The University of Texas at Dallas

Business Analytics with R
Project Report

Predicting Telco Customer Churn

Focused customer retention programs

Authors:
Amritha Rekha
Diana Francisco Almeida
Kinjal Chetan Sheth
Savrin Darash
Shyam Vishnu
(Group 3)

Professor:
Ling Ge

Spring 2019

Contents

1. Executive summary

A telecommunications company is concerned about the number of customers leaving their business for the competitors. They need to gain an insight into who is likely to leave by analyzing their customer data.

The data set provides info in helping predict and using that information retain customers. By analyzing all relevant customer data and developing a propensity model, focused on customer retention programs you can help save the company from losing more customers and profit.

To predict the customers that are going to churn we are using a large dataset which consist of numeric and categorical variable. Models are created on the dependent variable named 'Churn' which is a binary variable with 0/1 representing as non-churn & churn respectively which makes this a classification problem. This project is executed with two ways, with two variants of datasets. Using this technique, we tried to achieve increased accuracy in predicting the best model.

We started with data cleaning followed by data exploration to find the correlation between the predictor variables. The next step was data splitting into training and testing data sets. We started building models using elementary techniques like Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and then built models using complex techniques such as K-Nearest Neighbors (KNN), Decision Trees, Random Forest and Naïve-Bayes. We devised a cost-assumption plan of $1 per customer for advertising. Our goal was to select the model where we need to spend the least amount to retain a customer so choosing the final model would be a trade-off decision between accuracy and cost efficiency.

2. Project motivation/background

Customer churn occurs when customers or subscribers discontinuing with a company or service, also known as customer attrition. It is also referred as loss of clients or customers. Churn is one of the largest problems faced by most businesses. According to Harvard Business Review, it costs between 5 times and 25 times as much to find a new customer than to retain an existing one. Preventing customer churn is an important business function. All industries suffer churn, but telecommunication industry can be considered to be top, among the industries which suffer from this issue. Mobile Service Providers have implemented CRM (Customer Relationship Management) with intention to reduce the number of Customer Churn. However, still the Telecom Industry is facing high churn rate. Competitors in the telecommunication market is constantly tempting customers with more incentives to make churn over the service providers. That brings churn management, the process of retaining customers, a major challenge for the carriers to turn unreliable subscribers into customers. It is easier to retain the existing customer than to acquire new customer. We began this study by looking at the market size in the telecom industry and its customer churn impacts.

2.1 Research Question

By realizing how important the customer churn is in a telecom industry, this has motivated us to build a propensity model which will help us in predicting customer churn to reduce investment in advertising and thus, to retain the existing customer.

3. Data description

The dataset comprises of customer data which helps us predict customer churn. Each tuple in the dataset represents a customer and each column contains customer's information. The data set consists of 7043 observations and 21 variables (customer features). The dataset includes information such as:

- Customer Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Following are the variables in our dataset:

SL. No	Variable Name	Data Type	Variable Description
1	customerID	String	Customer ID
2	gender	String	Whether the customer is a male or a female
3	SeniorCitizen	Numeric	Whether the customer is a senior citizen or not (1, 0)
4	Partner	String	Whether the customer has a partner or not (Yes, No)
5	Dependents	String	Whether the customer has dependents or not (Yes, No)
6	tenure	Numeric	Number of months the customer has stayed with the company
7	PhoneService	String	Whether the customer has a phone service or not (Yes, No)
8	MultipleLines	String	Whether the customer has multiple lines or not (Yes, No, No phone service)
9	InternetService	String	Customer's internet service provider (DSL, Fiber optic, No)
10	OnlineSecurity	String	Whether the customer has online security or not (Yes, No, No internet service)
11	OnlineBackup	String	Whether the customer has online backup or not (Yes, No, No internet service)
12	DeviceProtection	String	Whether the customer has device protection or not (Yes, No, No internet service)

13	TechSupport	String	Whether the customer has tech support or not (Yes, No, No internet service)
14	StreamingTV	String	Whether the customer has streaming TV or not (Yes, No, No internet service)
15	StreamingMovies	String	Whether the customer has streaming movies or not (Yes, No, No internet service)
16	Contract	String	The contract term of the customer (Month-to-month, One year, Two year)
17	PaperlessBilling	String	Whether the customer has paperless billing or not (Yes, No)
18	PaymentMethod	String	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
19	MonthlyCharges	Numeric	The amount charged to the customer monthly
20	TotalCharges	Numeric	The total amount charged to the customer
21	Churn	String	Whether the customer churned or not (Yes or No)

4. Exploratory data analysis

At first, we did data cleaning. We checked for any null or blank values in the data. We found the column “Total Charges” had 11 rows of blank data, so we removed those rows from the dataset.

The Churn column tells us the number of customers who left within the last month. So, we visualized the Percentage of Customer Churned. As we can see from the pie chart, 26.54% have left within the last month.

