Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

Telco Customer Churn: Building a Recommendation System to Prevent Churn

I. Background

This data is part of the IBM telecom dataset. There are several number of telecommunication networks that are available, and we have the luxury to choose the one we want based on our requirements. The increased number of telecoms are a challenge to the telecom companies and many companies are facing huge revenue losses. To retain the customers many companies, invest a huge revenue in the beginning and thus it becomes very important for the customers to expand the business and get back the amount that has been invested in the business [8].

The increasing number of churning customers is the present-day challenge for the telecom industry and such customers create a financial burden for the company [3].

Churn: When one customer leaves one company and moves to another company it is referred to as churning.

II. Objectives

To predict which customers are most likely to churn

Analyze the reasons for customer churn by using different predicting models

To come up with new customized plans for those customers who are most likely to churn

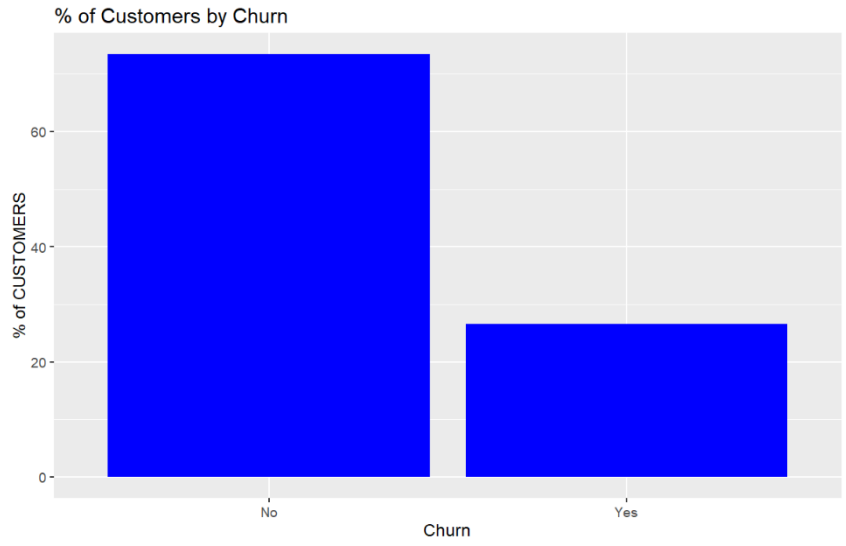
III. Data Exploration

Data Structure

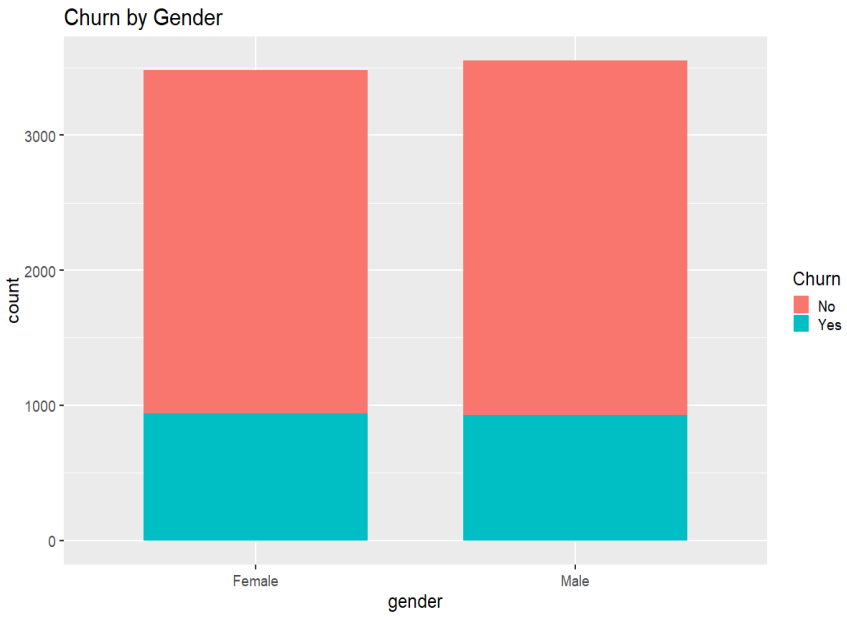
```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

It is a data set with 7043 rows and 21 variables. The column 'tenure', 'MonthlyCharges', and 'TotalCharges' are numerical variables, and the other are supposed to be categorical variables.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018



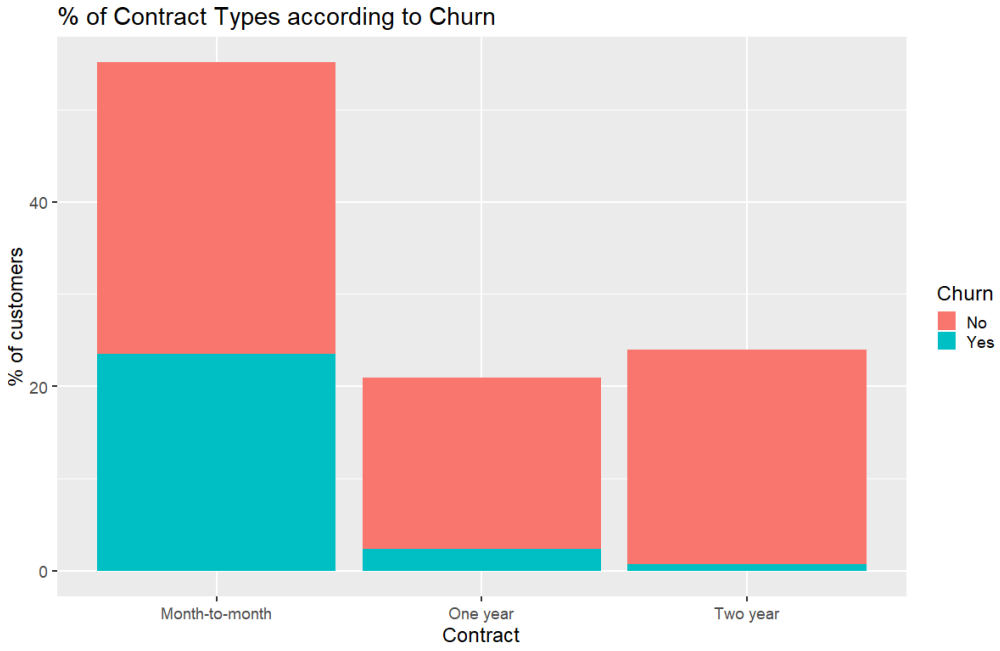
After cleaning, fitting and preprocessing, the dataset consists of 7032 observations of individual customers who have either decided to leave the company's network (churned away) or are still doing business with the company. As shown in the bar plot over 26.5% of the total customer observations, have churned.



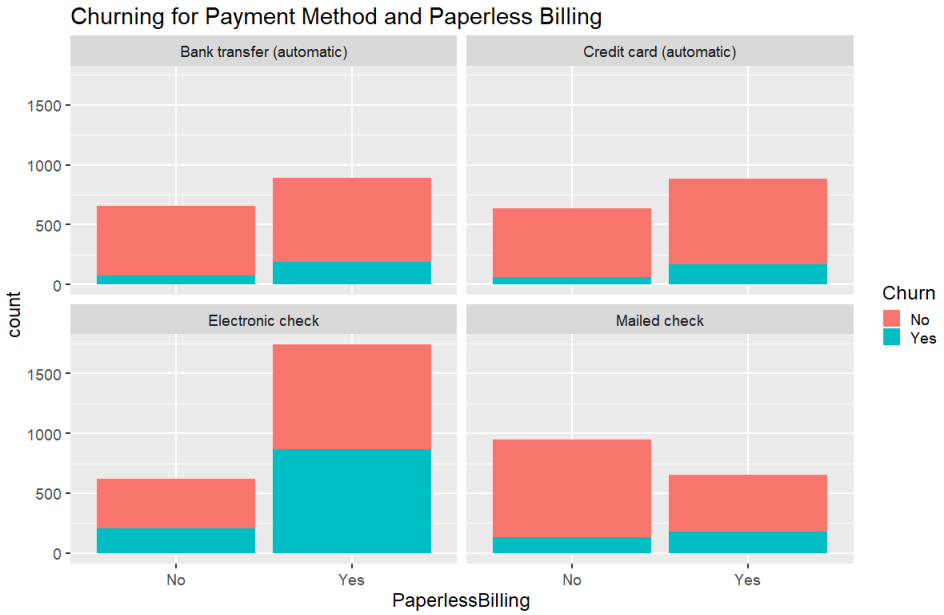
Several speculations can be made about why a customer decides to churn. There are several insights in data visualization. However, an initial bar plot shows that there is good reason to rule out gender as reason of variation in churn.

It should be noted at this time that without further analysis, no certain conclusions- especially for causal relationships can be drawn. However, certain plots that will follow have given insights of potential areas of interest.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

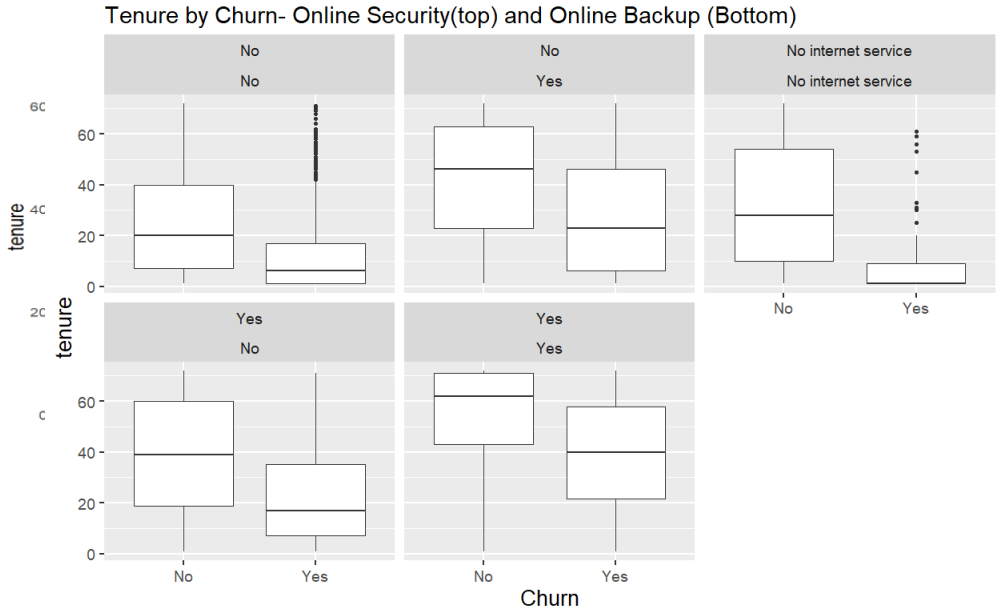


This bar plot shows one of the more intuitive ideas that customers who sign longer contracts find it easy to stay with the company and have the smallest proportion of churn. However, a key takeaway here is that over 50% of the company's client base are still month to month. Ideally, this client base should be moved to a longer contract.

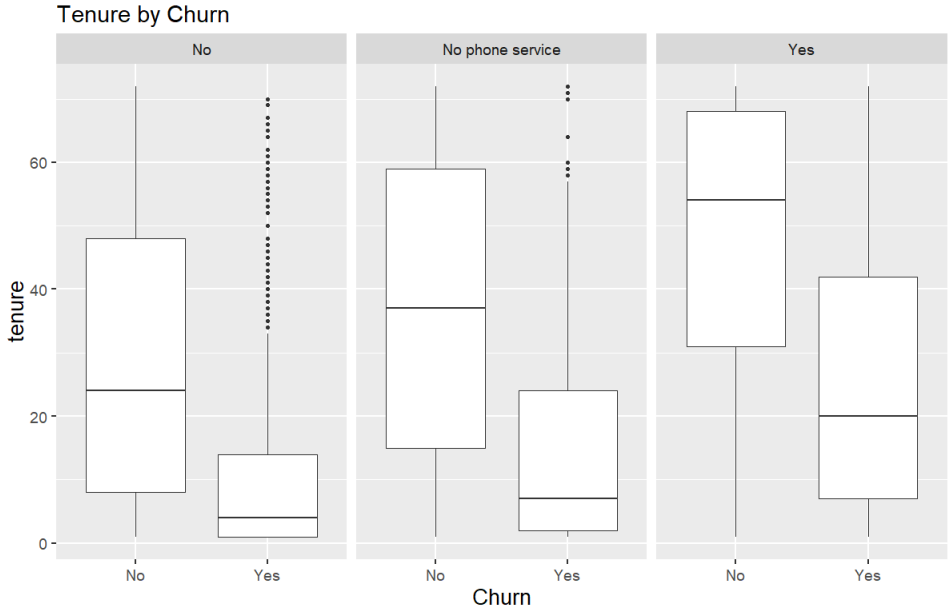


A less intuitive relationship is between churn and the mode of payment. The most prevalent mode of payment in the company's customer base is electronic check with over 1500 observations. However, this method also has the highest proportion of customers who churn away from the network. Similarly, paperless billing (No or Yes on x-axis) is another factor that seems to be associated with the decision to churn. As compared to customers who are still doing business with the company, a greater proportion of customers who decided t churn had been receiving paper bills.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018



