Telecommunication Company Churn Analysis

BUAN 6337.501

GROUP MEMBERS: -

- NOEL ABRAHAM (NGA100020)
- RAMESH DASARI (RXD210017)
- ROHAN SRIVASTAVA(RXS210007)
- SAURAV GUPTA (SXG200009)
- VAIBHAV GUPTA (VXG200039)

Topic

Churn Analysis; Predicting whether a customer will leave or continue the Telecom service and finding the most relevant features that contribute to churn.

Overview

This report contains a comprehensive data analysts' approach by which conclusions are drawn that aid the business to retain its customers. Based on this approach, the company can predict whether a customer will churn or not. The data used contains various parameters recorded for each customer with a unique ID. Though the data seemed clean on the off hand, further cleaning/wrangling is carried out to make the data more usable. This is followed by exploratory analysis of the dataset to understand the underlying patterns. A few models are then proposed which help understand the actual significance of each parameter and help predict churn. Finally, conclusions are made and based on these, some recommendations are proposed that would enable the company to increase its profits by reducing customer churn.

Table of Content

Objective

The aim of this project is to carry out a Churn analysis of the dataset containing records of customers of a Telecommunication company. Churn refers to the process of measuring the rate at which customers quit the service. It is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers. Our final objective is to answer the questions like Which variables are most responsible for customer churn? and what actions can be taken to reduce churn rate? and to predict whether a customer would churn or not given particular service parameters. We also aim to suggest actionable insights to allow the company to take action and increase its profit. At the end of this report, we expect lower churn rates leading to happier customers, larger margins, and higher profits.

Data Overview and Preprocessing

The dataset Used in this project is obtained from the Kaggle repository and can be accessed from the following link: - https://www.kaggle.com/datasets/blastchar/telco-customer-churn, available in the CSV format.

The size of the dataset is rather small, with only 7043 observations and 21 attributes. The list of attributes, with transformed data types for analysis, are in the table below. Most features are categorical and are coded as dummy variables such that base and reference values are set for each categorical feature. (Refer appendix for exact encoding)

SNo	Attribute	Data Type
1	Customer ID	Character
2	Gender	Categorical (Binary)
3	Senior Citizen	Categorical (Binary)
4	Partner	Categorical (Binary)
5	Dependents	Categorical (Binary)
6	Tenure	Integer
7	Phone Service	Categorical (Binary)
8	Multiple Lines	Categorical (Binary)
9	Internet Service	Categorical (Dummy)
10	Online Security	Categorical (Binary)
11	Online Backup	Categorical (Binary)
12	Device Protection	Categorical (Binary)
13	Tech Support	Categorical (Binary)
14	Streaming TV	Categorical (Binary)
15	Streaming Movies	Categorical (Binary)
16	Contract	Categorical (Dummy)
17	Paperless Billing	Categorical (Binary)
18	Payment Method	Categorical (Dummy)
19	Monthly Charges	Integer
20	Total Charges	Integer
21	Churn	Categorical (Binary)

Features of dataset along with converted data type

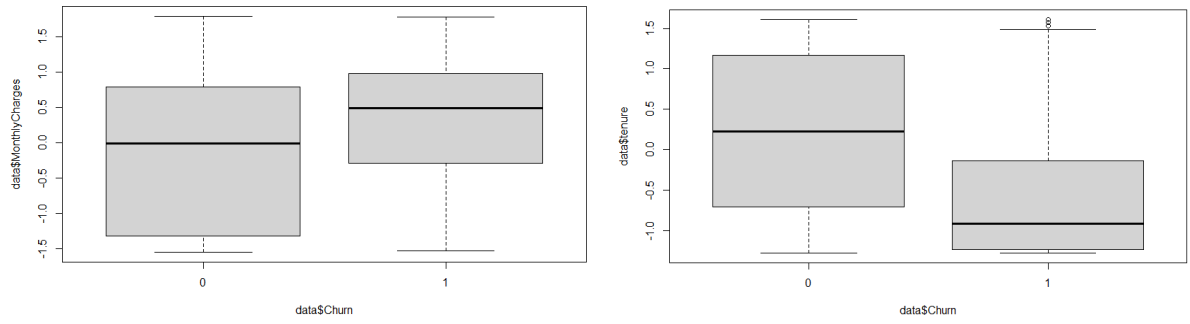
The attribute Customer ID is dropped as it would not serve any purpose for further analysis. Now, our data contains 19 independent predicting variables, which can be classified into 3 groups:

1. Demographics: Gender, Senior Citizen, Partner, Dependents
2. Company Observation Data: tenure, contract, paperless Billing, Payment Method, monthly Charges, Total Charges
3. Services: phone Service, Multiple Lines, Internet Service, Online Security, Online backup, Device Protection, Tech Support, Streaming TV, Streaming Movies

Further exploration revealed that the attribute 'Total Charges' had 11 missing values. Since it represents a very small portion of the data, these records were omitted from the dataset instead of mode/mean imputation.