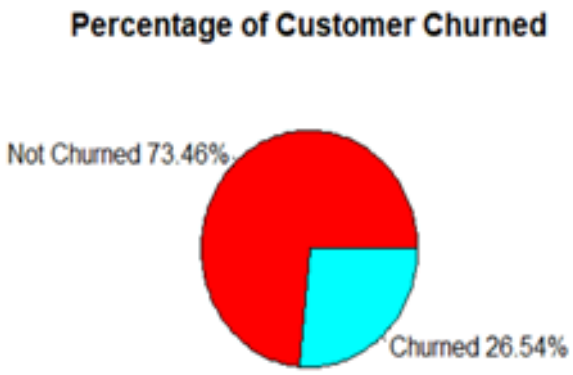

Figure 1

From Figure 2, the churn percent is almost the same in males and females i.e. around 25%. Senior Citizens tend to churn more, approximately 46%. Customers who have Partners and Dependents have low churn rate than compared to customers who don’t have Partners and Dependents.

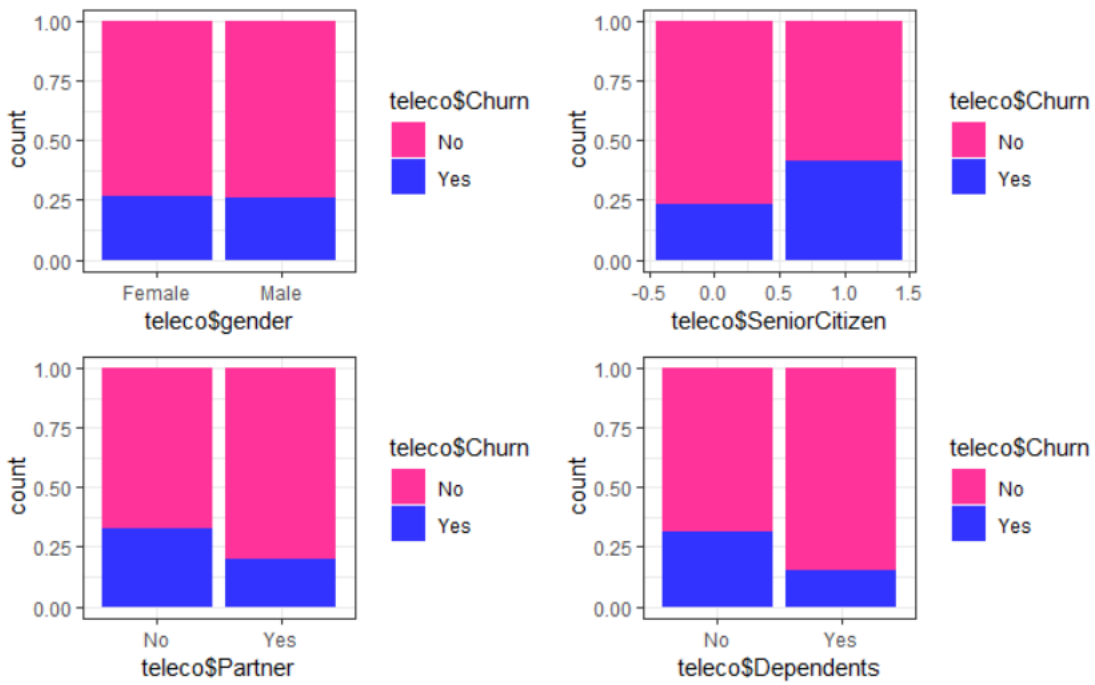

Figure 2

From Figure 3, we can see that in case of Fiber Optic "InternetServices" the churn rate is much higher almost 49% and around 25% of customers who do not have services like Phone Services and Multiple Lines have churned last month.

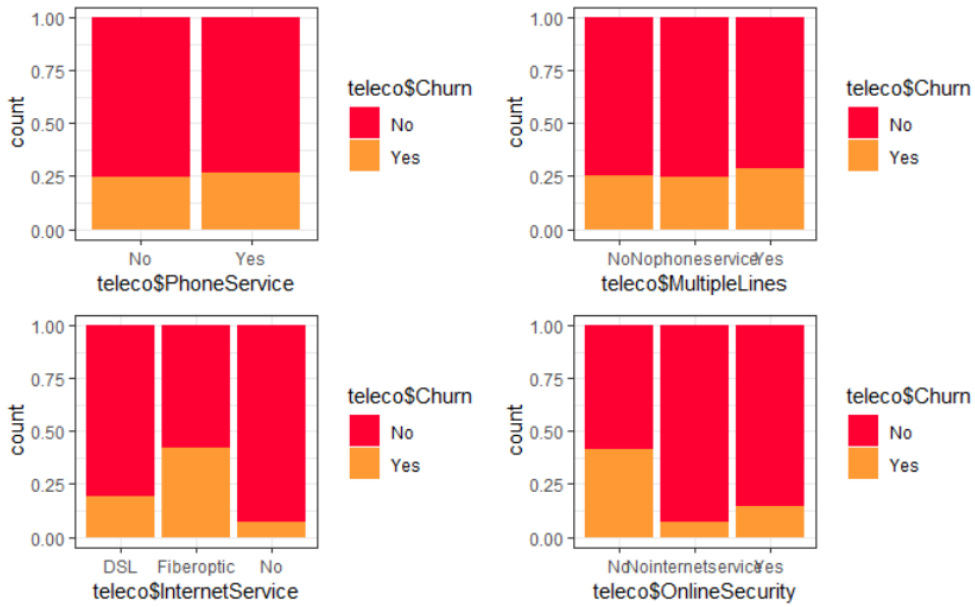

Figure 3

From Figure 4, we can see that almost 45% of customers with No online backup, no device protection and no tech support have churned.

