

**Telco Customer Churn Reduction Analysis**

**A Web & Social Analytics Project**



**INSY 5377 001**

**Web and Social Analytics**

**Summer 2020**

**Final Project Report**

**Submitted By: Team 9**

**Roman Basnet**

**Afroza Hossain**

**Wenhan Ji**

**Kanchan,Prathvi Narasimha**

**Yujie Hu**

**Abstract**

Our study seeks to find relevant correlations and relationships among (1) Female Gender with electronic checks will likely to be retained as customers. (2) Customer with partner and no dependent will not retained as customer to the company with high probabilities. (3) If we reduce the monthly rates of customers who have multiple lines (including phone service); we are likely to retain more customers. Telco is a telecommunications or communication service provider (CSP) company that transports information electronically through telephony and data communication services in the networking industry. Our objective is to find out what are the strategies Telco should follow to retain their customers (i.e. figure out what services are highly related to customer churn). We downloaded the dataset from Kaggle which has 7043 observations. Dependent variable is Churn and we have 19 independent features which are:

Gender, SenorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges , TotalCharges.

Variables we will be focusing on are demographical variables, contract-related variables: contract types payment-related variables: monthlycharges/totalcharges.

**Introduction**

Telco is tele company here in the [United States](https://en.wikipedia.org/wiki/United_States) from the beginning of [mobile telephony](https://en.wikipedia.org/wiki/Mobile_telephony), which now includes [wireless carriers](https://en.wikipedia.org/wiki/Wireless_carrier), or [mobile network operators](https://en.wikipedia.org/wiki/Mobile_network_operator) and internet. The objective of this project is to understand what strategies Telco should follow to retain their customers (i.e. figure out what services are highly related to customer churn). We plan to use multiple models like logistic regression, decision trees, random forest etc. for our analysis to predict customer retention and select the one that gives us the best results. We intend to use Tableau/Excel/SAAS for visualization to understand and analyze the data.

## **Data Preprocessing**

We downloaded the Telco dataset from the Kaggle website contains 7043 observations and 20 columns. After removing all the null values and checking for duplicate records, our dataset contained 5517 records. We used Churn(yes=1) as our dependent variable. And we expected to use the other 18 columns as independent variables. Following are the independent variables:

Gender - Whether the customer is a male or a female

Senior citizen - Whether the customer is a senior citizen or not

Partner - Whether the customer has a partner or not

Dependents - Whether the customer has dependents or not

Tenure - Number of months the customer has stayed with the company

PhoneService - Whether the customer has a phone service or not (Yes, No)

MultipleLines - Whether the customer has multiple lines or not (Yes, No, No phone service)

InternetService - Customer’s internet service provider (DSL, Fiber optic, No)

OnlineSecurity - Whether the customer has online security or not (Yes, No, No internet service)

OnlineBackup - Whether the customer has online backup or not (Yes, No, No internet service)

DeviceProtection - Whether the customer has device protection or not (Yes, No, No internet service)

TechSupport - Whether the customer has tech support or not (Yes, No, No internet service)

StreamingTV - Whether the customer has streaming TV or not (Yes, No, No internet service)

StreamingMovies - Whether the customer has streaming movies or not (Yes, No, No internet service)

Contract - The contract term of the customer (Month-to-month, One year, Two year)

PaperlessBilling - Whether the customer has paperless billing or not (Yes, No)

PaymentMethod - The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

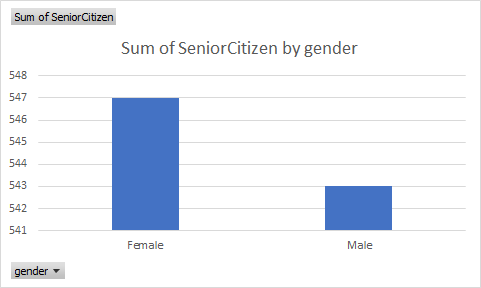
MonthlyCharges - The amount charged to the customer monthly

