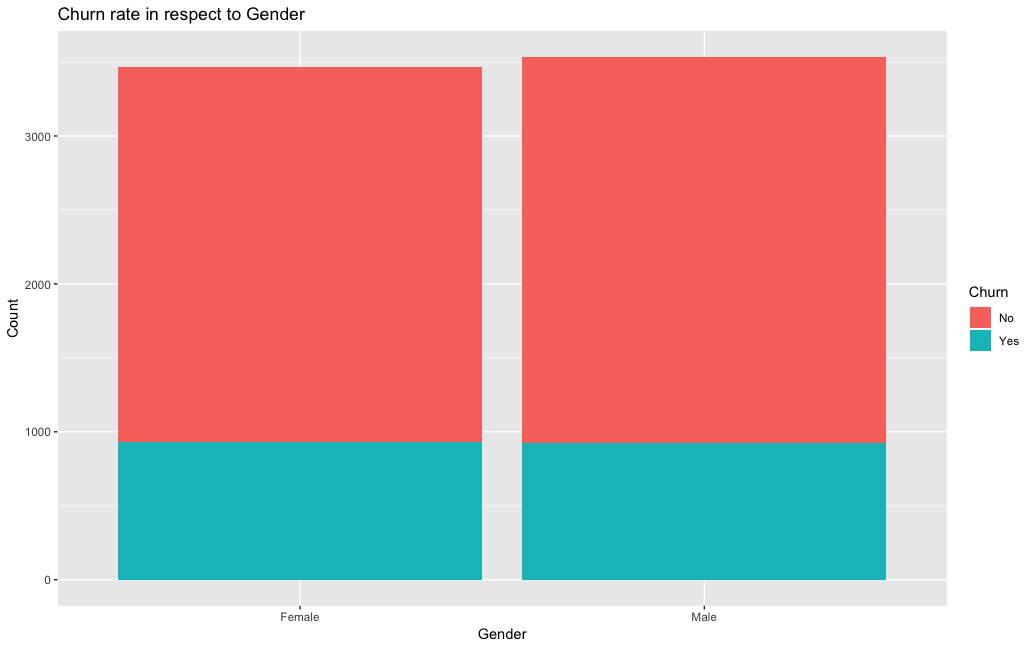
Research.cs.tamu.edu. (2019). [online] Available at: http://research.cs.tamu.edu/prism/lectures/pr/pr\_l10.pdf [Accessed 10 Jun. 2019].