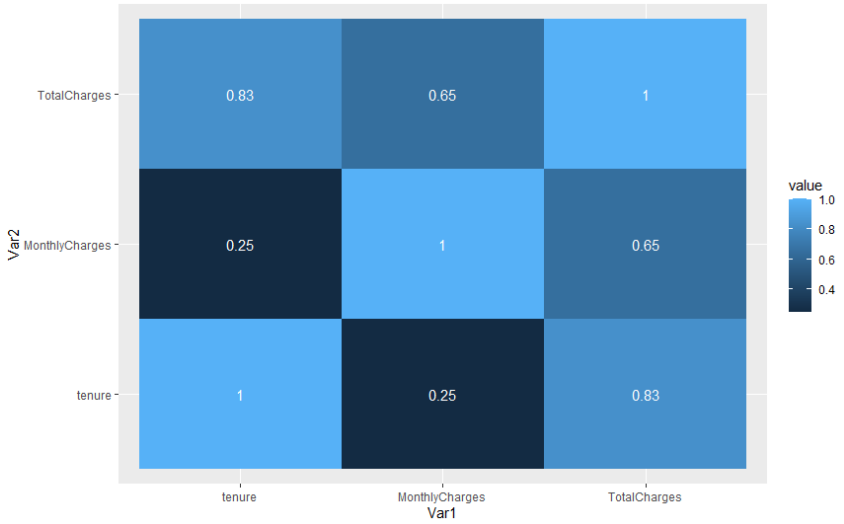
From the table above, we can also observe that there are 3 continuous variables in the dataset and before any further analysis is carried out, they are transformed to scaler form so that they can be utilized the same way across all of the records during model analysis.

There is a strong possibility of outliers skewing the data. Therefore, for each continuous variable, the data distribution box plots are generated. Inter-Quartile range is estimated and data is trimming is carried out to remove outliers. Since a large portion of the data would have to be appended in order to account for only a few outliers, the entire dataset is kept as it is.



Boxplot of Continuous Variables distribution and their outliers

There is also a possibility of correlation between the continuous variables. Intuitively, there is a high chance of collinearity between Tenure, Monthly charges and Total charges. To account for it, we can plot a heat map and check for the same.



Heat Map: Continuous Variables

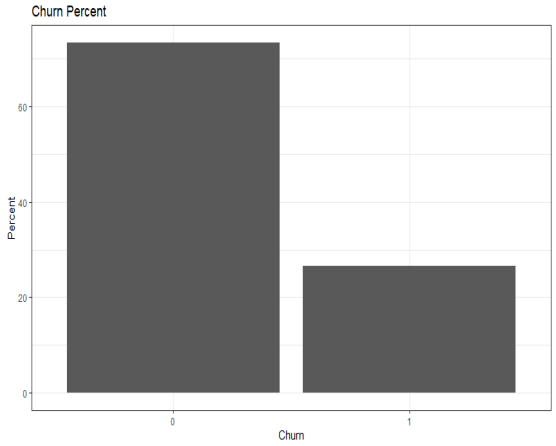
From the above heat map, it is evident that Total charges is highly correlated to both Monthly charges and tenure, and hence would cause an issue of multi-collinearity in any model. Therefore, we drop this attribute from the dataset.

Exploratory Data Analysis

In order to further understand the data and some underlying patterns, we perform and exploratory analysis of the dataset. The purpose of this section was to draw out how the Features affected to outcome Churn and if there was any evidence of significant effect of these features on churn. We looked for obvious errors, patterns within the data and analyzed relations among the variables. The purpose of this section is to ensure that the results produced are valid and applicable to our business outcomes and goals.