While churn is a binary variable for a customer's decision to stay with or leave the company's network, tenure is the amount of time in months that the customer has spent on the company's network. A simple box plot demonstrates that customers who do decide to churn away, spend, on average, over three times fewer months doing business with the company, as compared to customers who do not churn.

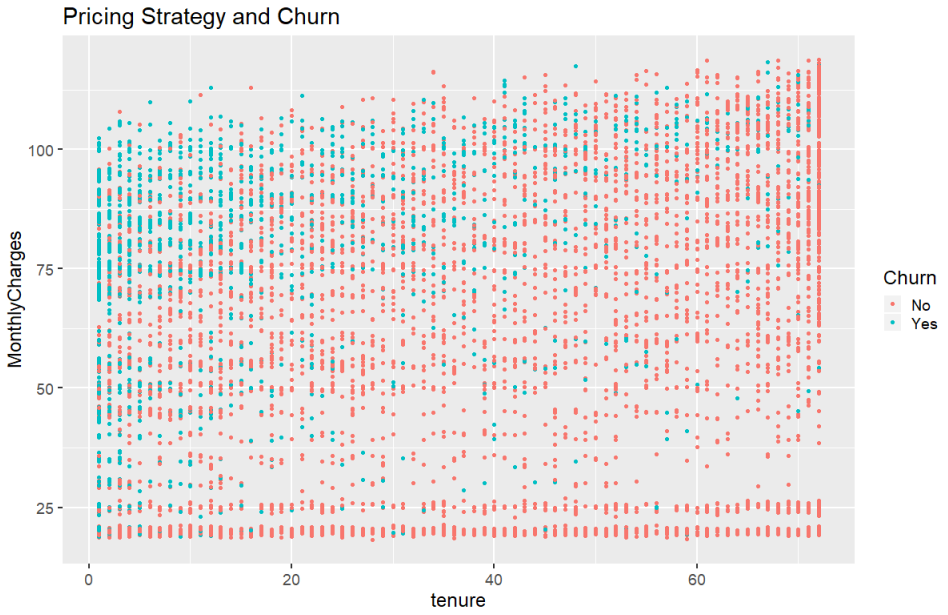


This box plot adds the dimension of having multiple lines to the previous plot. Churning customers having multiple lines spend twice the time doing business with the company as compared to churning not having multiple lines. Similarly, customers who do not churn and have multiple lines spend more than double the time with the company on average, as compared to customers non-churning customers who do not have multiple lines.

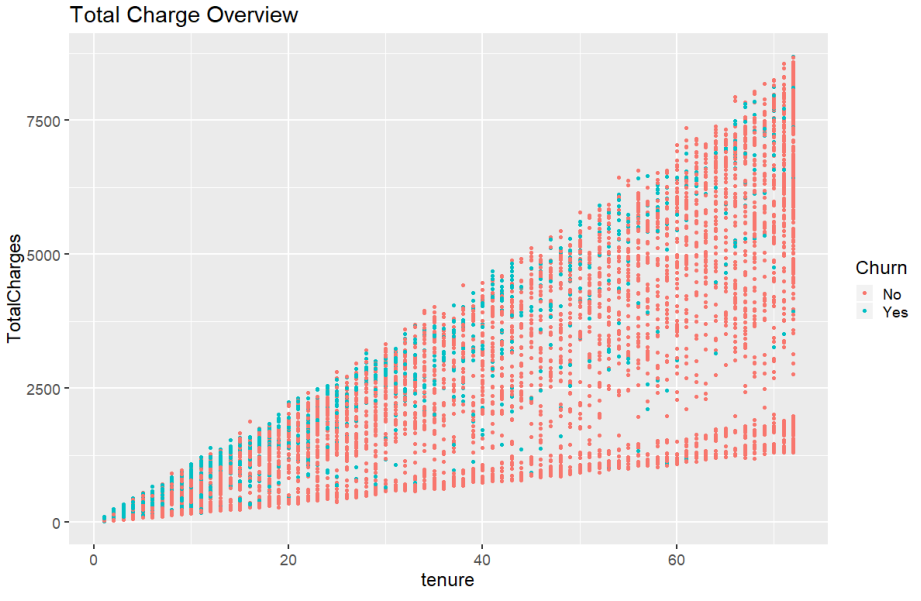
Similar conclusions can be drawn about other factors like whether a customer streams T.V, streams movies, has device protection, has internet service, has online backup, device protection, tech support,

etc. Hence, it can be observed that opting for more services from the company goes hand in hand with increased tenure.

Thus far we have noticed that customers that are more "involved" in using the company's service are on average spending more time doing business with the company. It would make sense for the company to push more services to its customers to extend Tenure and consequently delay and reduce Churn. However, buying more services from the company would also increase the monthly and total charges that a customer pays. Is the problem of Churn sensitive to price and monthly charges?



This figure supports the claim. At almost all levels of tenure, high monthly charges lead to churn. However, this is especially true for low levels of tenure. Once the customer has passed a certain point of tenure, it becomes less likely that he would churn away



Another interesting view of the previous figure is to look at the customer's sensitivity towards cost through total charges. It can be noted that the company loses its most valuable customers at all levels of tenure. However, it can be noted that the customers become less sensitive towards the total charges that they pay, as the time they spend with the company (tenure)

rises. At low levels of tenure, customers are churning the most at high total charges made to their accounts.

Data preprocessing

Removing missing and unnecessary data: Since there are only 11 missing data, we just remove it. We also remove the first column since it is ID.

Extract categorial columns and transform them in to dummy variables.

Combine the numerical and dummy data and separate them in to testing data and validating data.

Except for the dependent variable, we need to scale the data to prevent features domination. Now, we are ready to do the model prediction!

IV. Prediction Models

SVM (support Vector Machine): SVM are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

Logistic Regression with Binary Variable Churn

Linear Regression with Numerical Proxy Tenure

K Nearest Neighbor (KNN)

Linear Discriminant Analysis (LDA)

Validation Techniques

Confusion Matrix

K-Fold Cross Validation

Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (called the validation dataset or testing set). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset. Here we set the K fold to 5.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

Feature Analysis

To analyze the feature of our data, we run the summary of Logistic Regression