TotalCharges - The total amount charged to the customer

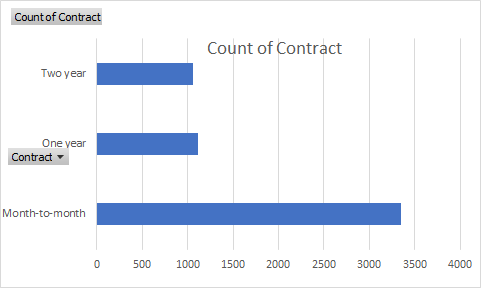
A good business model considers the variables which are independent of each other. As a result, we ignored the column "Totalcharges" because column "tenure" multiply by "Monthlycharges" equals to "Totalcharges".

## **Data Visualization**

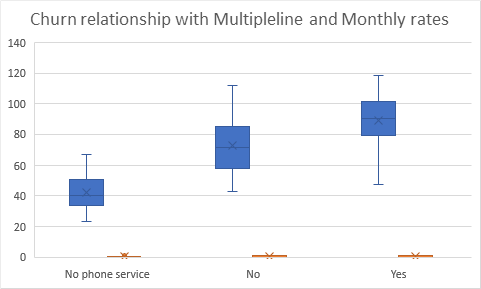
Visualization is an easy technique to visualize the relationships between the features in our data. Our goal is to decide which columns are more important to help our team predict what type of clients are more probable to Churn. From general observations we believed that some columns such as Monthlycharges; Multiplelines; Paymentmethod and Gender could give us some significant supports. As a result, we used bar chart, column chart, pie chart and box plot to visualize our data, expecting to find some patterns for our predict models. And the below are some of the examples we used to explore the data in general.



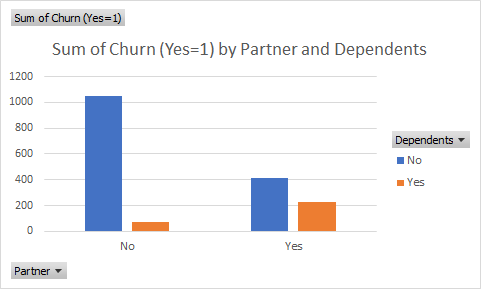
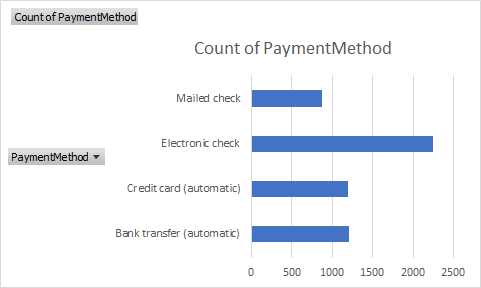
The first graph is about the senior citizen by gender, we can see that we have a lot of female senior citizens are Telco clients.



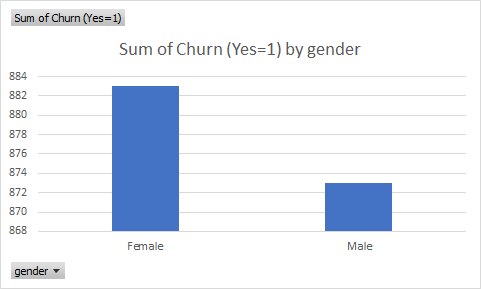
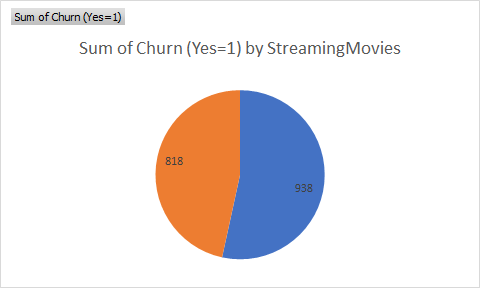
In this bar chart, we can clearly see that most people choose to use month-to-month contract.