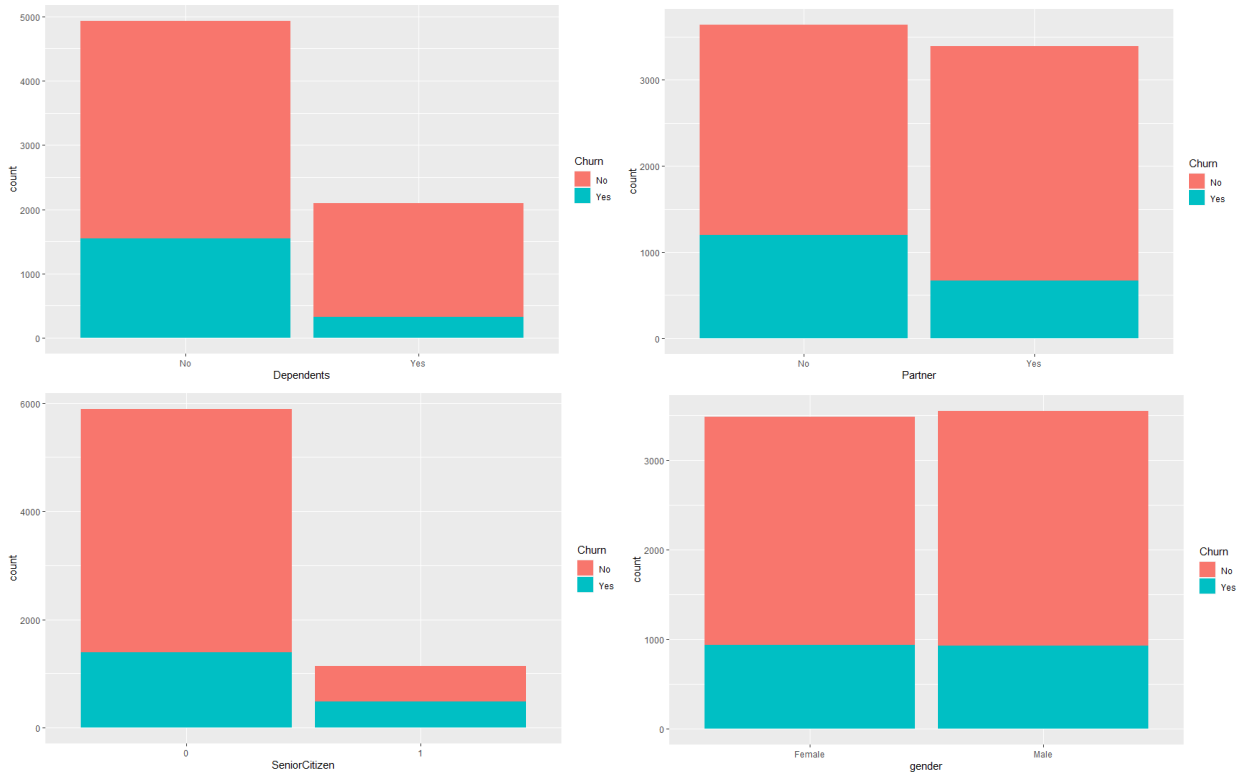
	vars	n	mean	sd	median	trimmed	mad	min	max
tenure	1	7032	0.00	1.00	-0.14	-0.04	1.33	-1.28	1.61
MonthlyCharges	2	7032	0.00	1.00	0.18	0.01	1.19	-1.55	1.79
TotalCharges	3	7032	0.00	1.00	-0.39	-0.14	0.80	-1.00	2.82
gender	4	7032	0.50	0.50	1.00	0.51	0.00	0.00	1.00
SeniorCitizen	5	7032	0.16	0.37	0.00	0.08	0.00	0.00	1.00
Partner	6	7032	0.48	0.50	0.00	0.48	0.00	0.00	1.00
Dependents	7	7032	0.30	0.46	0.00	0.25	0.00	0.00	1.00
PhoneService	8	7032	0.90	0.30	1.00	1.00	0.00	0.00	1.00
OnlineSecurity	9	7032	0.29	0.45	0.00	0.23	0.00	0.00	1.00
OnlineBackup	10	7032	0.34	0.48	0.00	0.31	0.00	0.00	1.00
DeviceProtection	11	7032	0.34	0.48	0.00	0.30	0.00	0.00	1.00
TechSupport	12	7032	0.29	0.45	0.00	0.24	0.00	0.00	1.00
StreamingTV	13	7032	0.38	0.49	0.00	0.36	0.00	0.00	1.00
StreamingMovies	14	7032	0.39	0.49	0.00	0.36	0.00	0.00	1.00
PaperlessBilling	15	7032	0.59	0.49	1.00	0.62	0.00	0.00	1.00
Churn	16	7032	0.27	0.44	0.00	0.21	0.00	0.00	1.00
MultipleLines*	17	7032	1.42	0.49	1.00	1.40	0.00	1.00	2.00
InternetService*	18	7032	2.22	0.78	2.00	2.28	1.48	1.00	3.00
Contract*	19	7032	2.03	0.67	2.00	2.04	0.00	1.00	3.00
PaymentMethod*	20	7032	2.42	1.06	2.00	2.40	1.48	1.00	4.00