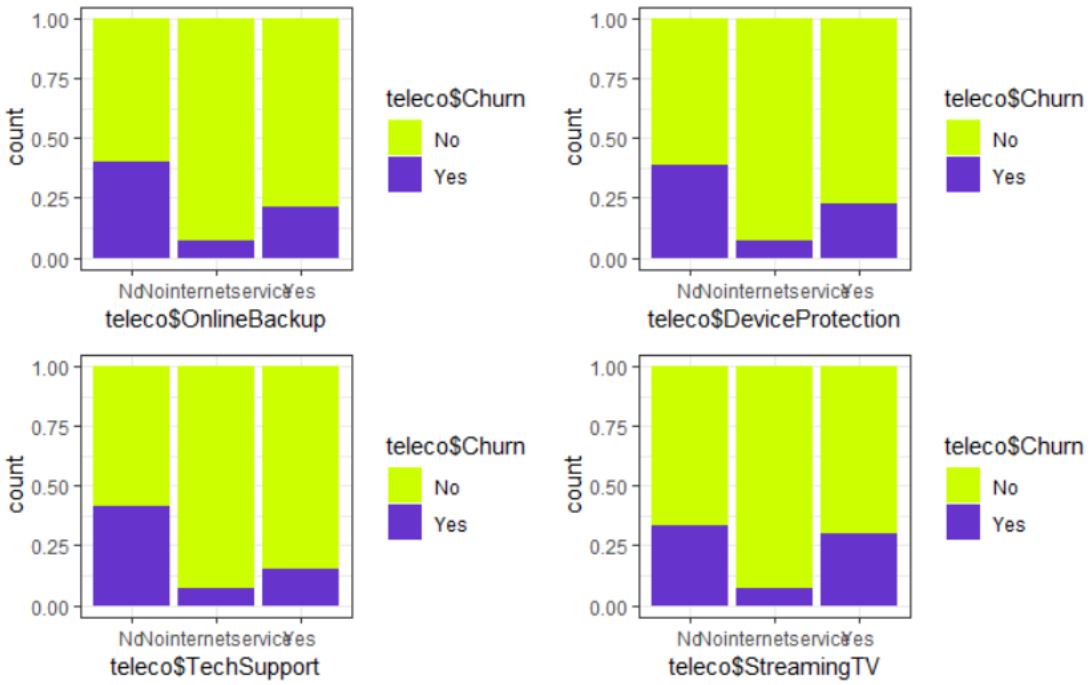

Figure 4

In Figure 5, we can see that customers with monthly subscription have churned, almost 45%, compared to customers with one- or two-year contract. For customers with paperless billing option churn rate is 35% and customers who have Electronic check as their Payment Method tend to leave more frequently when compared to others.

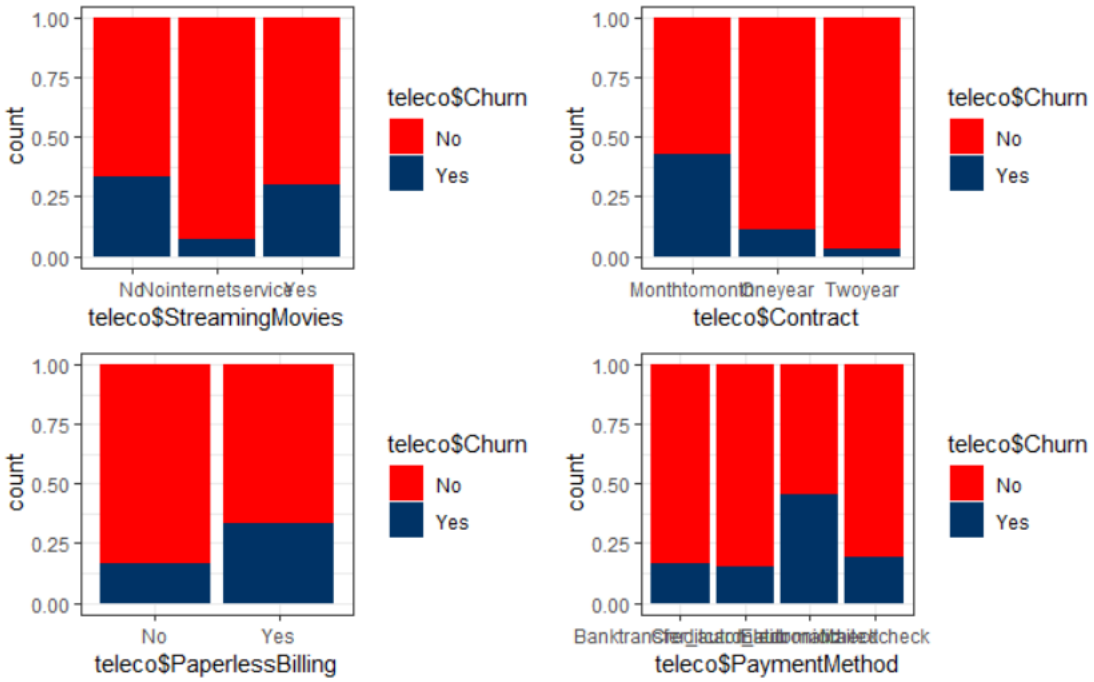

Figure 5

From Figure 6, the median tenure for customers who have churned is around 10 months and that of those who have not churned is approximately 40 months. Also, from the size of the box plots we can see that the number of churned people is less as compared to those who have not churned.