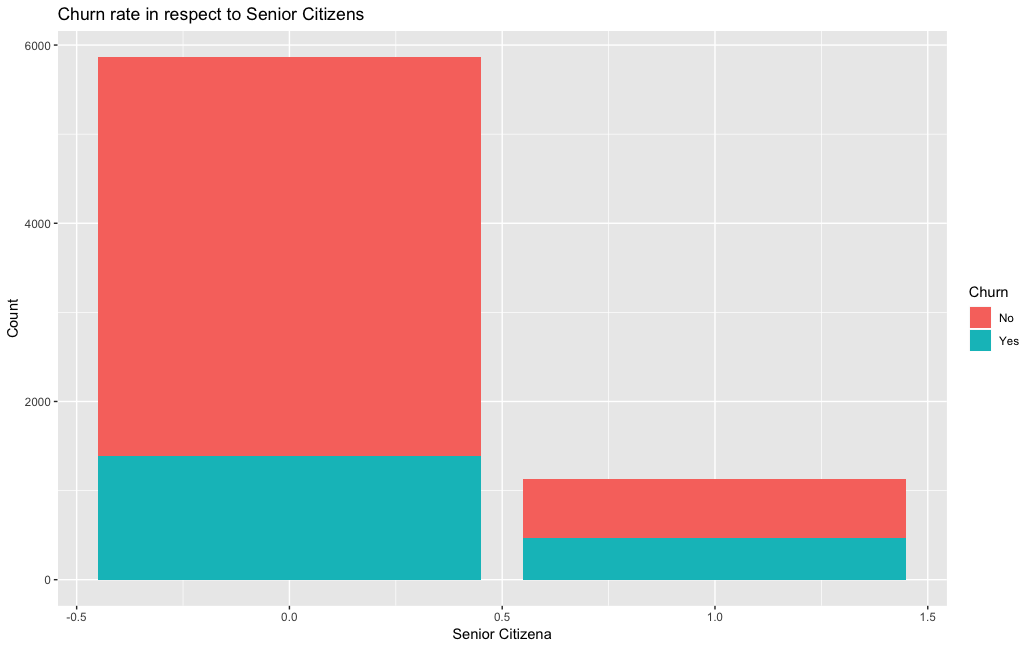
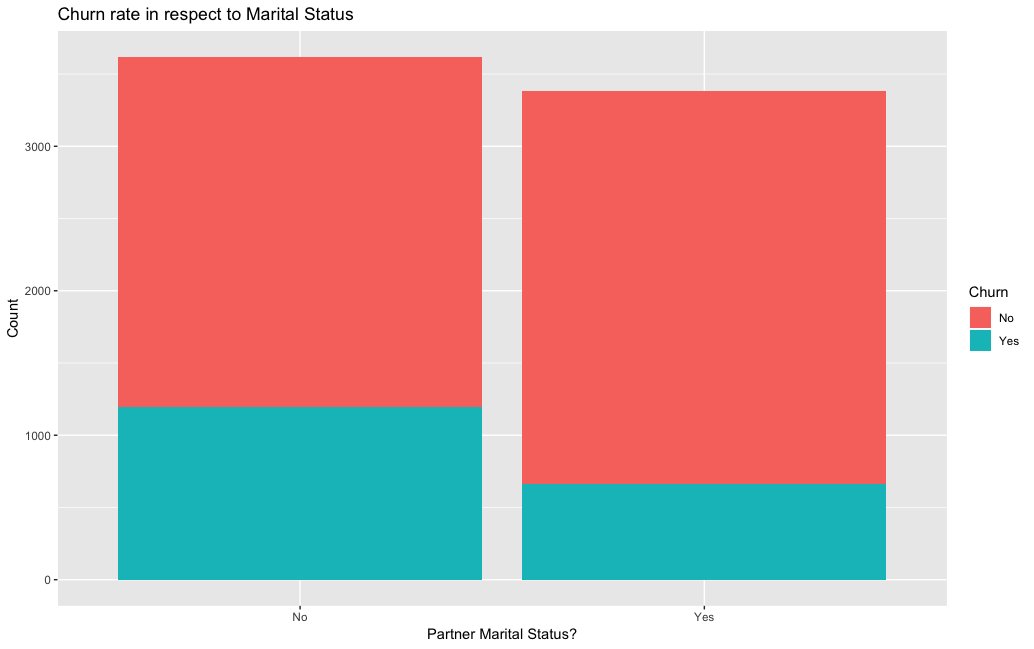
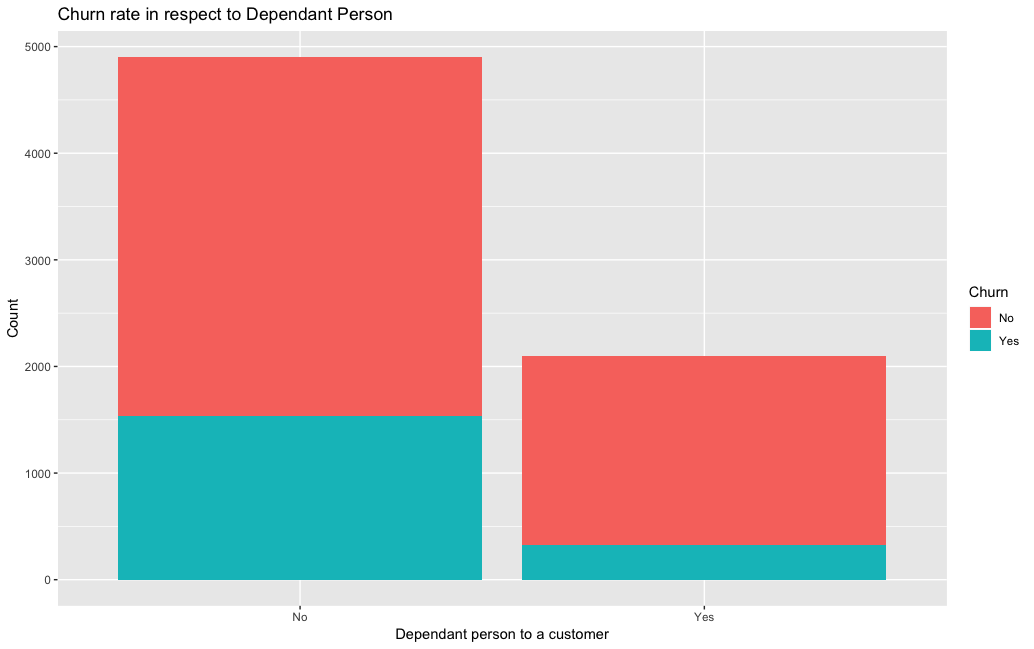
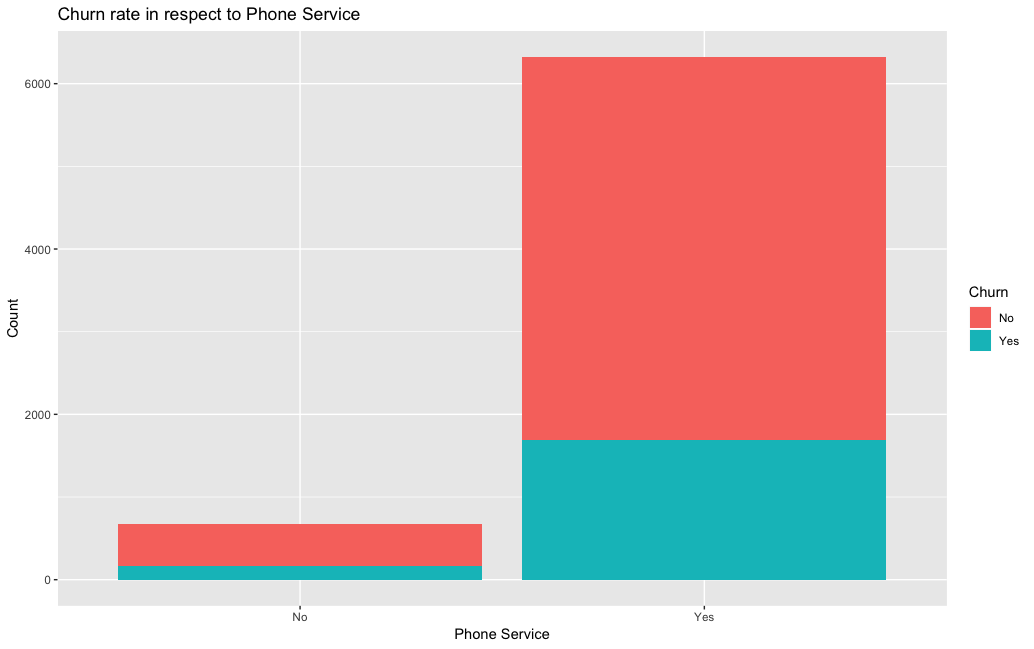
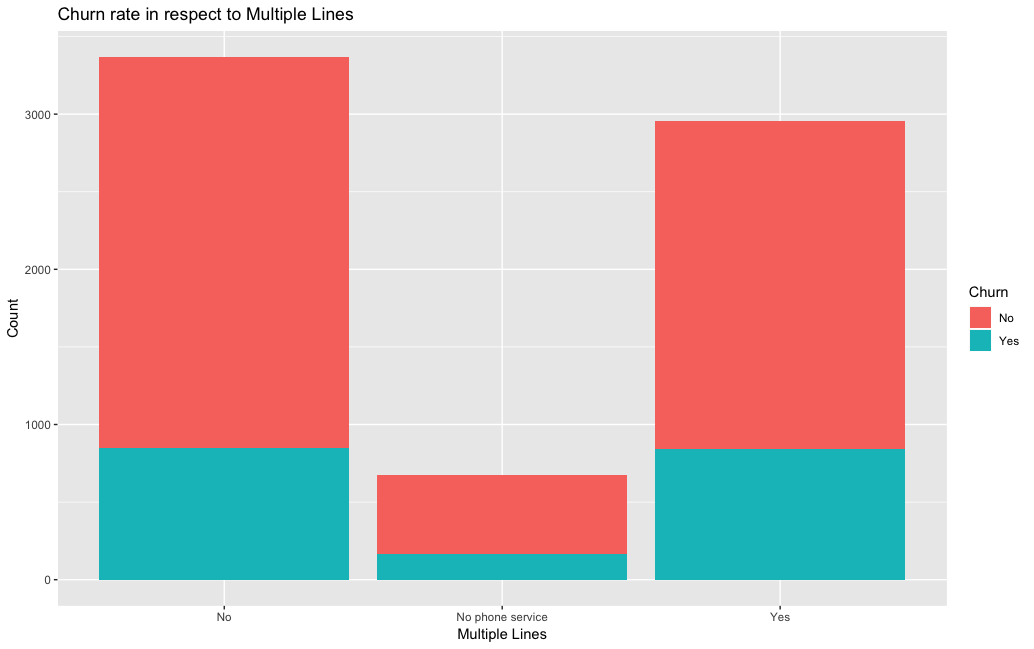