Data Summary



Customer Distribution by Churn Percent

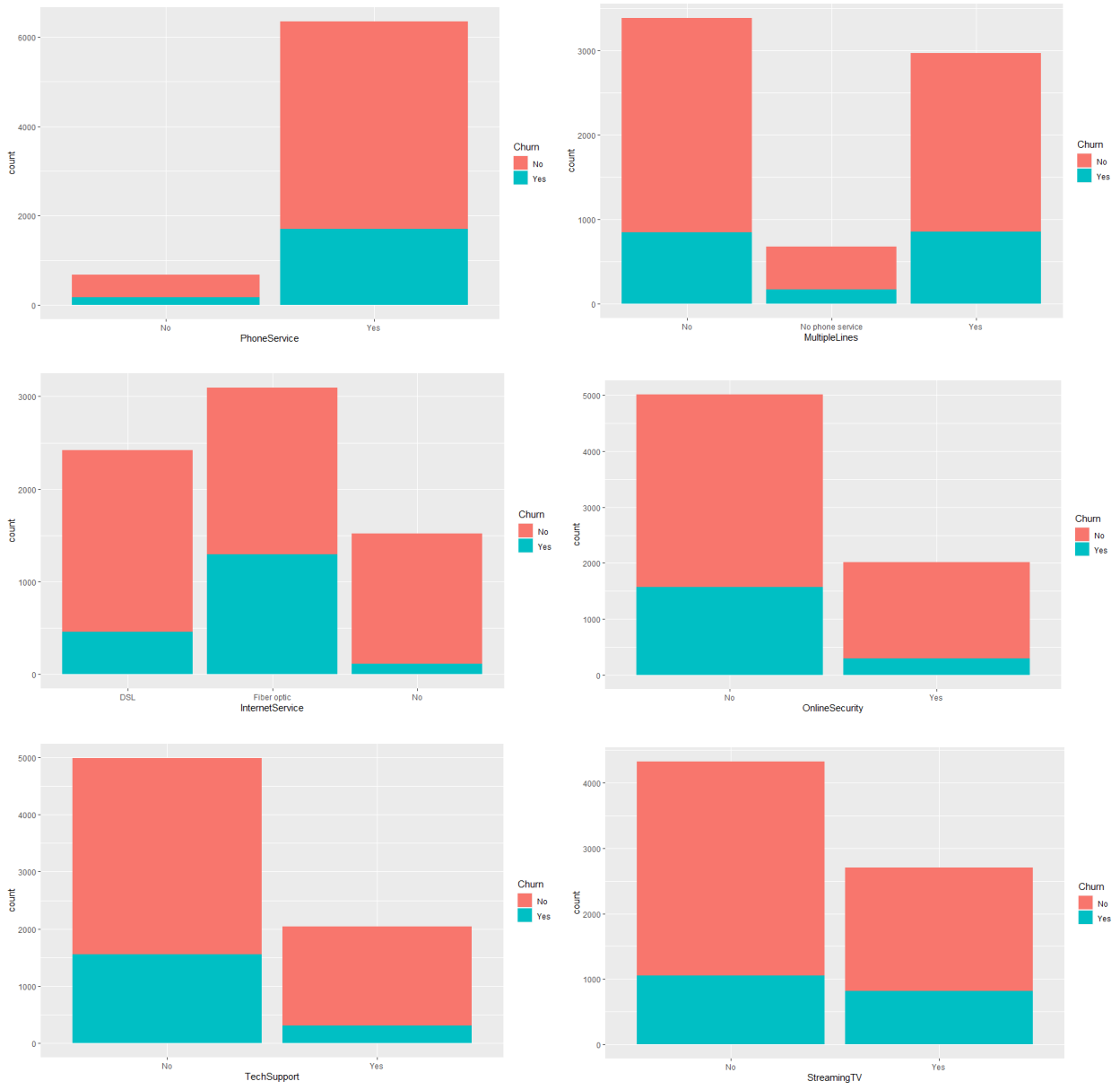
From the above figure, we can observe that the customer churn rate for the company is at 26%. This is ouer target market and to increase profits, we must minimize this 26% of the population set. Further, we can look at the demographical effect on churn.



Demographic Data Distribution

The market mainly comprises of people without dependents and who are not senior citizens. Churn rate for both these categories is comparatively higher and would probably be a good predictor in defining churn.

Gender seems to play no role in deciding the churn rate as the distribution is uniform even within groups. Gender might not be a good predictor for analyzing churn rate.

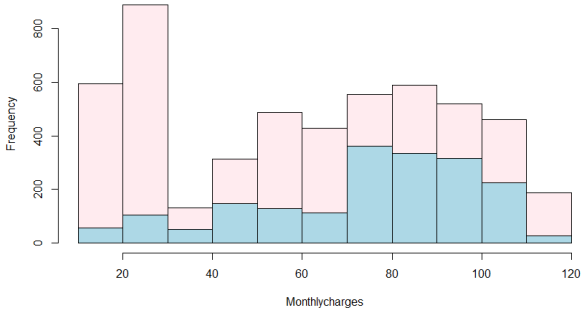


Company Services Data Distribution

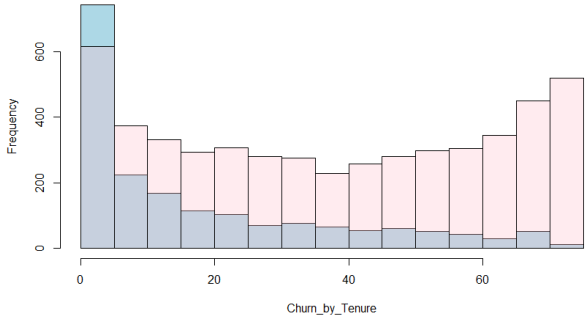
Type quality of services that the company provides has different effects across various variables. It can be observed that the Phone services are taken by majority of the people and about 25% of these people are lost due to churn. It would be advisable to modify the business strategy to assuage these customers.

It can also be seen that most people opt out of online security and the churn rate of this category is higher as compared to when they take security. Therefore, the company could reduce the cost of online security and make it mandatory for all customers. A similar pattern is observed for Tech Support.

More people tend to churn out when the take streaming services, therefore improving the quality of streaming services might be financially beneficial in the long run.



Churn Rate by Monthly Charges



Churn Rate by Tenure

Churn rate seems to be higher when the monthly charges are high and new customer, who have a smaller tenure, seem to churn out more.