```
Coefficients: (7 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.736494   0.065224 -26.624  < 2e-16 ***
tenure                                 -1.341043   0.178872  -7.497 6.52e-14 ***
MonthlyCharges                         -0.758123   1.132143  -0.670 0.503090
TotalCharges                            0.626485   0.187344   3.344 0.000826 ***
gender                                 -0.022842   0.038914  -0.587 0.557211
Partner                                 0.009889   0.046124   0.214 0.830231
Dependents                             -0.065370   0.048618  -1.345 0.178767
PhoneService                           -0.094803   0.226520  -0.419 0.675569
MultipleLines.xNo.phone.service               NA         NA      NA       NA
MultipleLines.xYes                      0.189703   0.103699   1.829 0.067345 .
InternetService.xFiber.optic            0.745441   0.469394   1.588 0.112266
InternetService.xNo                    -0.589107   0.393987  -1.495 0.134851
OnlineSecurity.xNo.internet.service           NA         NA      NA       NA
OnlineSecurity.xYes                    -0.135579   0.096139  -1.410 0.158468
OnlineBackup.xNo.internet.service             NA         NA      NA       NA
OnlineBackup.xYes                      -0.013312   0.098983  -0.134 0.893014
DeviceProtection.xNo.internet.service         NA         NA      NA       NA
DeviceProtection.xYes                   0.035534   0.099275   0.358 0.720391
TechSupport.xNo.internet.service              NA         NA      NA       NA
TechSupport.xYes                       -0.137611   0.096604  -1.424 0.154306
StreamingTV.xNo.internet.service              NA         NA      NA       NA
StreamingTV.xYes                        0.226333   0.187720   1.206 0.227935
StreamingMovies.xNo.internet.service          NA         NA      NA       NA
StreamingMovies.xYes                    0.210963   0.188669   1.118 0.263497
Contract.xOne.year                     -0.294726   0.052786  -5.583 2.36e-08 ***
Contract.xTwo.year                     -0.563043   0.089208  -6.312 2.76e-10 ***
PaperlessBilling                        0.179553   0.044334   4.050 5.12e-05 ***
PaymentMethod.xCredit.card..automatic. -0.125981   0.057044  -2.209 0.027209 *
PaymentMethod.xElectronic.check         0.137011   0.052801   2.595 0.009463 **
PaymentMethod.xMailed.check            -0.023653   0.056837  -0.416 0.677292
---
```

We can see that tenure, Total Charge, Contract Year, Payment Method are significant. Besides, Internet Service also have a high absolute value.

Result

With set.seed(170).

The accuracy rate in confusion matrix of SVM = 0.814

The accuracy rate of k-fold cross validation = 0.802

The accuracy rate in confusion matrix of Logistic Regression = 0.804

The accuracy rate in confusion matrix of Decision Tree = 0.784

The accuracy rate of k-fold cross validation of Decision Tree = 0.797

The accuracy rate in confusion matrix of KNN = 0.76

The accuracy rate of KNN validation of Decision Tree = 0.761

The accuracy rate in confusion matrix of LDA = 1

Summary

1. Features such as tenure group, Contract, Payment Method, Monthly Charges and Internet Service appear to play a role in customer churn.

2. Most models are equally fine. LDA has a very high accuracy. In order to make sure that the model has no problem, we tested LDA again with less variables, and we did get a much lower accuracy.

Linear Regression

Linear regression is a method of mapping the relationship between the dependent variable and the independent variables. This algorithm is often used in prediction or forecasting where a fitted training model can be used on a test dataset to predict the outcome in that situation [4].

In this project, linear regression was used on the four numerical variables. The independent variables in this situation were Senior Citizen, Monthly Charges, and Total Charges, while the dependent variables was tenure. Using the entire dataset as the training set, the model was fitted using Multiple Linear Regression (multiple because we have more than one independent variable). The screenshot below shows the summary of the Linear Regression Model:

```
##
## Call:
## lm(formula = final_df_numeric$final_df.tenure ~ ., data = final_df_numeric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.365  -5.796   0.764   4.413  33.150
##
## Coefficients:
##                              Estimate  Std. Error t value
## (Intercept)                 30.57732848  0.29443628 103.850
## final_df.SeniorCitizen       0.59223265  0.33701800   1.757
## final_df.MonthlyCharges     -0.41378983  0.00541487 -76.417
## final_df.TotalCharges        0.01250870  0.00007048 177.476
##                                 Pr(>|t|)
## (Intercept)                 <0.0000000000000002 ***
## final_df.SeniorCitizen                0.0789 .
## final_df.MonthlyCharges     <0.0000000000000002 ***
## final_df.TotalCharges       <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.15 on 7028 degrees of freedom
## Multiple R-squared:  0.829,  Adjusted R-squared:  0.8289
## F-statistic: 1.136e+04 on 3 and 7028 DF,  p-value: < 0.00000000000000022
```

Here, we see that the most significant predictor variables are Monthly Charges and Total Charges. When Monthly Charges is increased by one unit, tenure decreases by 0.4138 units and when Total Charges is increased by one unit, tenure increases by 0.0125 units.

The output below shows the accuracy of the model.

	ME	MAE
Test set	0.0000000000005716146	7.389777

Here, the term “Test set” implies the data that the linear regression model analyzed [7].

ME (mean error): The mean error is the average of all errors in the dataset. Due to the fact that it does not take into account absolute values, is not a very useful or accurate measure of the goodness of fit of the model. As a result, it is advisable to interpret the MAE instead.

MAE (mean absolute error): The mean absolute error in this situation is 7.38977. This value is not particularly large and thus proves that the model has a fairly good fit.

While the linear regression model is fairly good when working with the numerical variables in the dataset, the fact that it does not deal with factor variables is certainly a shortcoming. Using linear regression also means that another algorithm specifically tailored for categorical (factor) variables must be used in conjunction with it.

We use significant variables in the two regression models as the bases of our regression model.

Association Rules

Association Rules is a supervised machine learning algorithm that helps determine the relationship between variables in a dataset [1]. It seeks to find out "what goes with what" in a shopping cart for example. This algorithm is often used as the basis for recommendation systems which suggest certain items or services based on what the customer has already purchased or is looking at [2]. The Association Rules Algorithm used in this project is the Apriori algorithm which looks for frequently occurring itemset (the set comprising the antecedent and the consequent) using the bottom-up approach. A major caveat with using this approach is that, because the algorithm searches through the dataset multiple times, there is a depreciation in its performance [6]. However, considering the scope of the dataset and the knowledge of the team with regard to machine learning algorithms, this algorithm suffices.

The tables below show the different combinations used to find "what goes with what" using Association Rules.

Association Rules with Churn = No on the R.H.S.

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.PhoneService=No, final_df.Contract=Two year, final_df.PaymentMethod=Credit card (automatic)	final_df.Churn=No	0.0074	1	1.3620	52

2	final_df.PhoneService=No, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)	final_df.Churn=No	0.0080	1	1.3620	56
3	final_df.Dependents=Yes, final_df.PhoneService=No, final_df.Contract=Two year	final_df.Churn=No	0.0102	1	1.3620	72
4	final_df.PhoneService=No, final_df.StreamingMovies=No, final_df.Contract=Two year	final_df.Churn=No	0.0085	1	1.3620	60

5	final_df.Partner=No, final_df.PhoneService=No, final_df.Contract=Two year	final_df.Churn=No	0.0077	1	1.3620	54
6	final_df.MultipleLines=No phone service, final_df.Contract=Two year, final_df.PaymentMethod=Credit card (automatic)	final_df.Churn=No	0.0074	1	1.3620	52

Interpretation

Here, we see what services are availed by and what criteria apply to those people who do not churn. One of the first points is that people who do not churn, that is, those who stay with the service provider, have the facility of automated payment through credit card. Such subscribers also have a two-year contract. These two observations imply that, if the company wants to entice those customers who do churn or those who intend to churn to stay, they can offer longer contracts and recommend the option of paying automatedly through credit card direct deposit.

Association Rules with Two-year Contract on the R.H.S

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.Dependents=Yes, final_df.OnlineSecurity=Yes, final_df.OnlineBackup=Yes,	final_df.Contract=Two year	0.0051	1.0000	4.1733	36

	final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.PaperlessBilling=No, final_df.PaymentMethod=Credit card (automatic)					
2	final_df.Dependents=Yes, final_df.OnlineSecurity=Yes, final_df.OnlineBackup=Yes, final_df.DeviceProtection=Yes, final_df.TechSupport=Yes, final_df.StreamingTV=Yes, final_df.PaymentMethod=Credit card (automatic)	final_df.Contract=Two year	0.0092	0.98485	4.1101	65
3	final_df.Dependents=Yes, final_df.MultipleLines=Yes, final_df.OnlineSecurity=Yes, final_df.DeviceProtection=Yes, final_df.TechSupport=Yes, final_df.StreamingTV=Yes, final_df.PaymentMethod=Credit card (automatic)	final_df.Contract=Two year	0.0077	0.9818	4.0974	54
4	final_df.Dependents=Yes, final_df.MultipleLines=Yes, final_df.OnlineSecurity=Yes, final_df.DeviceProtection=Yes, final_df.TechSupport=Yes,	final_df.Contract=Two year	0.0058	0.9762	4.0739	41

	final_df.StreamingTV=Yes, final_df.PaymentMethod=Bank transfer (automatic)					
5	final_df.Dependents=Yes, final_df.MultipleLines=Yes, final_df.OnlineSecurity=Yes, final_df.DeviceProtection=Yes, final_df.PaperlessBilling=No, final_df.PaymentMethod=Credit card (automatic), final_df.Churn=No	final_df.Contract=Two year	0.0051	0.9730	4.0605	36
6	final_df.Dependents=Yes, final_df.MultipleLines=Yes, final_df.InternetService=DSL, final_df.OnlineSecurity=Yes, final_df.OnlineBackup=Yes, final_df.StreamingMovies=Yes, final_df.StreamingTV=Yes	final_df.Contract=Two year	0.0073	0.9623	4.0158	51

Interpretation: Here, we examine what services and criteria appear on the left-hand side when the Two-year contract appears on the RHS. The reason for examining this set of rules is to recommend those services availed by customers who have the longest contracts with the company to those customers who intend to churn. For example, we notice that customers who have two year contracts have facilities like online backup, online security, tech support, device protection, automated credit card payment, streaming television, and multiple lines. If a customer who intends to churn or who has churned does not have these services, it is advisable to recommend some of the services appearing in this table.

Association Rules with StreamingMovies=Yes on R.H.S.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.MultipleLines=Yes, final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.Contract=One year, final_df.Churn=Yes	final_df.StreamingMovies=Yes	0.0068	1	2.5749	48
2	final_df.Partner=Yes, final_df.Dependents=No, final_df.StreamingTV=Yes, final_df.Contract=One year, final_df.Churn=Yes	final_df.StreamingMovies=Yes	0.0060	1	2.5749	42
3	final_df.OnlineSecurity=No, final_df.OnlineBackup=Yes, final_df.TechSupport=No, final_df.StreamingTV=Yes, final_df.Contract=Two year	final_df.StreamingMovies=Yes	0.0065	1	2.5749	46
4	final_df.MultipleLines=Yes, final_df.InternetService=Fiber optic, final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.Contract=One year, final_df.Churn=Yes	final_df.StreamingMovies=Yes	0.0061	1	2.5749	43

5	final_df.MultipleLines=Yes, final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.Contract=One year, final_df.PaperlessBilling=Yes, final_df.Churn=Yes	final_df.StreamingMovies=Yes	0.0058	1	2.5749	41
6	final_df.Dependents=No, final_df.MultipleLines=Yes, final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.Contract=One year, final_df.Churn=Yes	final_df.StreamingMovies=Yes	0.0058	1	2.5749	41

Interpretation: Here, we examine those relationships wherein streaming movies (yes) appears on the right hand side. Curiously, customers who have the service of streaming service tend to churn. It should also be notice that some of these customers also have a one-year contract with the company. Based on this observation, the team recommends that the company take a survey of this set of customers who have this service and still churn and take action accordingly. It is also possible that the company needs to work towards improving their movie-streaming service.

Association Rules with DeviceProtection=Yes on R.H.S.

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.gender=Male, final_df.InternetService=Fiber optic, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.DeviceProtection=Yes	0.0058	1.0000	2.9082	41