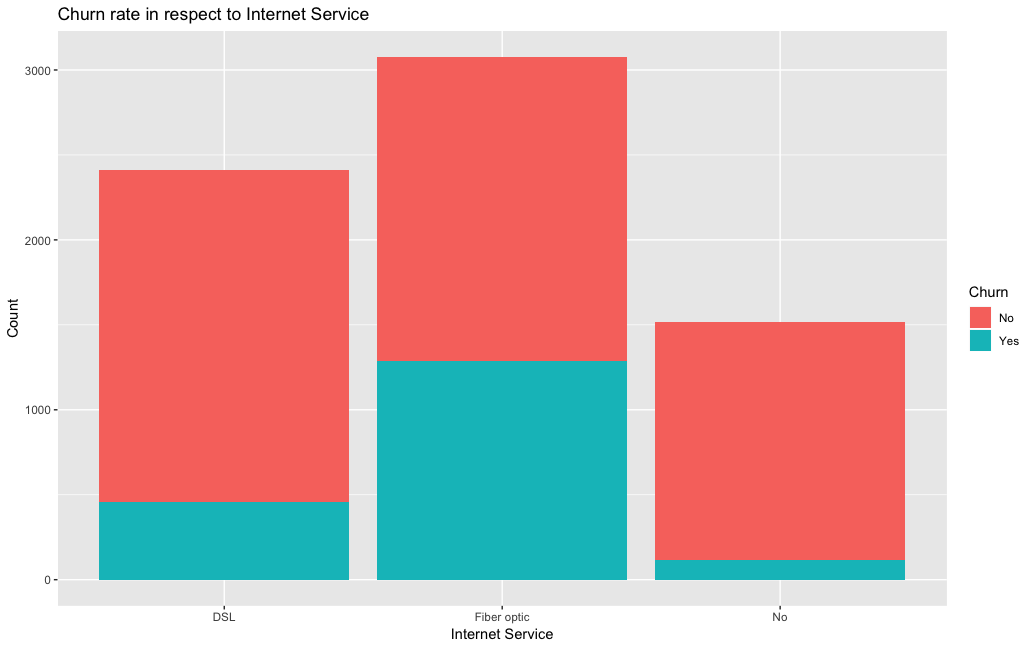
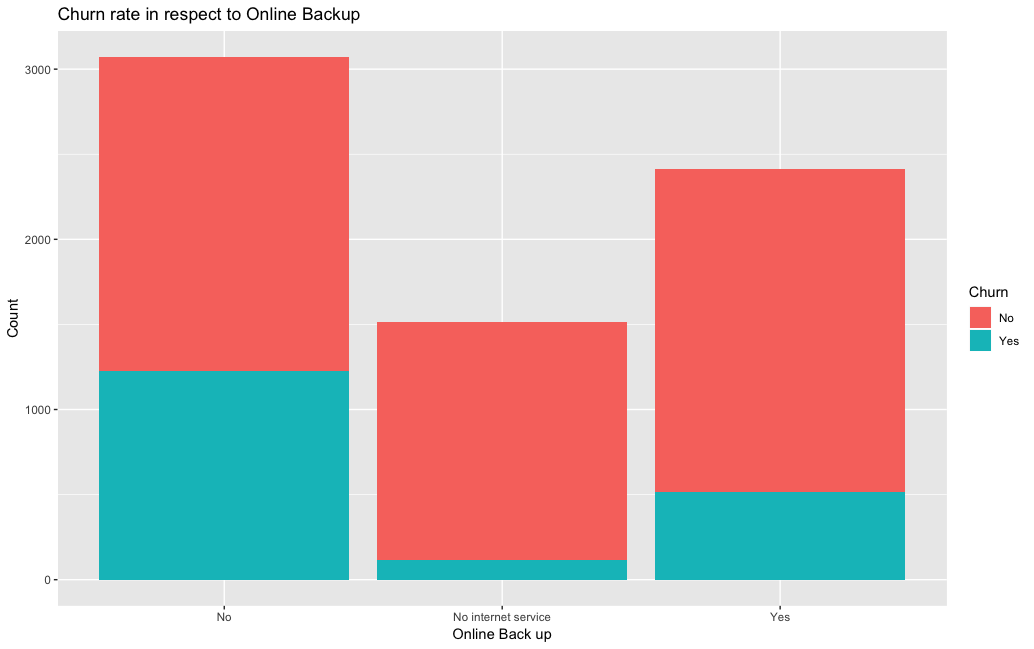
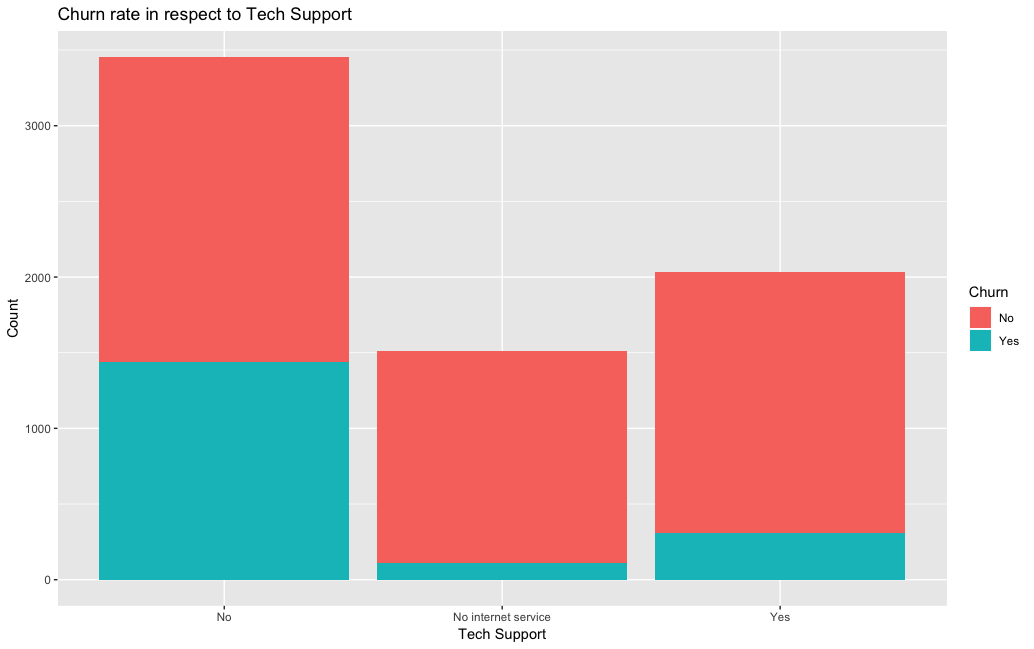
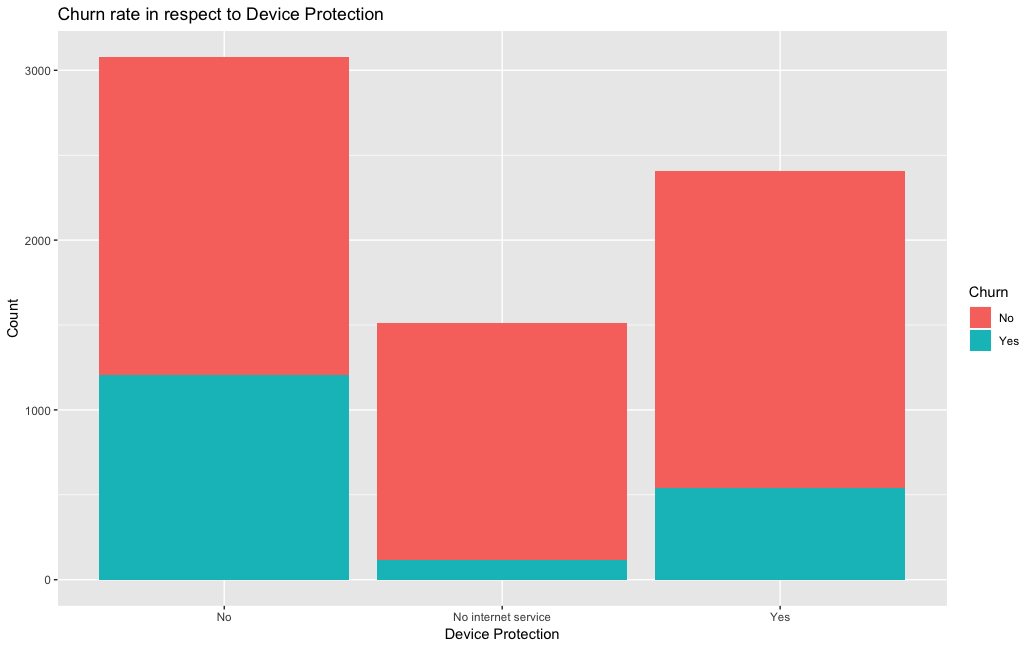