By looking at the plots above, the most relevant attributes for detecting churn seem to be Contract, Online Security, Tech Support, Internet Services while the least significant seem to be gender, Partner, Streaming Service. We expect these attributes to be discriminative in our future models.

Inferences from Churn Data

The above analysis gives us a general idea of how the data is distributed and a few superficial findings. In order to accurately and reliably predict customer churn, we need to implement predictive models. Since our data follows a format such that a binary dependent variable is explained by some continuous and some categorical variables. Therefore, to estimate the probability of customer churn, we implemented the following 3 models: Logistic regression, Probit regression and Random Forest. Pros and cons of each model are discussed and each model is optimized based on variables. The purpose of this optimization is to get rid of multi collinearity and heteroskedasticity by using GVIF score and Hyperparameter tuning. Finally, the best model is selected based on the predictive capability of the models.

The data was split into training and test set. The training data was used to train the model and based on the suggested estimates, the test set was used to determine the accuracy of the model. For this project, a 70 – 30 split was carried out.

Multinomial Logistic Regression

First, we ran a model consisting of all the variables to get a general idea of the estimates. The results recorded were as follows.

Predictors	*Estimates*	*CI*	*p*	TechSupport	-0.05	-0.11 – 0.01	0.087
(Intercept)	-0.20	-0.86 – 0.47	0.556	StreamingTV	0.06	-0.04 – 0.17	0.248
tenure	-0.12	-0.13 – -0.10	**<0.001**	StreamingMovies	0.07	-0.03 – 0.18	0.178
MonthlyCharges	-0.12	-0.43 – 0.19	0.445	PaperlessBilling	0.05	0.03 – 0.07	**<0.001**
gender	0.00	-0.02 – 0.02	0.860	MultipleLines [1]	0.06	-0.00 – 0.11	0.058
SeniorCitizen	0.02	-0.01 – 0.05	0.282	InternetService [1]	0.22	-0.04 – 0.48	0.095
Partner	-0.01	-0.03 – 0.02	0.636	InternetService [2]	0.48	-0.03 – 1.00	0.063
Dependents	-0.02	-0.05 – 0.01	0.161	Contract [1]	0.11	0.07 – 0.14	**<0.001**
PhoneService	0.01	-0.20 – 0.22	0.924	Contract [2]	0.05	0.01 – 0.08	**0.007**
OnlineSecurity	-0.04	-0.10 – 0.02	0.147	PaymentMethod [1]	0.08	0.05 – 0.11	**<0.001**
OnlineBackup	-0.02	-0.08 – 0.03	0.433	PaymentMethod [2]	0.02	-0.02 – 0.05	0.365
DeviceProtection	0.01	-0.05 – 0.07	0.741	PaymentMethod [3]	0.01	-0.03 – 0.04	0.693

Logistic Regression Estimates with all Variables

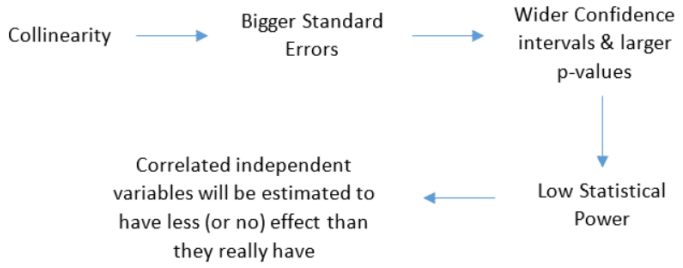
We can observe that the significance scores are not very ideal and the AIC (AIC: 4339) score is also noted. This is to be expected as the model may be suffering from collinearity, heteroskedasticity.

We also ran a naïve model based on Step AIC to get an overview of the significance of estimates. Observed AIC = 4330.57. Even though the score is lower, this approach is not reliable enough as it may even remove important features, and lead to omitted variable bias.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.16	-0.25 – -0.07	**0.001**
tenure	-0.11	-0.13 – -0.10	**<0.001**
MonthlyCharges	-0.10	-0.16 – -0.04	**0.001**
Dependents	-0.02	-0.05 – -0.00	**0.048**
OnlineSecurity	-0.05	-0.08 – -0.02	**0.002**
OnlineBackup	-0.03	-0.05 – 0.00	0.068
TechSupport	-0.05	-0.09 – -0.02	**<0.001**
StreamingTV	0.06	0.02 – 0.09	**0.001**
StreamingMovies	0.07	0.03 – 0.10	**<0.001**
PaperlessBilling	0.05	0.03 – 0.07	**<0.001**
MultipleLines [1]	0.05	0.02 – 0.08	**<0.001**
InternetService [1]	0.21	0.15 – 0.26	**<0.001**
InternetService [2]	0.46	0.35 – 0.56	**<0.001**
Contract [1]	0.11	0.07 – 0.14	**<0.001**
Contract [2]	0.05	0.01 – 0.08	**0.008**
PaymentMethod [1]	0.08	0.05 – 0.11	**<0.001**
PaymentMethod [2]	0.02	-0.02 – 0.05	0.366
PaymentMethod [3]	0.01	-0.03 – 0.04	0.678

Step-wise Logistic Regression Estimates

One of the issues that a binary logistic regression suffers is that the features might suffer from multicollinearity. It will cause unstable estimates and inaccurate variances that affect confidence intervals and hypothesis tests. Since the model at this stage might suffer from Multicollinearity, VIF scores for each feature in each model were estimated.