2	final_df.gender=Male, final_df.InternetService=Fiber optic, final_df.StreamingMovies=Yes, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.DeviceProtection=Yes	0.0055	1.0000	2.9082	39
3	final_df.gender=Male, final_df.MultipleLines=Yes, final_df.InternetService=Fiber optic, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.DeviceProtection=Yes	0.0055	1.0000	2.9082	39
4	final_df.gender=Male, final_df.InternetService=Fiber optic, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic),	final_df.DeviceProtection=Yes	0.0051	1.0000	2.9082	36

	final_df.Churn=No					
5	final_df.gender=Male, final_df.PhoneService=Yes, final_df.InternetService=Fiber optic, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.DeviceProtection=Yes	0.0058	1.0000	2.9082	41
6	final_df.gender=Male, final_df.Partner=Yes, final_df.MultipleLines=Yes, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.DeviceProtection=Yes	0.0064	0.9783	2.8450	45

Interpretation: Here, we examine the occurrence of services when device protection (yes) appears on the right hand-side. Some of the services that go with this service are automated payment through credit card, streaming television, paperless billing, and fiber optic. Based on this information, in the eventuality that a customer who wants to churn has the facility of device protection, the organization can recommend some of the services on the left-hand side in an effort to entice the customer to not churn.

Association Rules with OnlineSecurity=Yes on R.H.S.

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.gender=Female, final_df.Dependents=Yes, final_df.MultipleLines=Yes,	final_df.OnlineSecurity=Yes	0.0061	0.9773	3.4105	43

	final_df.DeviceProtection=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=No					
2	final_df.Dependents=Yes, final_df.MultipleLines=Yes, final_df.InternetService=DSL, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Credit card (automatic)	final_df.OnlineSecurity=Yes	0.0058	0.9762	3.4067	41
3	final_df.gender=Female, final_df.Dependents=Yes, final_df.PhoneService=Yes, final_df.InternetService=DSL, final_df.Contract=Two year, final_df.PaymentMethod=Credit card (automatic)	final_df.OnlineSecurity=Yes	0.0057	0.9756	3.4047	40
4	final_df.Dependents=Yes, final_df.PhoneService=Yes, final_df.InternetService=DSL, final_df.StreamingMovies=No, final_df.Contract=Two year, final_df.PaperlessBilling=No	final_df.OnlineSecurity=Yes	0.0057	0.9756	3.4047	40
5	final_df.MultipleLines=Yes,	final_df.OnlineSecurity=Yes	0.0053	0.9737	3.3980	37

	final_df.DeviceProtection=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Mailed check					
6	final_df.MultipleLines=Yes, final_df.DeviceProtection=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Mailed check, final_df.Churn=No	final_df.OnlineSecurity=Yes	0.0053	0.9737	3.3980	37

Interpretation: Here, we examine the services which appear on the left-hand side when online security (yes) appears on the right hand side. The aim of examining this set of rules is so that the organization can recommend some of these services (that appear on the left hand) to those customers who want to churn and have online security. It is observed that customers who have online security also have device protection, phone service, DSL internet service, and streaming TV. In order to entice customers to stay with the company, the organization could provide incentives like discounts on some of these services.

Association Rules OnlineBackup=Yes on R.H.S.

	LHS	RHS	Support	Confidence	Lift	Count
1	final_df.gender=Male, final_df.Partner=Yes, final_df.Dependents=No, final_df.PhoneService=Yes, final_df.DeviceProtection=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)	final_df.OnlineBackup=Yes	0.0057	0.9524	2.7617	40

2	final_df.gender=Male, final_df.Partner=Yes, final_df.Dependents=No, final_df.PhoneService=Yes, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)	final_df.OnlineBackup=Yes	0.0051	0.9474	2.7472	36
3	final_df.Partner=Yes, final_df.Dependents=No, final_df.MultipleLines=Yes, final_df.OnlineSecurity=Yes, final_df.DeviceProtection=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)	final_df.OnlineBackup=Yes	0.0058	0.9318	2.7021	41
4	final_df.Partner=Yes, final_df.Dependents=No, final_df.PhoneService=Yes, final_df.OnlineSecurity=Yes, final_df.Contract=Two year, final_df.PaperlessBilling=Yes, final_df.PaymentMethod=Bank transfer (automatic)	final_df.OnlineBackup=Yes	0.0058	0.9318	2.7021	41
5	final_df.Dependents=No, final_df.MultipleLines=Yes,	final_df.OnlineBackup=Yes	0.0058	0.9318	2.7021	41

	final_df.OnlineSecurity=Yes, final_df.DeviceProtection=Yes, final_df.StreamingTV=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)					
6	final_df.Partner=Yes, final_df.Dependents=No, final_df.MultipleLines=Yes, final_df.OnlineSecurity=Yes, final_df.TechSupport=Yes, final_df.Contract=Two year, final_df.PaymentMethod=Bank transfer (automatic)	final_df.OnlineBackup=Yes	0.0055	0.9286	2.6927	39

Interpretation: Here, the services that appear on the left hand side when online backup (yes) appears on the right hand side are examined. This is also done with the purpose of recommending some of these services (possibly at discounted values) to those customers who churn and who have online backup. It is observed that such customers also have online security, device protection, streaming television and a phone service.

V. Conclusion/Recommendation

It is a common claim, often rattled by marketing experts that it is five times costlier to get a new customer as compared to retaining an existing customer. In this dataset, we can see through LDA and data visualizations that the company is losing its most valuable customers through churn. Amongst many differences that segregate churning and non-churning customers, monthly and total charges are an important one. Hence, clients with largest propensity to consume and generate revenue for the company end up churning. The regression models discussed in this paper highlight important factors for churn and the association rules help provide recommendations to prevent churn. However how far do we want to go with these recommendations and incentives.

The dilemma is that once these customers are given incentives and discounts, it will cost the company some money. Moreover, the monthly charge of customers who churn is often higher than customers who remain with the firm. Is it wise to spend funds and effort to encourage customers to stay with the

company or is it better to milk profits from high value churning customers and let them leave when they finally do?

It is beyond the scope of this study to estimate cost of preventing churn. However, we can get an idea of the benefit of keeping a customer. The average life time value (average total revenue earned or average total charge) of a customer who does not churn is over 2555 USD. The value for customers who do churn is 1531.8 USD. This means that when the company convinces a customer who is about to churn, to not leave, it earns an average of over 1023.5 USD of additional revenue. Hence, although in the short run it may seem more profitable to milk out high value churning customers and letting them leave, in the long run it is not advisable. Moreover, as long as the cost of incentivizing a client is less than 1023.5 USD, efforts, like those suggested by the association rules, should be made to prevent churn.

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

Works Cited

1. "Association rule learning," *Wikipedia.org,* Nov. 15, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Association_rule_learning, [Accessed: Nov. 20, 2018].
2. "Apriori algorithm," Wikipedia.org, Oct. 10, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Apriori_algorithm, [Accessed: Nov. 20, 2018].
3. Gursoy, Ummain Tugba Simsek. "Customer Churn Analysis in Telecommunication Sector," *Istanbul University Journal of the School of Business Administration*, 2010 [Online]. Available: http://dergipark.gov.tr/download/article-file/98167, [Accessed Nov. 20, 2018].
4. "Linear regression," *Wikipedia.org,* Nov. 21, 2018. [Online]. Available: https://en.wikipedia.org/wiki/Linear_regression, [Accessed Nov. 23, 2018].
5. Rahman, Faraz. "Telco Customer Churn Logistic Regression," Kaggle. Available: https://www.kaggle.com/farazrahman/telco-customer-churn-logisticregression/notebook, [Accessed Oct. 15, 2018].
6. S. Chatterjee. BUAN 6356. Class Lecture, Topic: "Association Rules and Collaborative Filtering." Jindal School of Management, The University of Texas at Dallas, Richardson, TX, Nov. 7, 2018.
7. S. Chatterjee. BUAN 6356. Class Lecture, Topic: "Linear Regression." Jindal School of Management, The University of Texas at Dallas, Richardson, TX, Sept. 19, 2018.
8. Sebastian, Helen Treasa and Wagh, Rupali. "Churn Analysis in Telecommunication Using Logistic Regression," *Oriental Journal of Computer Science and Technology*, Mar. 24, 2017 [Online]. Available: http://www.computerscijournal.org/vol10no1/churn-analysis-in-telecommunication-using-logistic-regression/, [Accessed Nov. 20, 2018].

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

A. Appendix A: Code