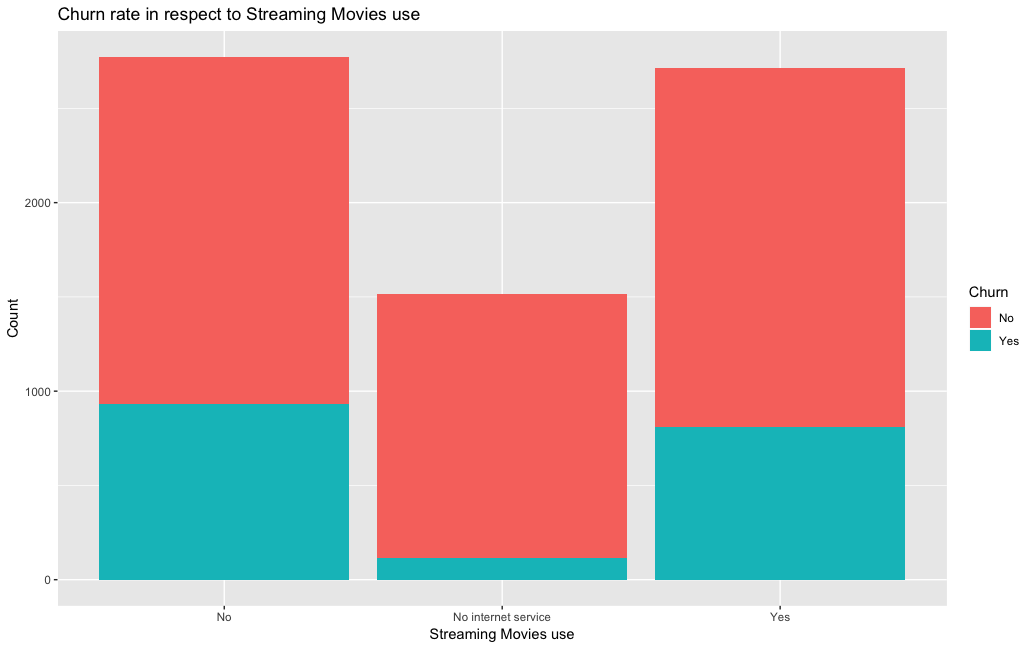
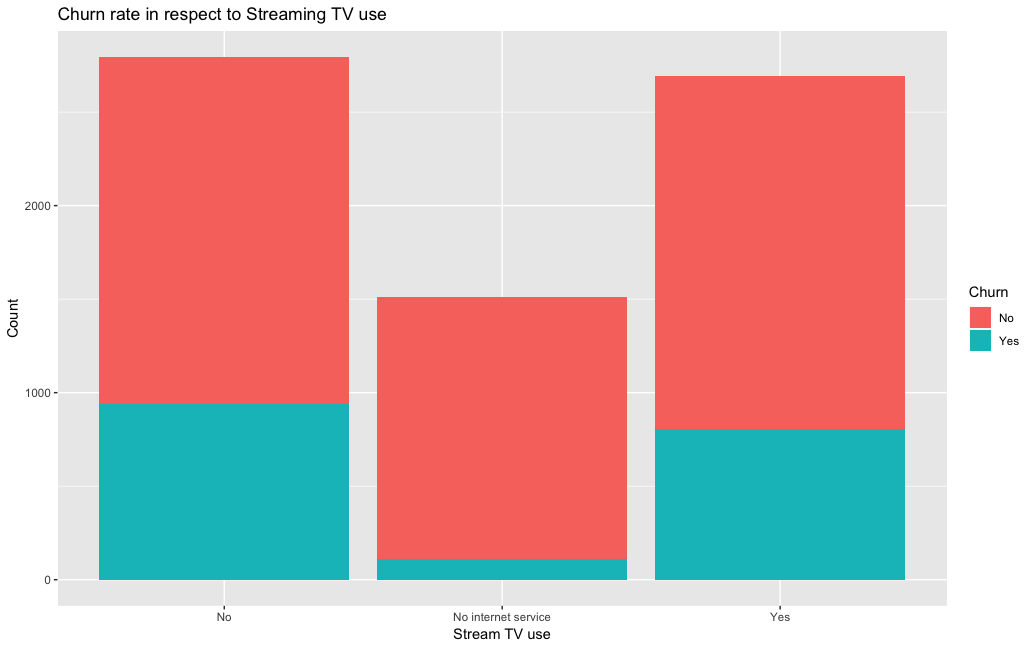
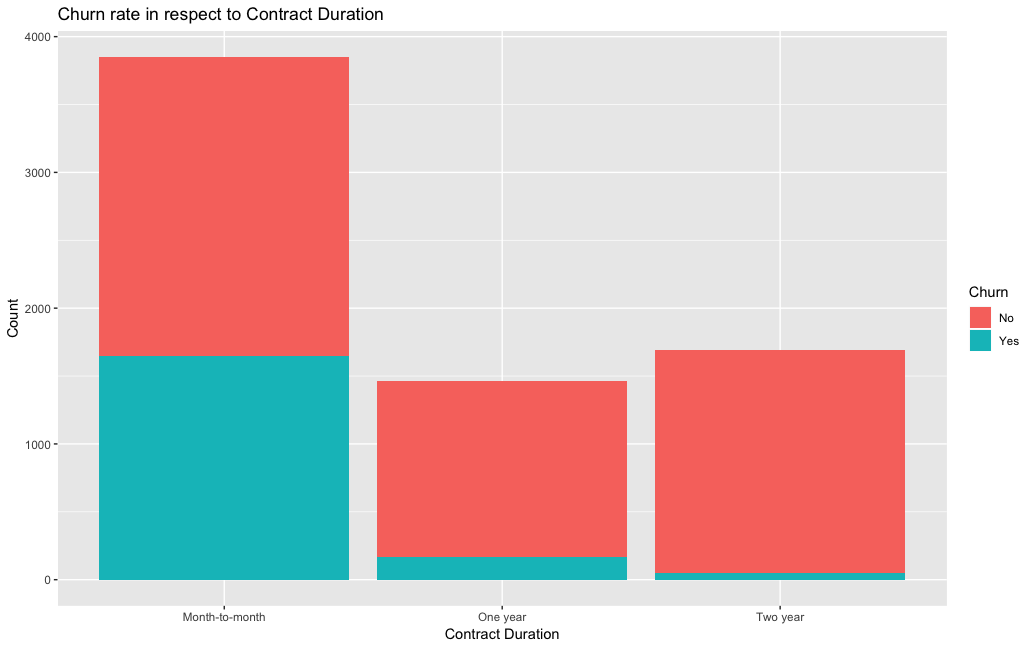
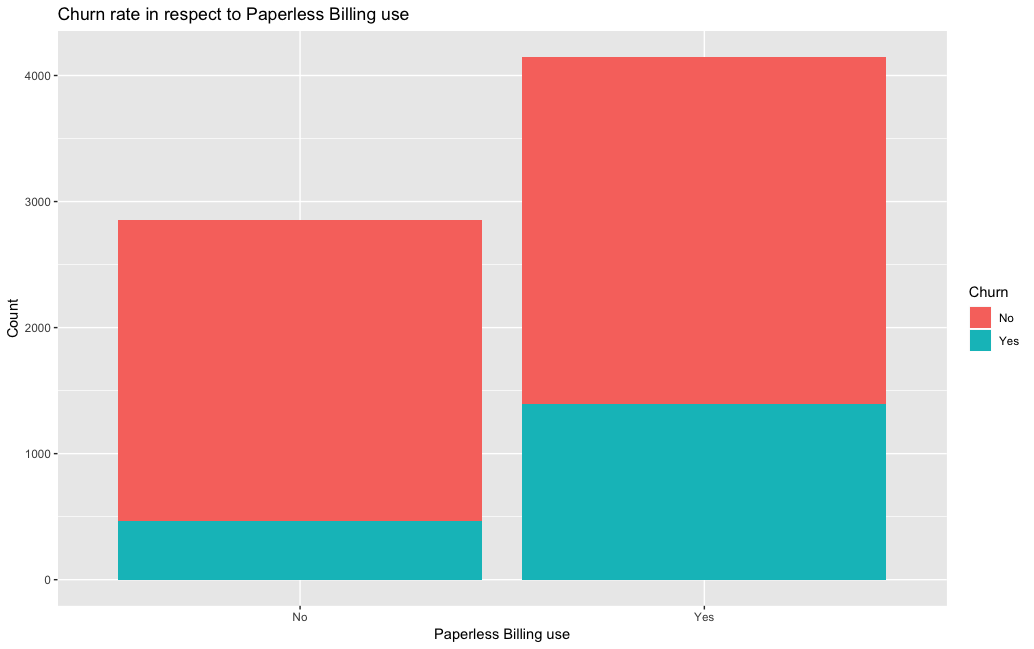
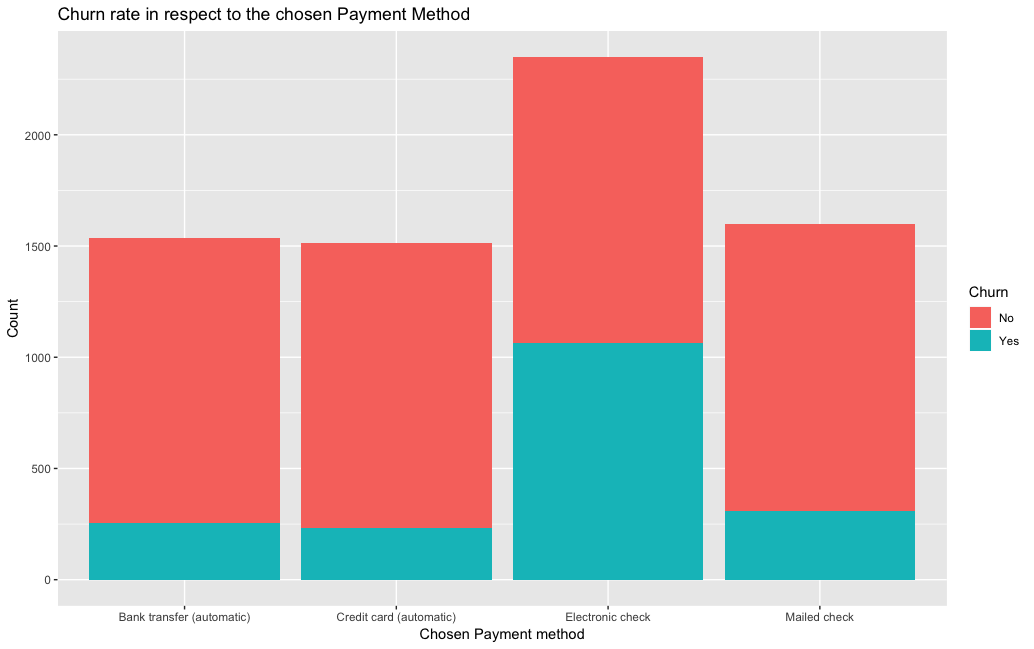
R-tutor.com. (2019). *Scatter Plot | R Tutorial*. [online] Available at: http://www.r-tutor.com/elementary-statistics/quantitative-data/scatter-plot [Accessed 9 Jun. 2019].