This is a boxplot, relationship between Multiplelines and Monthly rates. We can easily observe that higher rates have higher Churn in the graph.



The graphs above can also show some relationship between what type of payment the clients prefer to use. Also, how partner status and dependent situation relate to churn. We can see more people uses Electronic checks and a lot of clients are no dependents and no partners.



The pie chart gives the team a direct percentages comparation between about whether the clients StreamingMovies or not. And the bar chart shows which gender is more relate to Churn (Yes=1).

## **Prediction Models:**

### **Logistic Regression:**

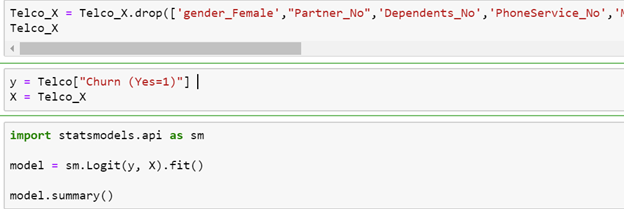
|  |  |  |  |
| --- | --- | --- | --- |
| Internet Service | No | Yes | Grand Total |
| DSL | 1962 | 459 | 2421 |
| Fiber optic | 1799 | 1297 | 3096 |
| No | 1413 | 113 | 1526 |
| Grand Total | 5174 | 1869 | 7043 |

From table above, we can see churn rate is very low (113/1526 = 7.2%) for people who don’t have internet service (1526 cases). The analysis of this project is focusing on customers who have internet service, which has a total of 5517 cases (churn rate = (459+1297)/ (2421+3096) = 32%)).

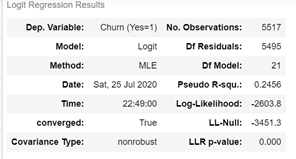
To find the most important variables which are affecting customer churn in this Telco company, we choose numerous prediction techniques--logistic regression model and decision tree model to predict churn. Python Jupyter notebook is used to build the models and analyze the results.

Since our target variable is a binary variable, logistic regression model is chosen first. dependent variable (y) is Churn (yes =1). Independent variables have been transformed into dummy variables if they are categorical variables. When the logistical regression model is built, n-1 dummy variables are included in the model. Below is python code for logistical regression model.





The findings of the logistic regression are shown below:





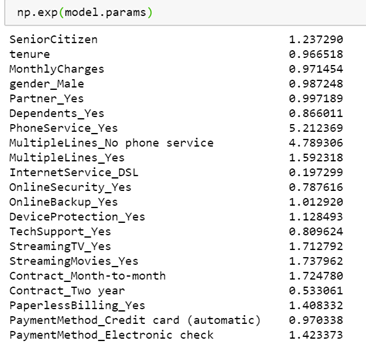


From the results shown above, predictors such as SeniorCitizen, tenure, Multiplelines, InternetService\_DSL, Contract\_Month-to-month, Contract\_Two year, PaperlessBilling, PaymentMethod\_Electronic check are statistically significant in explaining churn rate (since they are associated with very small P values which is less than 5%). Gender, Partner, Dependents, Online Streaming services are not significant towards to customer churn in this model analysis.

Coef. column shown in the picture above represents the coefficients for each of variables. If the coefficient is positive, it implies that the predictor has positive relationship with customer churn. If the coefficient has a negative value, then it means that it will decrease the probability of customer churn. For example, Senior Citizen and Contract (month-to-month) have positive relationships with customer churn, which indicates if customers choose to pay for services month-to-month instead of having one-year or two-year contracts with the company or customer is a senior citizen, the probability of churn is increasing.

If customers choose to use DSL Internet service, then the probability of customer churn is reduced, since this variable is associated with negative coefficients. Variable Tenure has a negative affect towards to customer churn because the coefficient associated with it is negative (Coef. = -0.0341). It also indicates with one unit increase in Tenure; the customer churn rate will decrease by 3.4%.