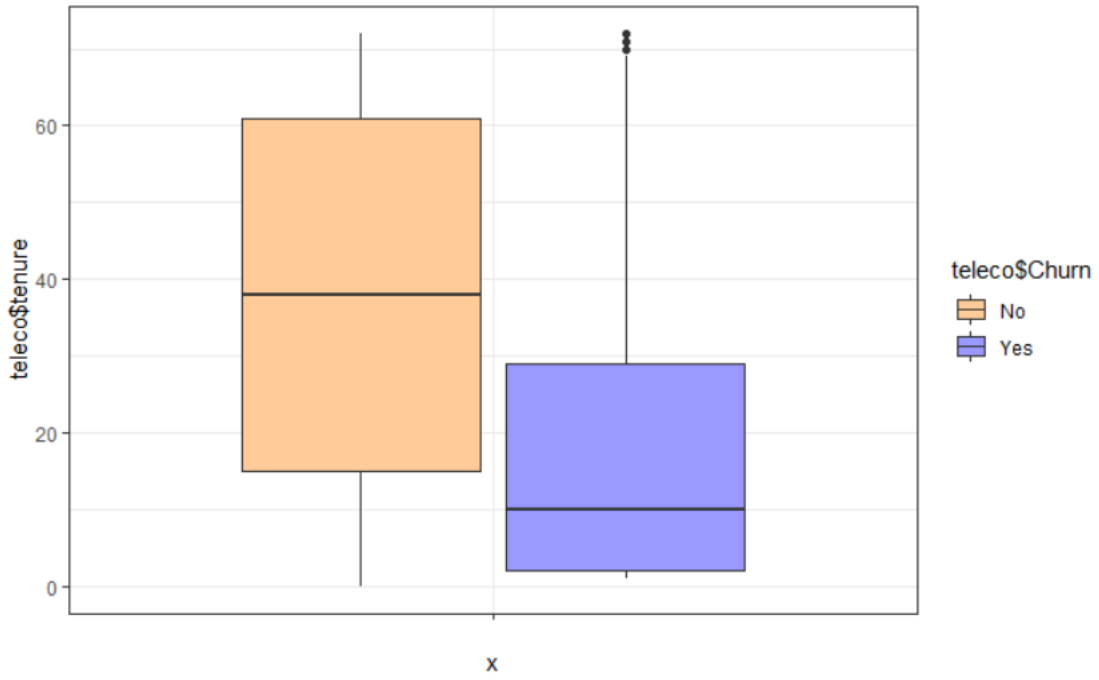

Figure 6

From Figure 7, The median of customers who have churned, below 1250, are people with low total charges and the median of the customers who have not churned is above 1250.

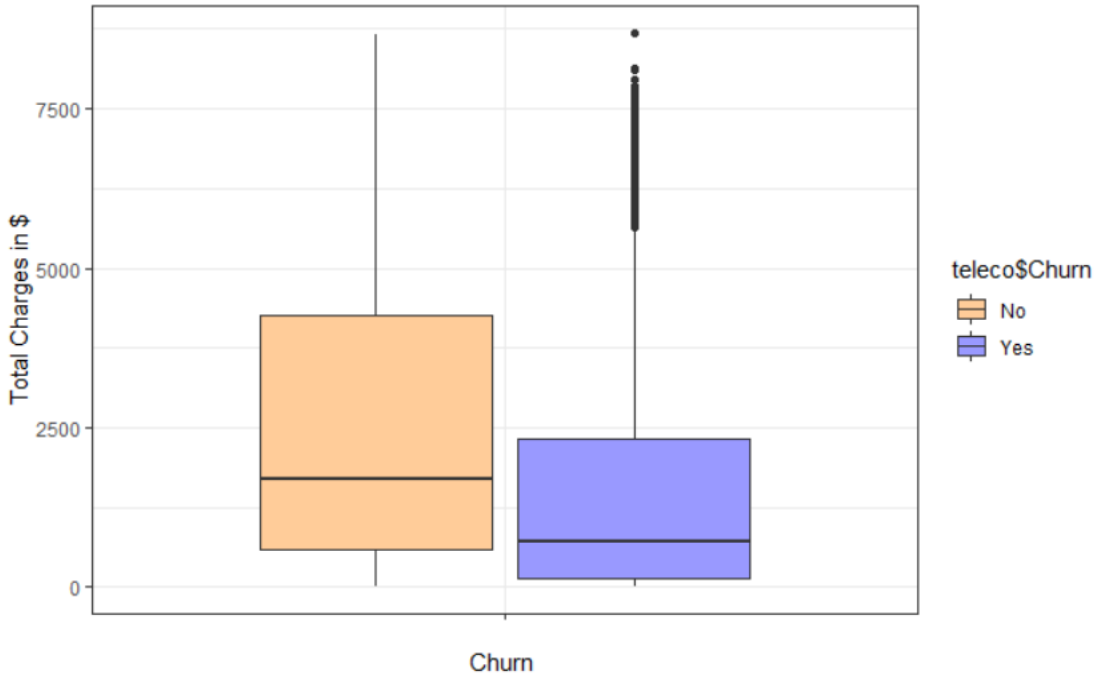

Figure 7

From Figure 8, we can see that Total Charges has positive correlation with Monthly Charges and tenure. There is little / no relation between Monthly Charges and tenure