# *10.0 Appendix*

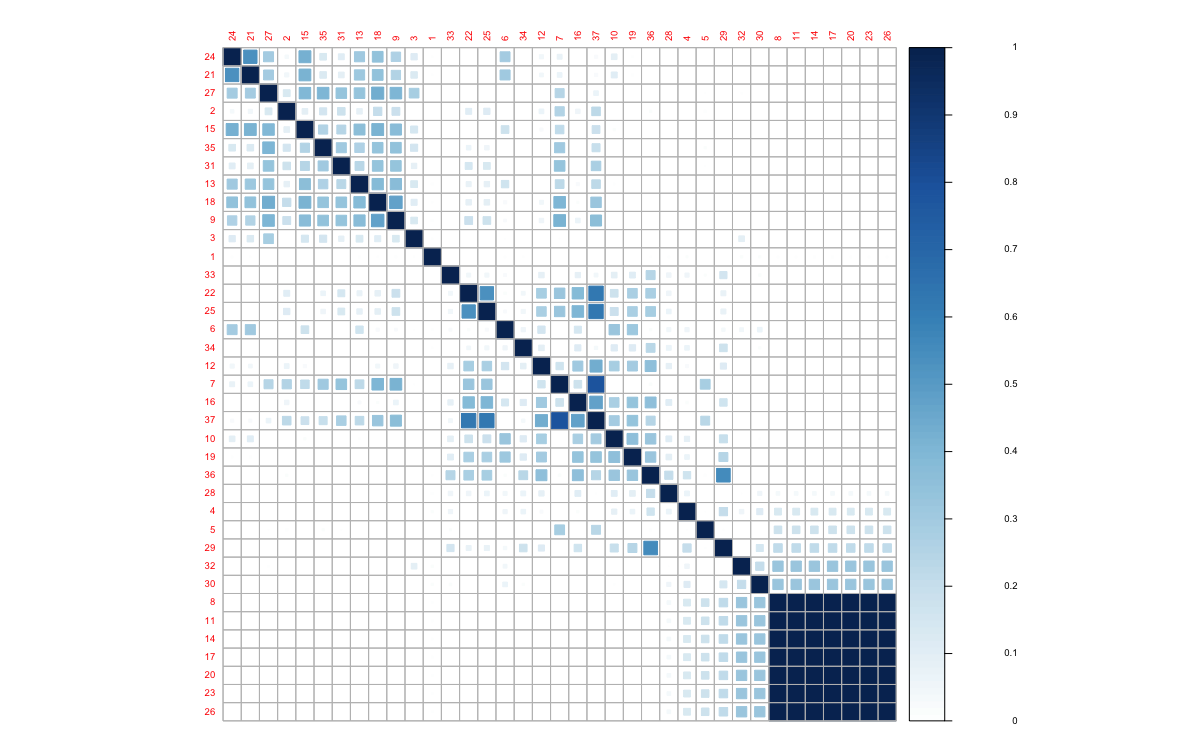
*Visualisations and barplot Graphs*





*Field Correlation plot and field analysis*



*Correlated field and its removal*

"Before redundancy check Fields= 37"

[1] "Following fields are correlated"

row col

MonthlyCharges\_rounded 37 7

OnlineSecurity\_No\_internet\_service 11 8

OnlineBackup\_No\_internet\_service 14 8

DeviceProtection\_No\_internet\_service 17 8

TechSupport\_No\_internet\_service 20 8

StreamingTV\_No\_internet\_service 23 8

StreamingMovies\_No\_internet\_service 26 8

InternetService\_No 8 11

OnlineBackup\_No\_internet\_service 14 11

DeviceProtection\_No\_internet\_service 17 11

TechSupport\_No\_internet\_service 20 11

StreamingTV\_No\_internet\_service 23 11

StreamingMovies\_No\_internet\_service 26 11

InternetService\_No 8 14

OnlineSecurity\_No\_internet\_service 11 14

DeviceProtection\_No\_internet\_service 17 14

TechSupport\_No\_internet\_service 20 14

StreamingTV\_No\_internet\_service 23 14

StreamingMovies\_No\_internet\_service 26 14

InternetService\_No 8 17

OnlineSecurity\_No\_internet\_service 11 17

OnlineBackup\_No\_internet\_service 14 17

TechSupport\_No\_internet\_service 20 17

StreamingTV\_No\_internet\_service 23 17

StreamingMovies\_No\_internet\_service 26 17

InternetService\_No 8 20

OnlineSecurity\_No\_internet\_service 11 20

OnlineBackup\_No\_internet\_service 14 20

DeviceProtection\_No\_internet\_service 17 20

StreamingTV\_No\_internet\_service 23 20

StreamingMovies\_No\_internet\_service 26 20

MonthlyCharges\_rounded 37 22

InternetService\_No 8 23

OnlineSecurity\_No\_internet\_service 11 23

OnlineBackup\_No\_internet\_service 14 23

DeviceProtection\_No\_internet\_service 17 23

TechSupport\_No\_internet\_service 20 23

StreamingMovies\_No\_internet\_service 26 23

MonthlyCharges\_rounded 37 25

InternetService\_No 8 26

OnlineSecurity\_No\_internet\_service 11 26

OnlineBackup\_No\_internet\_service 14 26

DeviceProtection\_No\_internet\_service 17 26

TechSupport\_No\_internet\_service 20 26

StreamingTV\_No\_internet\_service 23 26

InternetService\_Fiber\_optic 7 37

StreamingTV\_Yes 22 37

StreamingMovies\_Yes 25 37

[1] "Removing the following fields"