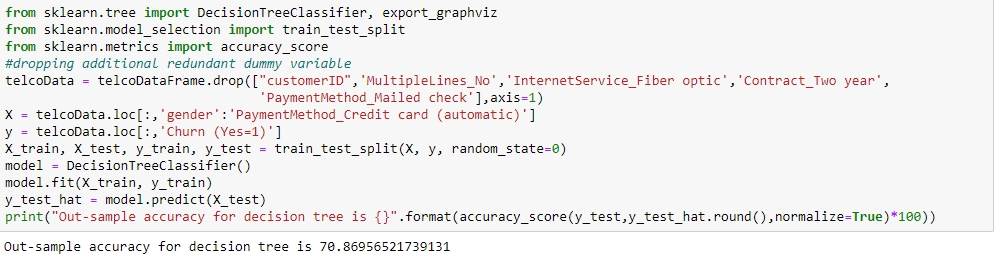
Since several variables included in this logistical model are categorical variables, odds are calculated for explaining how a 1 unit increase or decrease in a variable affects the odds of customer churn. For example, we can expect the odds of churn to decrease by about 80% if people switch their Internet service with DSL. If people sign a two-year contract, odds of churn will decrease 1-0.533=46.7%.



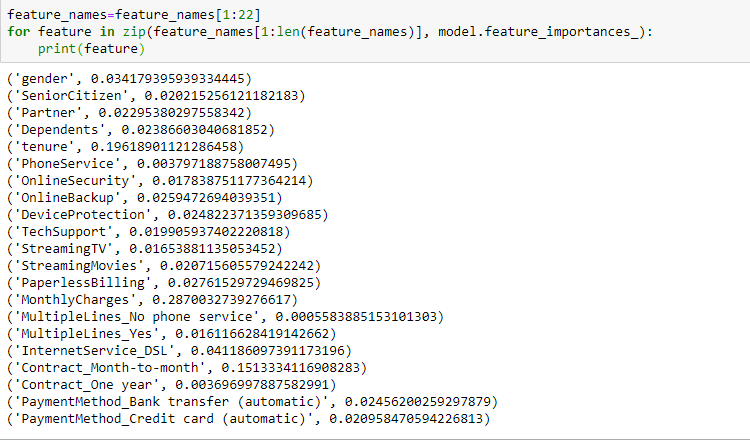
In terms of getting predictive accuracy of the logistic regression model, data has been partitioned to 70% training and 30% testing. Training data is being used to build the model, then the fitted model is used to predict testing data. Accuracy of predicting test data is 77% when this logistic model is used.

**Decision Tree:**

Decision Tree is one of the commonly used tools in machine learning analysis. This exploratory analysis can help us in short-listing features that determine customer churn effectively. We used Decision Tree Classifier from sci-kit library to develop the prediction model. Please find below the code snippet.



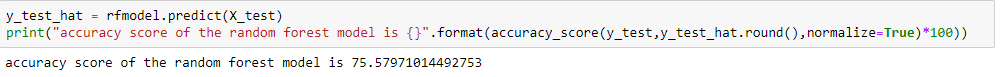
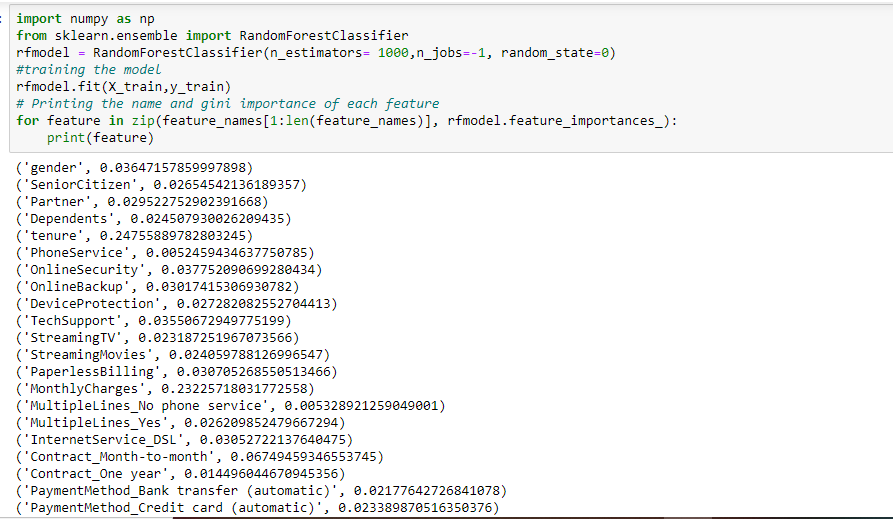
From the above result we can see that if we consider all the features, the model prediction is around 71%. Let us see the important features as per this model.