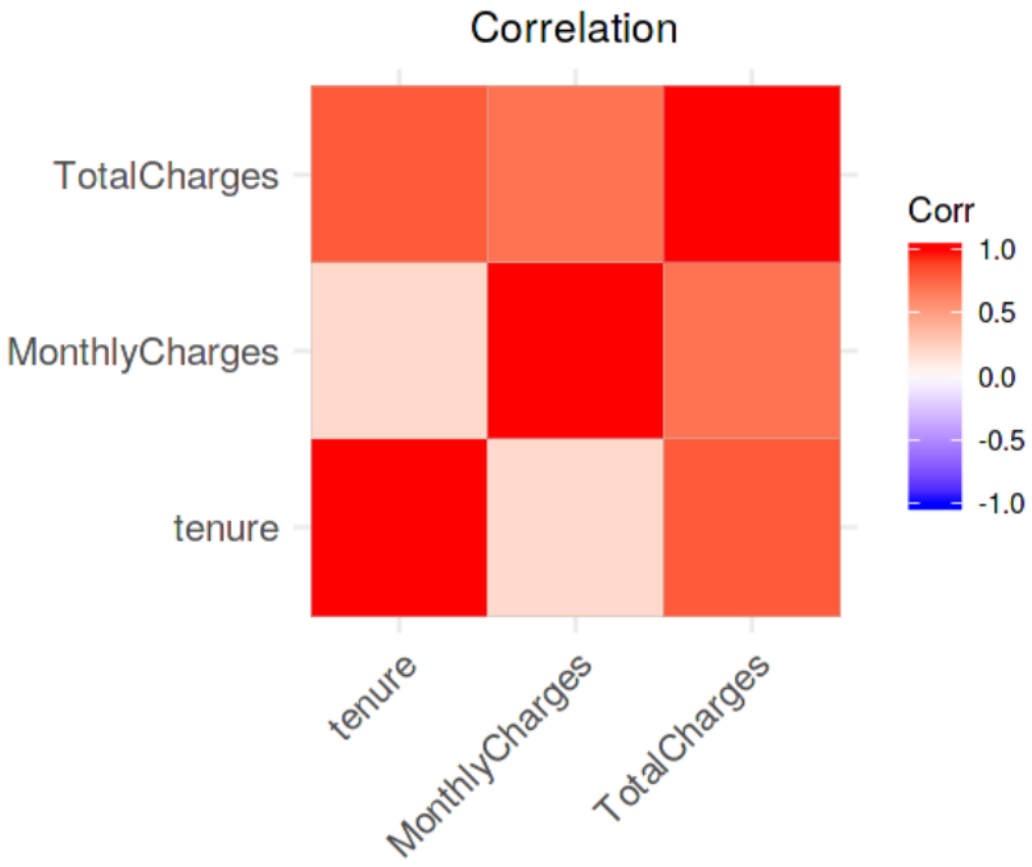

Figure 8

4. Models and analysis

We split the data into train and test data with a ratio of 0.8 – 80% training data and 20% testing data. We ran seven models on our train and test data. Following are the seven models:

4.1 Logistic Regression

Logistic Regression model is used when the dependent or target variable is categorical.