[1] "InternetService\_No"

[2] "InternetService\_No"

[3] "OnlineSecurity\_No\_internet\_service"

[4] "InternetService\_No"

[5] "OnlineSecurity\_No\_internet\_service"

[6] "OnlineBackup\_No\_internet\_service"

[7] "InternetService\_No"

[8] "OnlineSecurity\_No\_internet\_service"

[9] "OnlineBackup\_No\_internet\_service"

[10] "DeviceProtection\_No\_internet\_service"

[11] "InternetService\_No"

[12] "OnlineSecurity\_No\_internet\_service"

[13] "OnlineBackup\_No\_internet\_service"

[14] "DeviceProtection\_No\_internet\_service"

[15] "TechSupport\_No\_internet\_service"

[16] "InternetService\_No"

[17] "OnlineSecurity\_No\_internet\_service"

[18] "OnlineBackup\_No\_internet\_service"

[19] "DeviceProtection\_No\_internet\_service"

[20] "TechSupport\_No\_internet\_service"

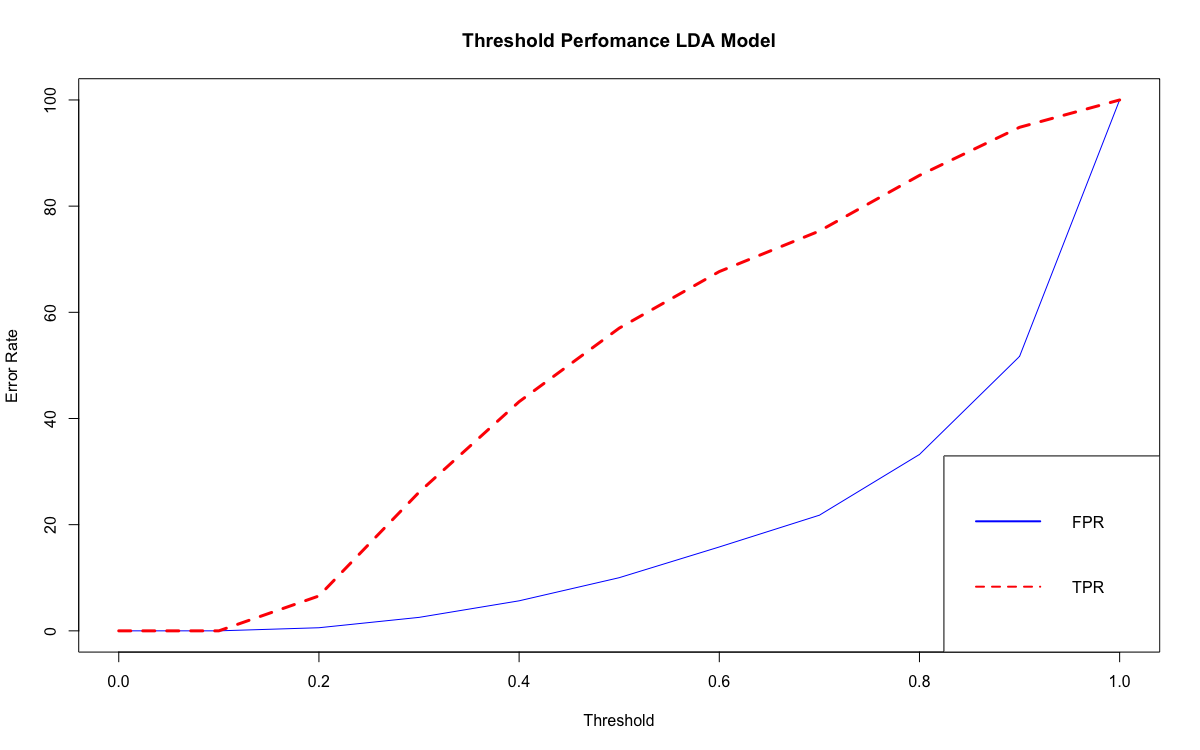
[21] "StreamingTV\_No\_internet\_service"

[22] "InternetService\_Fiber\_optic"

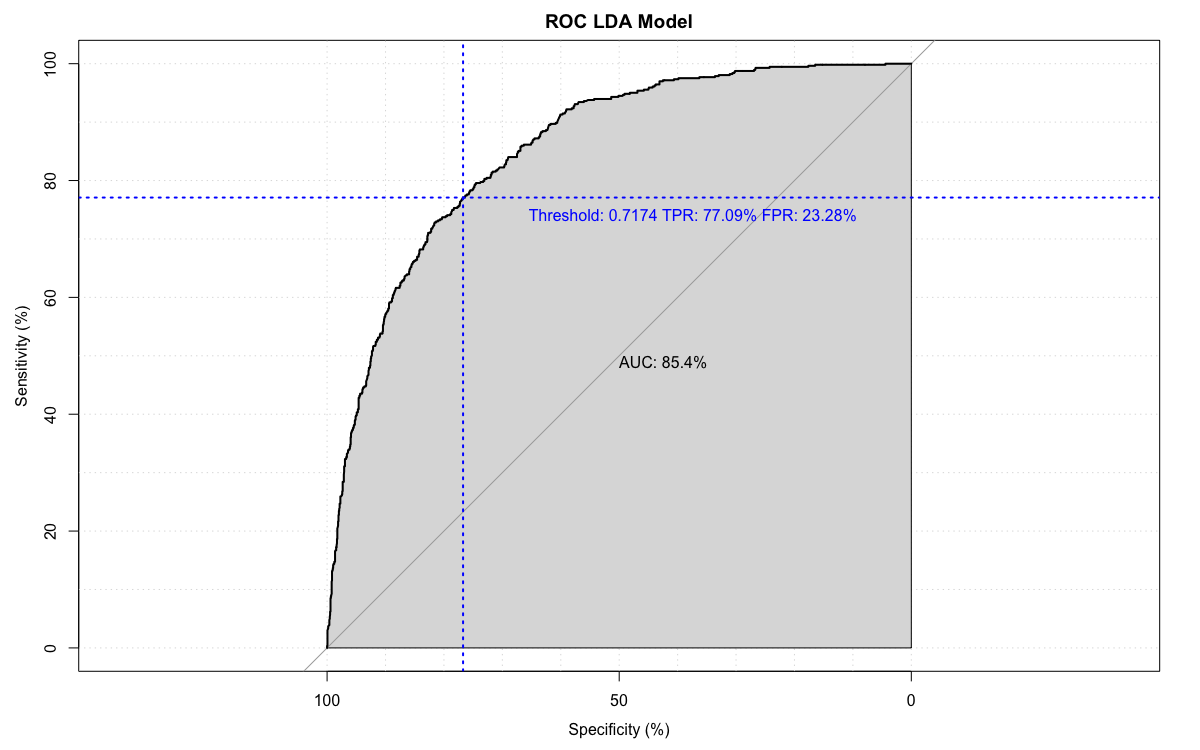
[23] "StreamingTV\_Yes"