As per the above feature importance values, we choose the features having high Gini Importance. So, Monthly Charges, tenure, Contract Month-to-Month are important features.

### **Random Forest Classifier**

Random forest analysis is another machine learning classification method which creates multiple decision trees and constructing models based on summary statistics of these decision trees.



Based on the above results, accuracy of random forest is almost 76% and tenure, monthly charges are significant features. If we round up the Gini Importance value for Contract\_Month-to-month it is around 0.1 approximately. Hence, Contract\_Month-to-month is also significant feature contributing to customer churn.

The logistic model gives us the numeric insight about the impact on customer churn per unit increase or decrease in the important independent variables and the random forest/decision tree models helps us shortlist what are the important independent features.

**Discussion**

Based on our logistic regression model and the results we got, we can say that customers who have two-year contracts are more likely to stay with the company compared to those who are on a month-to-month contract. If you think about it, that makes sense because customers who are on a long-term contract are often offered low rates, leading to lower churn rate among those customers. Also, if customers want to get out of their long-term contracts, they are usually charged a fee which stops them from leaving, compared to people with month-to-month contracts who don’t have that obligation and are free to leave.

Another important finding from our logistic regression model is that customers with DSL internet have low churn rate compared to customers with fiber optics. As a business, having this information, they can do two things:

1. Encourage your customers to switch to service with DSL internet and tie them to two-year contract.
2. Start new promotions, provide discounts, bundle fiber optics internet service with some product for low rates and observe for few months if that decreases the churn rate.

Based on our predictive models, decision tree and random forest, we can say that monthly charges, tenure and month-to-month contract are important factors for predicting churn. If you think about it, it makes sense as well because if monthly charges are high, people tend to leave the company and stay if the charges are low. It’s the same with tenure, companies tend to offer low rates to old customers which makes them want to stay more. About month-to-month contract being significant for churn rate; like we said earlier people with long-term contract tend to stay longer because of the low rates offered and fees they must pay to break the contract; same concept applies here as well.

The results of the prediction models are aligned with one of our hypotheses that if we reduce monthly charges of customers with multiple lines then the company is likely to retain more customers. We can say that this hypothesis is true because monthly charges is one of the significant factors predicting churn and if we decrease that we are likely to retain more customers.

**Conclusion**

Our objective was to find out what are the strategies Telco should follow to retain their customers, we have done that using different models. From our data visualization and logistic regression results we can conclude that customers who have two-year contracts are more likely to stay, because long-term contracts are often offered for lower monthly rates. So as data analysts, we would like to recommend Telco to work with their marketing team and prepare strategies that primarily targets customers with month to month contract and tempt them in signing long term contracts. We also found that we have a lot of female senior Citizens, most people choose to use month-to-month contract and found that higher rates have higher Churn. We would also recommend the company to work with their customer service team and try contacting their customers with month to month contracts and find out what will make them sign long term contracts and provide what the customers want if it’s feasible. Based on our predictive models, decision tree and random forest, we can conclude monthly charges, tenure and month-to-month contract are important factors for predicting churn. In order to compete more efficiently, Telco is required to divert their attention to find ways to provide excellent customer service, secure loyalty and reduce churn.

**Reference**

<https://www.kaggle.com/blastchar/telco-customer-churn>