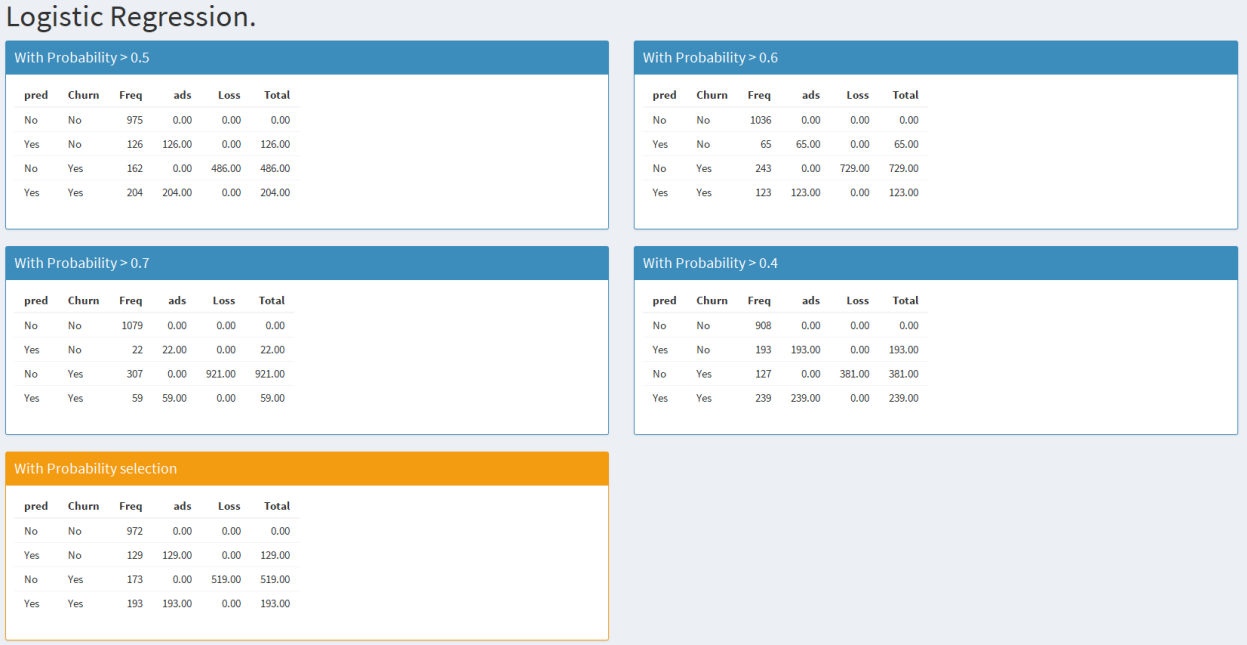

Figure 9

For the logistic model by increasing the threshold the accuracy increases but the cost also increases which is not ideal. The model with 7 predictors has the best accuracy of 79.41% and the model with all the variables with threshold of 0.4 has the least cost. The optimal model is one with least cost and relatively high accuracy which is the logistic model with all the variables with the threshold of 0.5 because the accuracy is 79% while its cost difference is minimal.

4.2 Linear Discriminant Analysis

Discriminant analysis is used to determine which variables discriminate between two or more naturally occurring groups. Logistic regression is a classification algorithm traditionally limited to only two-class classification problems. If you have more than two classes, then Linear Discriminant Analysis is the preferred linear classification technique.

LDA

LDA confusion matrix

lda_classify	Var2	Freq	ads	Loss	Total
No	No	961	0.00	0.00	0.00
Yes	No	140	140.00	0.00	140.00
No	Yes	163	0.00	489.00	489.00
Yes	Yes	203	203.00	0.00	203.00

LDA confusion matrix_Selection

lda_classify_selection	Var2	Freq	ads	Loss	Total
No	No	966	0.00	0.00	0.00
Yes	No	135	135.00	0.00	135.00
No	Yes	170	0.00	0.00	0.00
Yes	Yes	196	196.00	588.00	784.00

Figure 10

In our LDA models with and without applying selection method to limit our predictors we can see that there is a significant difference in cost and accuracy. By choosing the least cost over better accuracy we can conclude that the LDA model without selection with the cost of 832 and the accuracy of 75.46% is the optimal LDA model.

4.3 Quadratic Discriminant Analysis

QDA is closely related to linear discriminant analysis (LDA), but unlike LDA, in QDA there is no assumption that the covariance of each of the classes is identical.

QDA

QDA confusion matrix

qda_classify	Var2	Freq	ads	Loss	Total
No	No	824	0.00	0.00	0.00
Yes	No	277	277.00	0.00	277.00
No	Yes	83	0.00	249.00	249.00
Yes	Yes	283	283.00	0.00	283.00

QDA confusion matrix_Selection

qda_classify_Selection	Var2	Freq	ads	Loss	Total
No	No	852	0.00	0.00	0.00
Yes	No	249	249.00	0.00	249.00
No	Yes	99	0.00	297.00	297.00
Yes	Yes	267	267.00	0.00	267.00

Figure 11

In our QDA models with and without applying selection method to limit our predictors we can see that there isn't a significant difference in cost and accuracy. However, by choosing the least cost over better accuracy we can conclude that the QDA model without selection with the cost of 809 is the optimal QDA model.

4.4 K-Nearest Neighbors

KNN is a non-parametric, supervised algorithm that does not make any assumptions on the underlying data distribution and can be used for both classification and regression problems.