```
#Data Visualization Code

#Data Preprocessing

library(ggplot2)

df=read.csv("Project_Data_Churn.csv")
head(df)

final_df <- na.omit(df)
final_df[,1] <- list(NULL)

final_df$SeniorCitizen <- as.factor(final_df$SeniorCitizen)
M.plot<- ggplot(final_df)


#This bar plot shows the % of Customers by Churn
bar_churn<- M.plot+ geom_bar(aes(x=Churn, y=((..count..)/sum(..count..)*100) ), fill= "blue")
bar_churn<-bar_churn+ggtitle("% of Customers by Churn") + ylab("% of CUSTOMERS")


#This bar plot shows the churn with respect to gender
bar_gender<- M.plot + geom_bar(aes(x=gender, fill = Churn ) , stat="count", width=0.7) + ggtit
le("Churn by Gender")
bar_gender



#This bar plot shows the Churn for senior citzens according to whether they have a phone service
or not

bar_phone_senior<- M.plot+ geom_bar(aes(x=SeniorCitizen, fill = PhoneService), stat="count"
)
bar_phone_senior



#This bar plot shows the Churn for senior citzens according to whether they have internet service
or not

bar_Int.Sen <- M.plot + geom_bar(aes(x=SeniorCitizen,y = ((..count..)/sum(..count..)*100), fill
= InternetService))
bar_Int.Sen

facet_CC<- bar_Int.Sen + facet_wrap(Contract~Churn)
facet_CC
```

```
theme_set(theme_grey(base_size = 18))



#These Box plots specify the tenure by churn
tenurebar<- ggplot(final_df,aes(Churn,tenure)) + geom_boxplot()+ theme(axis.title.y = element
_text(size = 20))+ theme(axis.title.x =  element_text(size = 20)) +ggtitle("Tenure by Churn")

Mline.Tenure<-tenurebar +facet_wrap(.~MultipleLines)

Int.Tenure<-tenurebar +facet_wrap(.~InternetService)

OnlSec.Tenure<-tenurebar +facet_wrap(.~OnlineSecurity)

OnBck.Tenure<-tenurebar +facet_wrap(OnlineSecurity~OnlineBackup)+ggtitle("Tenure by Ch
urn- Online Security(top) and Online Backup (Bottom)")