[24] "StreamingMovies\_Yes"

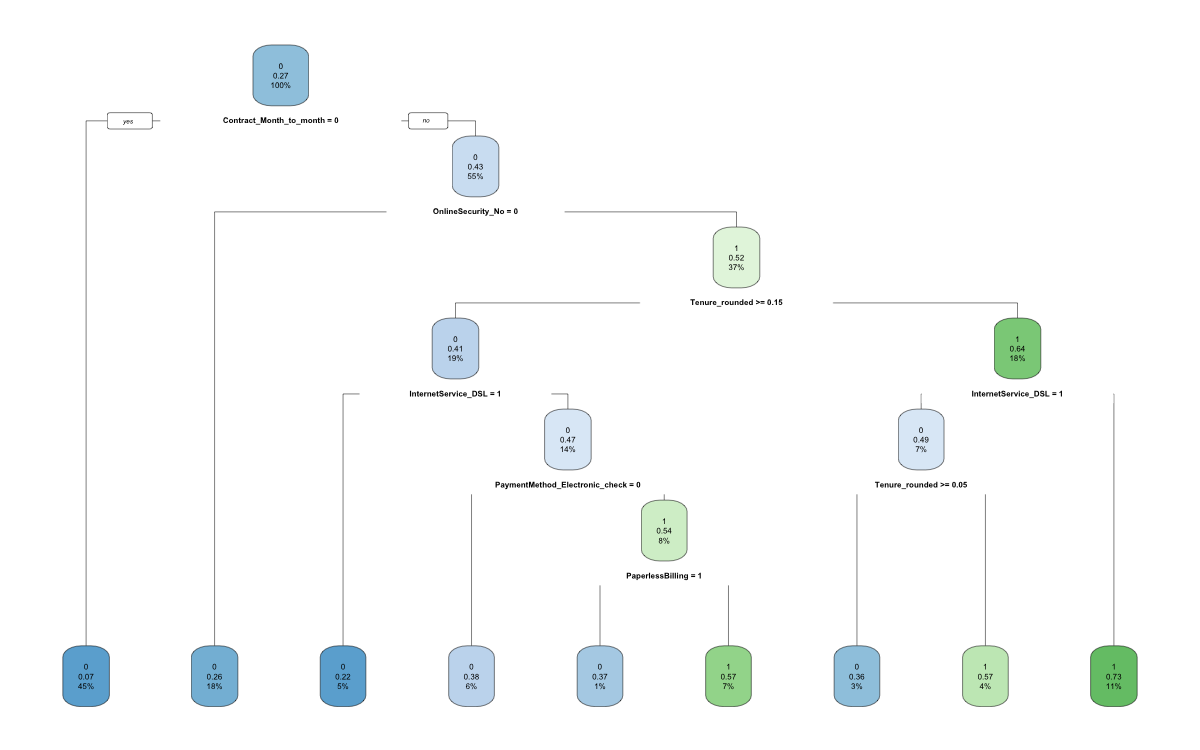
*LDA model threshold and error rate relationship*



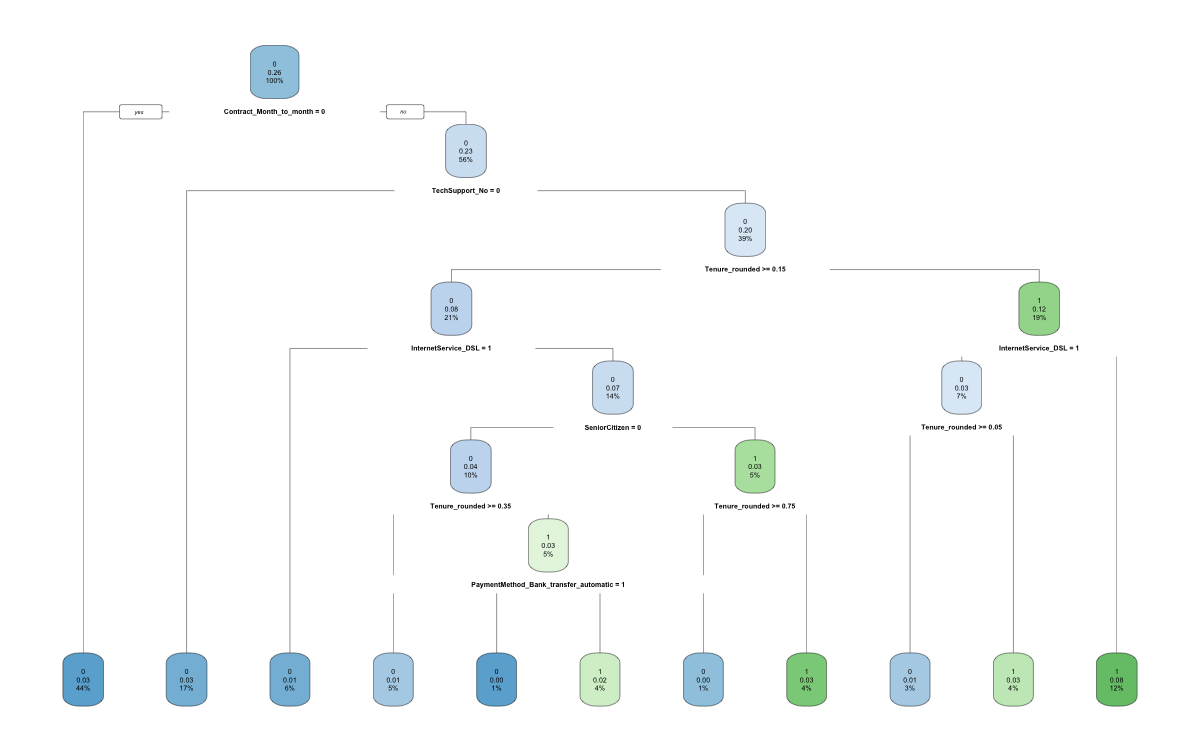
*LDA ROC curve and optimum threshold level*



*Decision tree applied on training data*



*Decision tree on testing model*



*Decision Tree ROC curve applied on testing model*