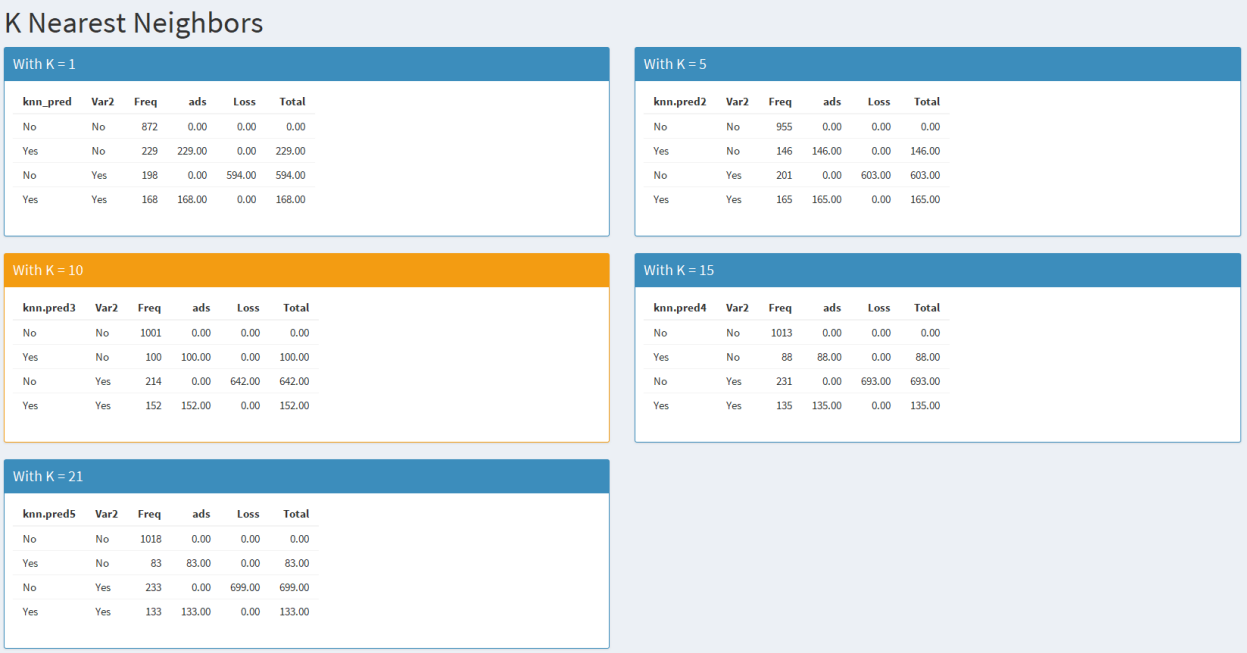

Figure 12

We ran the KNN model with inclusion of all variables and for different K's we have found out that the k=10 with the cost of 894 and the accuracy of 78.59% is the best KNN model.

4.5 Decision Trees

Decision tree is a type of supervised algorithm that is mostly used in classification problems and easy to understand. In this, we split the population or sample into two or more homogeneous sets based on most significant splitter in input variables.

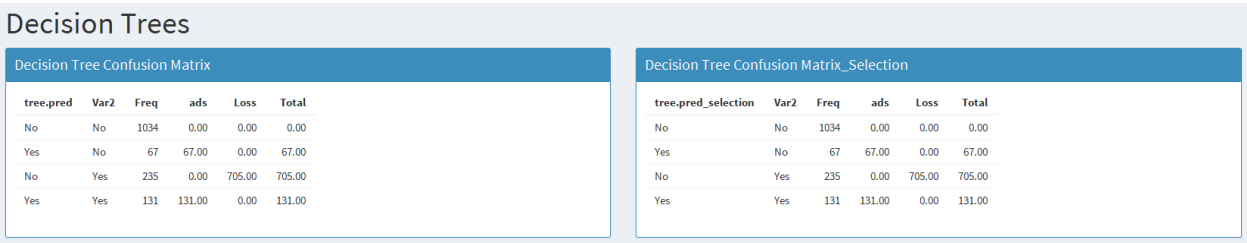

Figure 13

Using decision model to predict our churn with or without selection gave us the same accuracy of 79.41% and the least cost of $903.

4.6 Random Forest

Random forest is a supervised algorithm that builds multiple decision trees trained with bagging method and merges them together to get a more accurate and stable prediction.

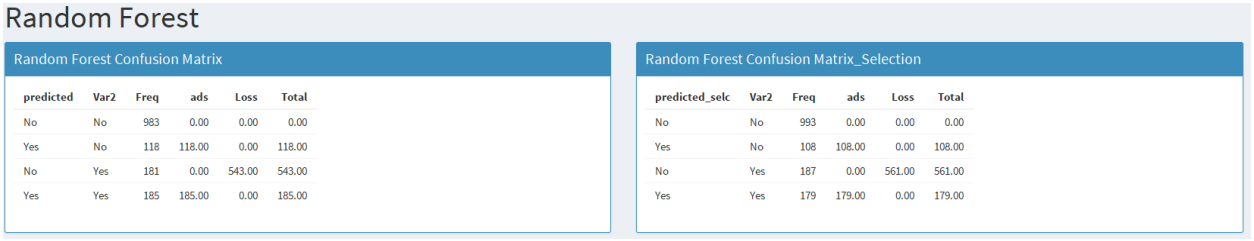

Figure 14

Using naïve Bayes to predict our target variable without selection method gave us a better accuracy of 79.61% and least cost of $846.

4.7 Naïve Bayes

Naive Bayes model is easy to build and particularly useful for very large data sets. It is a classification technique based on Bayes' Theorem and assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

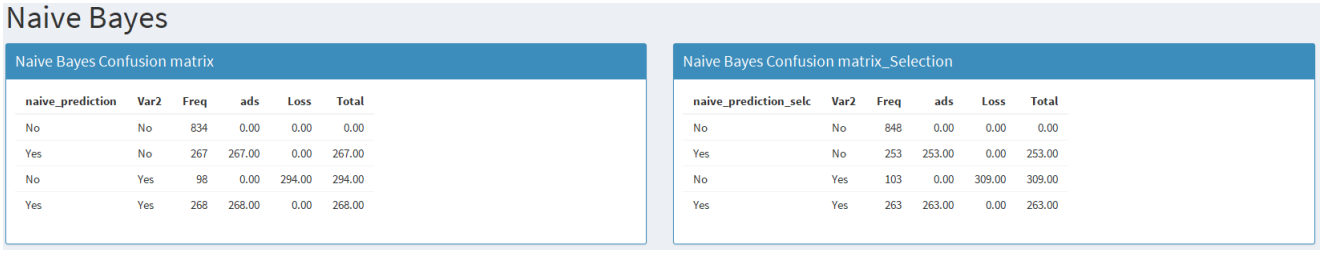

Figure 15

Using naïve Bayes to predict our target variable with selection method gave us a better accuracy of 75.27% and least cost of $825.