DevProt.Tenure<-tenurebar +facet_wrap(.~DeviceProtection)

TechSup.Tenure<-tenurebar +facet_wrap(.~TechSupport)

MovieorTV.Tenure<-tenurebar +facet_wrap(StreamingTV~StreamingMovies)+ggtitle("Tenure
Vs Streaming TV or Movies")

Contr.churn<- ggplot(final_df)+ geom_bar(aes(x=Contract, y=((..count..)/sum(..count..)*100), f
ill=Churn)) + ggtitle("% of Contract Types according to Churn") +ylab("% of customers")



Paperless.churn<-ggplot(final_df)+ geom_bar(aes(PaperlessBilling, fill=Churn))
Paperless.churn

Payment.churn<-ggplot(final_df)+ geom_bar(aes(PaperlessBilling, fill=Churn)) +facet_wrap(.
~PaymentMethod) + ggtitle("Churning for Payment Method and Paperless Billing")
Payment.churn

Payment.churn2<-ggplot(final_df)+ geom_bar(aes(x=PaymentMethod,y = (..count..)/sum(..cou
nt..), fill=Churn))
Payment.churn2

Charges<- ggplot(final_df,aes(MonthlyCharges, TotalCharges)) + geom_point()
Charges

Charge.Tenure<-ggplot(final_df, aes(tenure, MonthlyCharges, colour=Churn))+geom_point()
Charge.Tenure + ggtitle("Pricing Strategy and Churn")

Charge.Tenure2<-ggplot(final_df, aes(tenure, TotalCharges, colour=Churn))+geom_point()
Charge.Tenure2+ ggtitle("Total Charge Overview")
```

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

```
#Association Rules and Linear Regression Code

library(arules)

library(caret)

package: ggplot2

library(forecast)
df <- read.csv("Project_Data_Churn.csv")
final_df <- na.omit(df)
final_df[,1] <- list(NULL)
names(final_df)

str(final_df)

#Association Rules
final_df_factor <- data.frame(final_df$gender,final_df$Partner,final_df$Dependents,final_df$P
honeService, final_df$MultipleLines, final_df$InternetService, final_df$OnlineSecurity, final_df
$OnlineBackup, final_df$DeviceProtection,final_df$TechSupport, final_df$StreamingMovies, fi
nal_df$StreamingTV, final_df$Contract, final_df$PaperlessBilling, final_df$PaymentMethod,fin
al_df$Churn)

str(final_df_factor)

#Rule Set 1: Churn= No as consequent
rules <- apriori(final_df_factor, parameter = list(minlen=1, supp=0.005, conf=0.8),
                 appearance=list(rhs=c("final_df.Churn=No"), default="lhs"),
                 control = list(verbose=F))

rules_sorted <- sort(rules, by="lift")
rules_one <- inspect(head(rules_sorted))

#Rule Set 2: Contract=Two year as consequent
rules_Contract <- apriori(final_df_factor,parameter = list(minlen=2, supp=0.005, conf=0.8),
                          appearance = list(rhs=c("final_df.Contract=Two year"),default="lhs"),
                          control = list(verbose=F))

rules_Contract_sorted <- sort(rules_Contract, by="lift")
inspect(head(rules_Contract_sorted))

#Rule Set 3: Streaming Movies=Yes as consequent
rules_churn_streamingMovies <- apriori(final_df_factor, parameter=list(minlen=2, supp=0.005,
```

```
conf=0.8),
                              appearance=list(rhs=c("final_df.StreamingMovies=Yes"),default="lhs"),
                              control=list(verbose=F))

sorted_rules_streamingMovies <- sort(rules_churn_streamingMovies, by="lift")
inspect(head(sorted_rules_streamingMovies))

#Rule Set 4: Device Protection = Yes as consequent
rules_churn_device_protection <- apriori(final_df_factor, parameter=list(minlen=2, supp=0.005
, conf=0.8),
                               appearance=list(rhs=c("final_df.DeviceProtection=Yes"),default="lhs"
),
                               control=list(verbose=F))

sorted_rules_device_protection <- sort(rules_churn_device_protection, by="lift")
inspect(head(sorted_rules_device_protection))

#Rule Set 5: Online Security as consequent
rules_churn_onlinsecurity <- apriori(final_df_factor, parameter=list(minlen=2, supp=0.005, con
f=0.8),
                              appearance=list(rhs=c("final_df.OnlineSecurity=Yes"),default="lhs"),
                              control=list(verbose=F))

sorted_rules_onlinesecurity <- sort(rules_churn_onlinsecurity, by="lift")
inspect(head(sorted_rules_onlinesecurity))

#Rule Set 6: Online Backup as consequent
rules_churn_onlinebackup <- apriori(final_df_factor, parameter=list(minlen=2, supp=0.005, con
f=0.8),
                             appearance=list(rhs=c("final_df.OnlineBackup=Yes"),default="lhs"),
                             control=list(verbose=F))

sorted_rules_onlinebackup <- sort(rules_churn_onlinebackup, by="lift")
inspect(head(sorted_rules_onlinebackup))



#Linear Regression with Numerical Data
#Tenure is the dependent variable
#Since there are multiple predictor variables, we perform Multiple Linear Regression and analyz
e accuracy using MAE and MAPE
final_df_numeric <- data.frame(final_df$SeniorCitizen, final_df$MonthlyCharges, final_df$Tot
alCharges, final_df$tenure)
linear_regression <- lm(final_df_numeric$final_df.tenure~., data=final_df_numeric)
options(scipen=999)
summary(linear_regression)
```

```
prediction <- predict(linear_regression, final_df_numeric)
accuracy(prediction, final_df_numeric$final_df.tenure)
```

#Models: SVM, Logistic Regression, Decision Tree, KNN, LDA with SVM

```
#Import libraries
library(tidyverse)

library(MASS)

library(car)

library(e1071)
library(caret)

library(cowplot)

library(caTools)
library(pROC)

library(ggcorrplot)
library(class)
library(randomForest)

#read the data as "telco"
telco <- read.csv("Project_Data_Churn.csv")

#count the na in the data of each column
na_count <-sapply(telco, function(y) sum(length(which(is.na(y)))))
##There are only 11 na data
#remove na data
telco <- na.omit(telco)
####################################################
#see the structure of the data
str(telco) ##We realize that columns 6,19,20 are numerical data;column 1 is just ID. Column 3 s
upposed to be factor. and the other are Factors

# turn all the Factors into dummy varialbes
telco_catgorial <- telco[,-c(1,6,19,20)]
```

```
dummy<- data.frame(sapply(telco_catgorial,function(x) data.frame(model.matrix(~x-1,data
=telco_catgorial))[,-1]))
str(dummy) ##We now have 27 variables in dummy dataset

telco_inter <- telco[,c(6,19,20)]
telco_combine <- cbind(telco_inter,dummy)

str(telco_combine) # we have 30 variables in our combined data, including the churn in column
30.