Understanding Collinearity

We observed a High VIF score for the following features: Monthly Charges, Phone Service, Streaming TV, Streaming Movies, Internet Service. A range of solutions can be applied to the data to fix this issue like increasing the sample size, dropping some variables and combining variables to an index. As a simple as, we chose to drop variables with high collinearity.

	GVIF	Df	GVIF^(1/(2*Df))
tenure	2.822050	1	1.679896
MonthlyCharges	863.046589	1	29.377655
gender	1.003515	1	1.001756
SeniorCitizen	1.152271	1	1.073439
Partner	1.459327	1	1.208026
Dependents	1.371387	1	1.171062
PhoneService	34.384385	1	5.863820
OnlineSecurity	6.259774	1	2.501954
OnlineBackup	6.845387	1	2.616369
DeviceProtection	6.910459	1	2.628775
TechSupport	6.388511	1	2.527550
StreamingTV	23.878008	1	4.886513
StreamingMovies	24.054218	1	4.904510
PaperlessBilling	1.214240	1	1.101926
MultipleLines	7.212157	1	2.685546
InternetService	606.011402	2	4.961582
Contract	2.600809	2	1.269922
PaymentMethod	1.587993	3	1.080127

VIF Scores Pre-Feature Omit

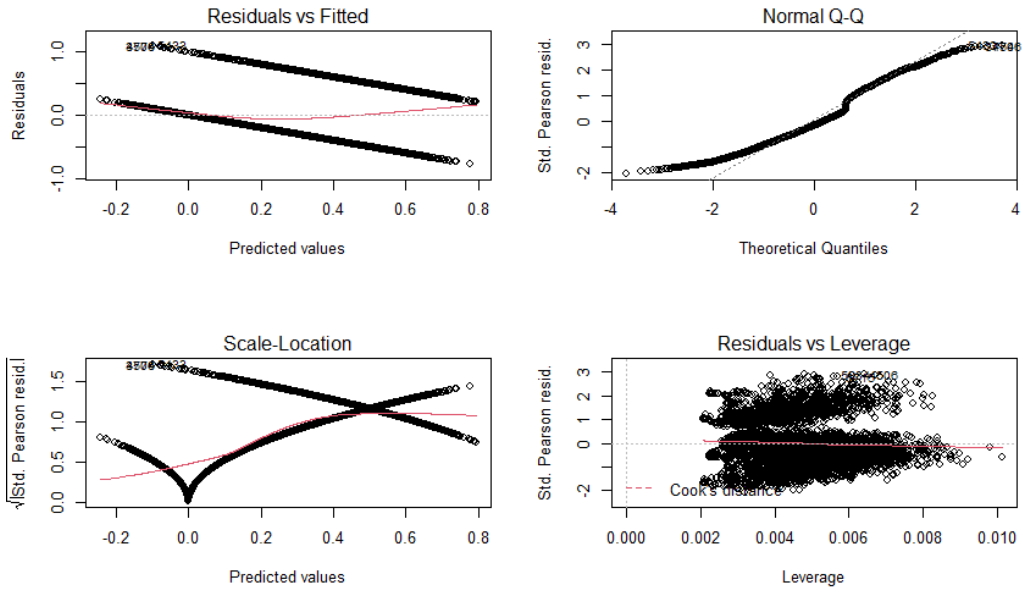
	GVIF	Df	GVIF^(1/(2*Df))
tenure	2.792926	1	1.671205
MonthlyCharges	5.351180	1	2.313262
gender	1.003018	1	1.001508
SeniorCitizen	1.150716	1	1.072714
Partner	1.458610	1	1.207729
Dependents	1.370487	1	1.170678
PhoneService	1.481840	1	1.217308
OnlineSecurity	1.363575	1	1.167722
OnlineBackup	1.450222	1	1.204252
DeviceProtection	1.568880	1	1.252549
TechSupport	1.446633	1	1.202761
StreamingTV	1.955238	1	1.398298
StreamingMovies	1.976248	1	1.405791
PaperlessBilling	1.212266	1	1.101029
MultipleLines	1.537750	1	1.240061
Contract	2.554884	2	1.264279
PaymentMethod	1.582529	3	1.079507

VIF Scores Post-Feature Omit

Dropping one variable at a time to achieve the least AIC and with no VIF score beyond the cutoff threshold, we arrived at the model:

Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.42	0.35 – 0.49	**<0.001**
tenure	-0.11	-0.13 – -0.10	**<0.001**
MonthlyCharges	0.18	0.15 – 0.20	**<0.001**
gender	0.00	-0.02 – 0.02	0.898
SeniorCitizen	0.02	-0.01 – 0.05	0.259
Partner	-0.01	-0.03 – 0.02	0.667
Dependents	-0.02	-0.05 – 0.01	0.150
PhoneService	-0.18	-0.22 – -0.13	**<0.001**
OnlineSecurity	-0.10	-0.13 – -0.07	**<0.001**
OnlineBackup	-0.08	-0.10 – -0.05	**<0.001**
DeviceProtection	-0.04	-0.07 – -0.01	**0.003**
TechSupport	-0.11	-0.13 – -0.08	**<0.001**
StreamingTV	-0.04	-0.07 – -0.01	**0.019**
StreamingMovies	-0.03	-0.06 – 0.00	0.094
PaperlessBilling	0.05	0.03 – 0.07	**<0.001**
MultipleLines [1]	0.01	-0.02 – 0.03	0.620
Contract [1]	0.10	0.07 – 0.14	**<0.001**
Contract [2]	0.05	0.01 – 0.08	**0.005**
PaymentMethod [1]	0.08	0.05 – 0.11	**<0.001**
PaymentMethod [2]	0.02	-0.02 – 0.05	0.385
PaymentMethod [3]	0.01	-0.03 – 0.04	0.735

Proposed Logistic Regression model



Residual Plot for Proposed Logistic Regression

Multinomial Probit Regression

As in the case of Logistic regression, the estimates of the Probit model with the proposed Features are as in the table below. The reported AIC score of this model is 4118.6.

Predictors	*Risk Ratios*	*CI*	*p*				
(Intercept)	0.81	0.60 – 1.10	0.178	TechSupport	0.72	0.64 – 0.80	**<0.001**
tenure	0.62	0.57 – 0.67	**<0.001**	StreamingTV	0.91	0.81 – 1.02	0.114
MonthlyCharges	1.95	1.77 – 2.16	**<0.001**	StreamingMovies	0.96	0.85 – 1.08	0.454
gender	1.01	0.93 – 1.11	0.758	PaperlessBilling	1.23	1.11 – 1.36	**<0.001**
SeniorCitizen	1.04	0.92 – 1.17	0.525	MultipleLines [1]	1.04	0.93 – 1.16	0.512
Partner	0.98	0.88 – 1.09	0.688	Contract [1]	1.44	1.25 – 1.65	**<0.001**
Dependents	0.90	0.80 – 1.02	0.091	Contract [2]	0.75	0.61 – 0.91	**0.004**
PhoneService	0.49	0.41 – 0.59	**<0.001**	PaymentMethod [1]	1.17	1.03 – 1.34	**0.018**
OnlineSecurity	0.73	0.65 – 0.81	**<0.001**	PaymentMethod [2]	0.98	0.84 – 1.14	0.753
OnlineBackup	0.81	0.73 – 0.91	**<0.001**	PaymentMethod [3]	0.95	0.81 – 1.10	0.490

Probit Regression Model

An advantage of the Logit model is that as the Logit model has a closed form solution and the Probit does not, interpretation of estimates can be precisely interpreted for the Logit model.

Interpreting the Results

Taking into account the AIC score, presence of multi-collinearity and the interpretability of the all the above models discussed, the Proposed Logistic Regression yields the best results for interpretation of the data. A few key findings based upon which we would recommend the Company for actions are as follows:

Most of the features like Tenure, Monthly charge, Phone Service, Online Security, Online Backup, Device Protection, Tech Support, Paperless Billing are significant even at a 1% level. Some feature like gender and partner are not significant enough.

Further, since we would prefer no churn ("0 churn rather than 1"), by comparing the estimates, we can conclude that:

- Prolonged tenure has a slightly positive impact on customer retention.
- Gender plays no role in determining the churn rate.
- Providing features like Phone Service, Online Security, Online Backup, Device Protection, Tech Support have a positive impact on churn rate.
- People tend to continue with the service less often if offered paperless billing.
- Customers opt to continue with the service more often if offered an option to pay annually rather than on a monthly basis or for two years.
- Surprisingly, mailed cheque is the more preferred method of payment over e-check.
- Tenure, Phone service, Online Security and Monthly Charges seem to impact the churn rate the most.

Predicting Customer Churn

Using the data for binary classification, we can predict whether a customer will churn or not. In this section we will propose various models and compare their accuracy for predicting customer churn.

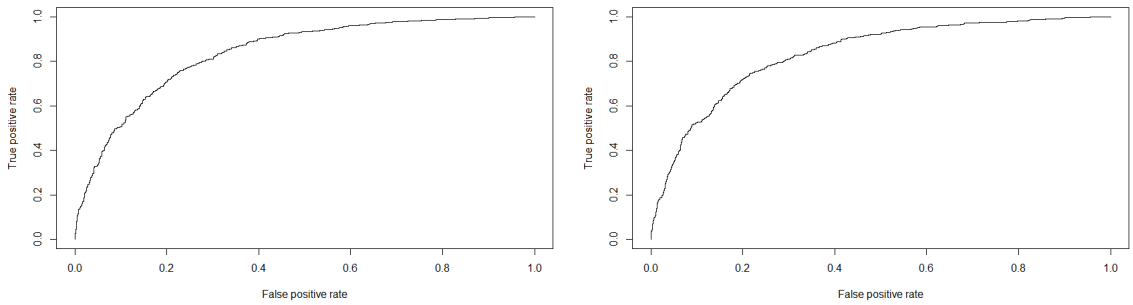
Random Forest Classifier

Random forest is a Supervised Machine Learning Algorithm that can be used for both Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. For the purpose of this project, a Random Forest Model was implemented with the following parameters: proximity = FALSE, importance = FALSE, ntree=500, mtry=4. and the accuracy was observed.