5. Findings and managerial implications

Forward selection	Backward selection	logistic regression significance
tenure	-	99%
InternetService_Fiberoptic	InternetService_Fiberoptic	-
PaymentMethod_Electroniccheck	PaymentMethod_Electroniccheck	-
OnlineSecurity_No	OnlineSecurity_No	-
TechSupport_No	-	NS
Contract_Monthtomonth	Contract_Monthtomonth	-
-	PaperlessBilling_yes	99%
-	Monthly charges	NS
-	Total charges	99%

Figure 16

We have run the models with all the predictors, then have computed the cost and accuracy for every model (logistic, QDA, LDA, KNN, Random forest, Decision Tree, naïve Bayes). Using maximum accuracy and minimum cost for the selection of best model we have selected QDA model with the accuracy of 75.46% and minimum cost of $809 as the best model including all the predictors.

We have used selection methods to reduce our predictors to 7 and have run all the models using the chosen predictors to find out if it will improve accuracy or reduce the cost. Using maximum accuracy and minimum cost for the selection of best model we have selected QDA model with the accuracy of 76.27% and minimum cost of $813 as the best model including the7 predictors.

Doing a trade-off between accuracy and cost by giving priority to minimum cost over accuracy we have concluded that the QDA model with all the predictors seems to be the best model to predict Telco customer churn while minimizing the advertising cost.

6. Conclusions

We ran seven models on our training and testing data. We calculated the accuracy of the models and also calculated the cost of advertising for the customer who is going to churn or the amount that is lost if the customer leaves. Following are the outputs and conclusions of the models:

LOGISTIC REGRESSION MODEL					
THRESHOLD	**0.5**	0.6	0.7	0.4	**0.5 without selection**
COST	**816**	917	1002	813	841
ACCURACY	**0.7941**	0.79	0.7757	0.7818	**0.7941**

Table 1

For Logistic Regression, we found the best model by changing the threshold from 0.4 – 0.8, with and without selection. As we kept increasing the threshold value from 0.5 to 0.7, the accuracy of the model kept decreasing. So, we ran the model for threshold of 0.4 and found that the accuracy was still less compared to the threshold value of 0.5. Also, the costs kept increasing. From the table above, we can see that the logistic regression model with threshold of 0.5 and with and without selection has the same accuracy but considerably different cost. So, we will choose the model with selection, threshold of 0.5, accuracy of 79.41% and cost of $816.

LDA MODEL		
	WITHOUT SELECTION	**WITH SELECTION**
ACCURACY	0.7546	**0.7627**
COST	**832**	919

Table 2

In LDA model, we can see that although the model with selection has a higher accuracy of 76.27% but the model without selection has lower cost. So, we will choose the model without selection having an accuracy of 75.46% and cost of $832.

QDA MODEL		
	WITHOUT SELECTION	**WITH SELECTION**

ACCURACY	0.754601227	**0.762781186**
COST	**809**	813

Table 3

For QDA, we can see that the model with selection has the highest accuracy of 76.27% but also has a higher cost of $813. Although, the QDA model without selection has lower accuracy of 75.46% but has lower cost of $809.

KNN					
k	1	5	**10**	15	21
COST	988	915	**894**	917	915
ACCURACY	0.7089	0.7634	**0.7859**	0.7825	0.7845

Table 4

For KNN, we can see clearly from the table above, the model with the nearest neighbors, k equal to 10 has the best accuracy of 78.59% and the lowest cost of $894.

DECISION TREE MODEL		
	WITHOUT SELECTION	**WITH SELECTION**
ACCURACY	**0.794137696**	0.794137696
COST	**903**	903

Table 5

For Decision Tree, from the table we can see that the models with and without selection both produce the same results of accuracy 79.41% and a cost of $903.

NAÏVE BAYES		
	WITHOUT SELECTION	**WITH SELECTION**
ACCURACY	0.751192911	**0.752777778**
COST	829	**825**

Table 6

For the Naïve Bayes model, from the table we can see that the models with selection produce results with better accuracy of 75.27% and lower cost of $825 as compared to the model without selection.

RANDOM FOREST		
	WITHOUT SELECTION	**WITH SELECTION**
ACCURACY	**0.796182686**	0.798909339
COST	**846**	848

Table 7

For the Random forest model, from the table we can see that both models with and without selections produce approximately same results. We will choose the model without selection having accuracy of 79.61% and cost of $846.

From the above analysis, we conclude that since the QDA model without selection gives us an accuracy of 75.46% and the lowest cost of $809, among all the models, we will select this model.

7. Appendix

We build an interactive web application using an R package called Shiny. A Shiny application is simply a directory containing a user-interface definition, a server script, and any additional data, scripts, or other resources required to support the application. We have included all our code in a file format .rmd file, that makes dynamic documents with R and contains chunks of embedded R code.

8. References

1. Kaggle – Dataset - https://www.kaggle.com/blastchar/telco-customer-churn
2. James, G., D. Witten, T. Hastie, and R. Tibshirani. An introduction to statistical learning. Vol. 112. New York: springer, 2013. Website: www.StatLearning.com
3. Wickman, H. and G. Grolemund. R for Data Science. O'Reilly, 2017 Website: http://r4ds.had.co.nz/introduction.html