#Split the data in to training and validation data
split <- sample.split(telco_combine$Churn, SplitRatio = 0.8)
##############################################################################
set.seed(13)
training_set <- subset(telco_combine, split == TRUE)
test_set <- subset(telco_combine, split == FALSE)

# Feature Scaling
training_set[-30] <- scale(training_set[-30])
test_set[-30] <- scale(test_set[-30])

#SVM - is used to train a support vector machine. It can be used to carry out general regression
and classification (of nu and epsilon-type), as well as density-estimation. A formula interface is p
rovided.



#Settng the equation for SVM
classifier = svm(formula = Churn ~ .,
                 data = training_set,
                 type = 'C-classification',
                 kernel = 'linear')
y_pred = predict(classifier, newdata = test_set[-30])

cm = table(test_set[, 30], y_pred)
cm  #79.5%

 (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])

#Applying K-fold cross validation for svm
folds = createFolds(training_set$Churn, k = 10)
cv = lapply(folds, function(x) {
  training_fold = training_set[-x, ]
  test_fold = training_set[x, ]
  classifier = svm(formula = Churn ~ .,
                   data = training_fold,
                   type = 'C-classification',
```

```
                kernel = 'radial')
  y_pred = predict(classifier, newdata = test_fold[-30])
  cm = table(test_fold[, 30], y_pred)
  accuracy = (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])
  return(accuracy)
})
accuracy = mean(as.numeric(cv)) #accuracy 0.798
accuracy

#Logistic regression

#Setting the equation for Logistic Regression
classifier <- glm(Churn ~ ., data = training_set, family = "binomial")
summary(classifier)

#Features such as tenure_group, Contract, PaperlessBilling, MonthlyCharges and InternetServi
ce appear to play a role in customer churn.
prob_pred <- predict(classifier, type= 'response',newdata= test_set[-30])

y_pred = ifelse(prob_pred > 0.5  ,1,0)
#Making the confusion matrix
cm = table(test_set[, 30], y_pred >0.5 )
cm  #80.7%

#Decision tree

require(rpart)

## Loading required package: rpart

#Setting Churn to a factor variable
training_set$Churn <- factor(training_set$Churn)
test_set$Churn <-factor(test_set$Churn)
classifier = rpart(formula= Churn~.,data = training_set)
y_pred = predict(classifier,newdata = test_set[-30],type = 'class')
#create confusion matrix for decision tree
cm = table(test_set[, 30], y_pred )
cm #78.9%accuracy

 (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])

#Applying K-fold cross validation for decision tree
folds = createFolds(training_set$Churn, k = 10)
cv = lapply(folds, function(x) {
  training_fold = training_set[-x, ]
  test_fold = training_set[x, ]
  classifier = rpart(formula = Churn ~ .,data = training_fold)
```

```
  y_pred = predict(classifier,newdata = test_fold[-30],type = 'class')
  cm = table(test_fold[, 30], y_pred)
  accuracy = (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])
  return(accuracy)
})
accuracy = mean(as.numeric(cv)) #0.7889
accuracy

# Applying Grid Search to find the best parameters for Decision Tree

#KNN

y_pred = knn(train = training_set[, -30],
             test = test_set[, -30],
             cl = training_set[, 30],
             k = 5,
             prob = TRUE)

# Making the Confusion Matrix
cm = table(test_set[, 30],y_pred)
(cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])#0.759

#Applying K-fold cross validation
folds = createFolds(training_set$Churn, k = 10)
cv = lapply(folds, function(x) {
  training_fold = training_set[-x, ]
  test_fold = training_set[x, ]
  y_pred = knn(train = training_fold[, -30],
               test = test_fold[, -30],
               cl = training_fold[, 30],
               k = 5,
               prob = TRUE)
  cm = table(test_fold[, 30], y_pred)
  accuracy = (cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])
  return(accuracy)
})
accuracy = mean(as.numeric(cv)) #0.76
accuracy



#LDA with svm

#Setting the equation for Linear Discriminant Analysis
lda = lda(formula = Churn ~ ., data = training_set)
```

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

```
## Warning in lda.default(x, grouping, ...): variables are collinear

training_set = as.data.frame(predict(lda, training_set))
training_set = training_set[c(4,1)]
test_set = as.data.frame(predict(lda, test_set))
test_set = test_set[c(4, 1)]
# Fitting SVM to the Training set
classifier = svm(formula = class ~ .,
                 data = training_set,
                 type = 'C-classification',
                 kernel = 'linear')

# Predicting the Test set results
y_pred = predict(classifier, newdata = test_set[-2])

# Making the Confusion Matrix
cm = table(test_set[, 2], y_pred)
(cm[1,1] + cm[2,2]) / (cm[1,1] + cm[2,2] + cm[1,2] + cm[2,1])
```

Team Two
BUAN 6356.002
Dr. Sourav Chatterjee
November 30 2018

Teenaz Ralhan Individual Report

Working on this project has been a positive experience for me. Through this project, I learned about different machine learning algorithms of both types: supervised and unsupervised. The project has helped me hone my programming skills, thereby, making a confident programmer in R as well. The dataset that my team chose was the telecom churn dataset. Having worked for a short period of time in the telecommunication industry, I came in with a certain amount of domain knowledge. This project has helped me understand the impact of telecom churn on the industry and to some extent, estimate the cost of retaining a customer through incentives like discounts versus the cost of letting one customer go and attempting to entice another individual to become a customer. A very important component of working on a team project is the attitude that each member brings with them. The other four members in my team brought with them a very positive attitude towards the project and whatever challenges it could bring with it and a willingness to explore new methods of performing analysis. Everyone in the group seemed to have had a good rapport with one another and there was always a clear division of tasks with room for flexibility. This meant that we were free to help each other out with challenges we faced in tasks assigned to us and vice-versa. Since there was a clear division of labor, it was pretty easy to determine the flow of tasks for the project. We met at least once every week for about two-three hours to discuss what tasks needed to be completed next and brainstorm for each one of them. A facet of the project that seemed to be a challenge and a success was determining a convenient meeting time. There were some week wherein an in-person face-to-face meeting was not a viable solution. Our team turned this challenge into a success by scheduling online meetings through Skype or detailed email exchanges regarding the challenges from the previous week's set of tasks and the assignment of tasks for the following week. Another challenge that we faced but turned into a success was deciding which methods to use for data analysis and prediction. We tried different methods and decided upon the few methods outlined and detailed in the report above. My contribution to this group project was to take care of housekeeping tasks like organizing group meetings and documenting the discussion in each meeting, aiding with data preprocessing, aiding with data interpretation, programming the association rules and linear regression, and developing the report. Adithya contributed to the project by assisting with Association Rules and KNN, consolidating the code, putting together the tables and diagrams,

and developing the report. Ali contributed to the project through his domain knowledge, by aiding in data preprocessing, data visualization, coding decision trees, and aiding in developing the report. Joseph contributed to this project by aiding in data preprocessing and visualization, generating different models based on algorithms like SVM, Logistic Regression and Linear Discriminant Analysis, and aiding in developing the report. Finally, Vardhan contributed to the project by aiding with data preprocessing, helping with the logistic regression algorithm, aiding in code consolidation, and developing the report. Overall, everyone contributed to the project equally and took responsibility in delivering the final product.