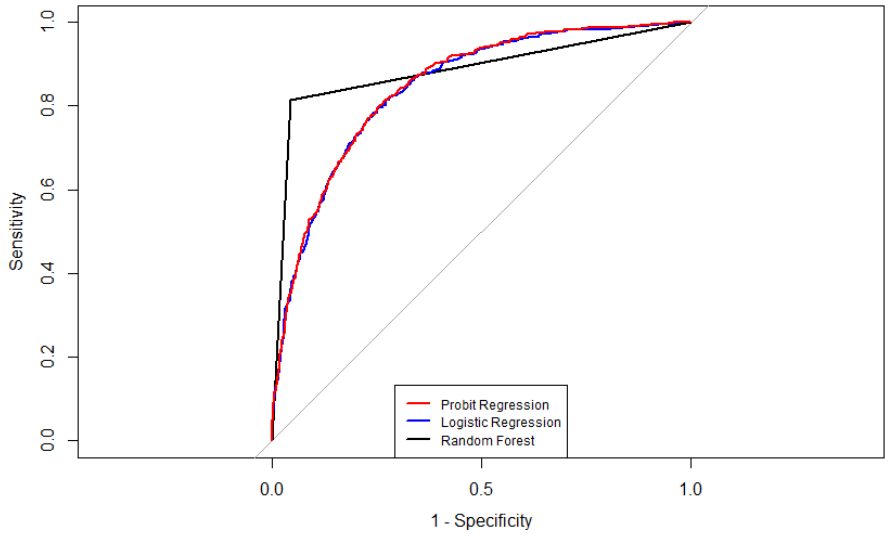
One of the key advantages of Random Forest is that it solves the problem of overfitting as output is based on majority voting/averaging. Also, it is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

Comparing the accuracy score for the models, we get:

S.No.	Model	Accuracy
1	Logistic Regression	0.74
2	Probit Regression	0.78
3	Random Forest	0.79



ROC Curves for a. Logistic b. Probit Regression



ROC Curve Comparison

Since the Random Forest model has the highest accuracy score and the best observed ROC curve, the proposed Random Forest Model would be the best model with an accuracy of 79%, to predict customer churn.

Conclusion

As we examined that data, we saw some notable observations to point out. First, we observe a 26% churn rate for our customers. Additionally, the market as a whole is mainly comprised of people who don't have dependents and who aren't senior citizens. Phone service was the most popular amongst customers, whereas having online security led to lower churn rate. The most relevant attributes for detecting churn were contract, online security, tech support, and internet services while the least significant were gender, partner, and streaming.

After, running our models we see a few more observations. Prolonged tenure has a slightly Positive impact on customer retention. People tend to continue with the service less if offered paperless billing. Customers opt to continue with the service more often if offered an option to pay on an annual basis rather than monthly or for two years. Surprisingly, mailed cheque is the most preferred mode of payment followed by E-cheque. Tenure, phone service, online security and monthly charges seem to impact the churn rate the most.

Recommendations

In order to reduce customer churn and maximize retention we would recommend the options listed further below. Before doing so however, it is very important to point out a huge limitation of this dataset, which is that we're not considering competitor(s) services or actions as reasons for customer churn. This is also not a time series dataset; therefore, we cannot point out or combat certain events that led to significant and concurrent customer churn.

That being said, we would start off our recommendations with creating and offering a tiered loyalty program. Customers could progress into higher tiers by a combination of subscribing to more services or staying with the company longer. The gamification and marketing of this aspect, perhaps introducing a points system and ability to share, could increase both retention and introduction of new customers. The benefits of the program could vary from discounts, premium or 24/7 support, to free or lower cost additional services, exclusive or early access to new or high demand devices or content (streaming), or if possible, partnering with other non-competing companies and offering benefits to mutual customers. Noting the fact that we're focused on retention of existing customers; the company could also offer a variety of rewards for bringing in new customers which would further increase the success and renown of the loyalty program.

Additionally, we could recommend discounts on contractual service rather than month to months and perhaps an additional discount on payment up front for the entire length of the contract. Similarly, we could recommend employee benefits for upselling contracts. Furthermore, we would recommend increasing quality or speed for high-tier services such as fiber optic and streaming and possibly the network as a whole since all services are reliant on it. Similarly, offering the latest or best hardware could achieve the same result. We would also recommend improving the customer service quality and speed, which would minimize the chain effect of one customer leaving and spreading the word and more people doing the same. Finally, as a last-ditch effort, the company could offer something for cancelling customers to coerce them to stay (or return). Refer Code Raw or Code R Mark down for analysis procedure.