AIM 5005 Machine Learning

**Project Final Report**

Due date: by August 10 (Tuesday), 2021

**Title:** **Customer Churn Analysis and Prediction**

Name **Xiaolan Li**

1. **Introduction**
   1. Problems

- Specify the problems that you would like to solve from this project. i.e., Specify why you do this project

Problems:

**- Q1: Which services should company improve to avoid current customer leaving?**

**- Q2: What kind of services are more related to whether a customer leave?**

**- Q3: Who are more likely to leave and what the confidence of that predictions are?**

Reason**:**

**In this project, I’m going to explore the reason behind customer churn and predict who’s about to leave currently. This can help the company to maximize the customer experience for preventing customer attrition. Therefore, it can help companies reduce losses and even increase the number of customers to increase profits. Additionally, this can be applied in every company, with a high impact value.**

* 1. Related (prior) works

- Review the literatures

- Describe the current status or previous relevant works done by others

- Describe how your project will be different from the prior works

The paper “A hybrid classification model for churn prediction based on customer clustering” used a hybrid classification combined with multiple layer perceptron, k-means and gradient boosting decision tree to lower the dimension, seperate customer into 4 dataset and then predict the result.

I will use multiple methods to do the prediction and evaluation.

I will use the PCA to do the dimension deduction then use the best performance algorithm from (LogisticRegression, RandomForest, Naive Bayes, and KNeighbours)to predict the result. If the result is not good, I will also add a cluster in the middle.

* 1. Possible outcomes of this project

- Final products (outcomes): Describe what you will produce when you finished this project

1. **Finding out the best model to predict if a customer is going to leave.**
2. **Finding out the features that lead to a customer leave and do an order of them.**
3. **Finding out the most important reason lead to current customer leave and provide some suggestions.**

1. **Technical Plans** 
   1. Dataset that you will use

**Downloaded from IBM community. Publish date May,2020.**

<https://community.ibm.com/accelerators/?context=analytics&query=telco%20churn&type=Data>

The data set is detailed in

[https://community.ibm.com/community/user/businessanalytics/blogs/stev en-macko/2019/07/11/telco-customer-churn-1113](https://community.ibm.com/community/user/businessanalytics/blogs/stev	en-macko/2019/07/11/telco-customer-churn-1113 )

**The data set includes information about:**

**-- Customers who left within the last month – the column is called 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**

* 1. Describe about learning methods that you will use for this project (if any)

**Logistic Regression**

**KNN model**

**Naive Bayes**

**Random Forest**

**PCA**

1. **Data**

**Data Prepossessing:**

At the very beginning, I checked the dataset that includes 33 columns and 7043 observations.

Among of all data, 11 data in Total Charges and 5174 data in Churn Reason columns are Nan value. 'Count', 'Country', 'State', 'CustomerID', 'City', 'Zip Code', 'Lat Long', 'Latitude', 'Longitude', 'Churn Label', 'Churn Score' columns are not helpful in this project. Thus after cleaned the data, I have 20 features, 1 label and 7032 observations on the dataset.

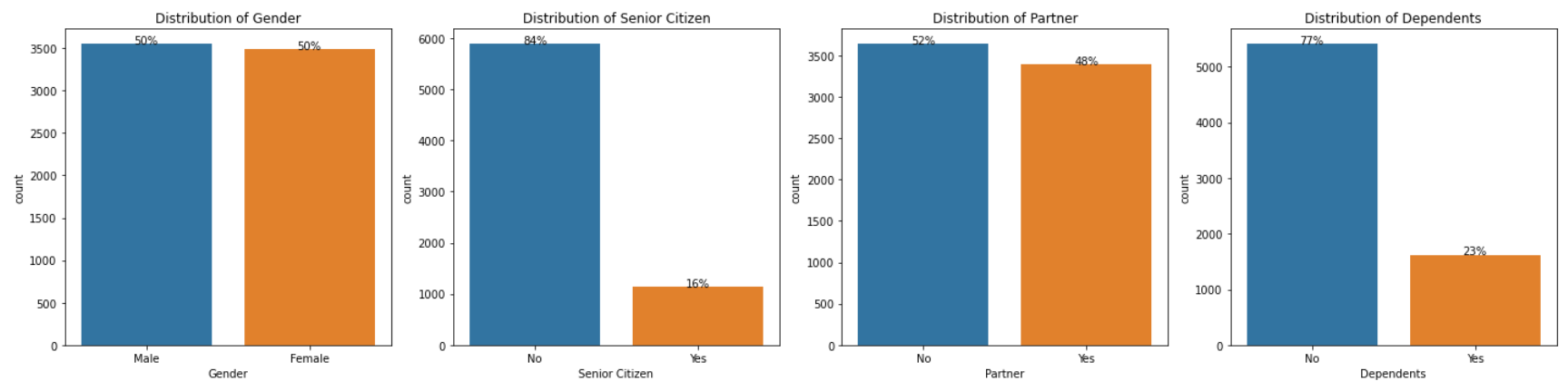
**Data Exploration Analysis**

**Label Data:**

There’re 5163 Churn label observations and 1869 Stay label observations which is quite imbalanced distribution

**Categorical Data:**

***Demographic\_group***

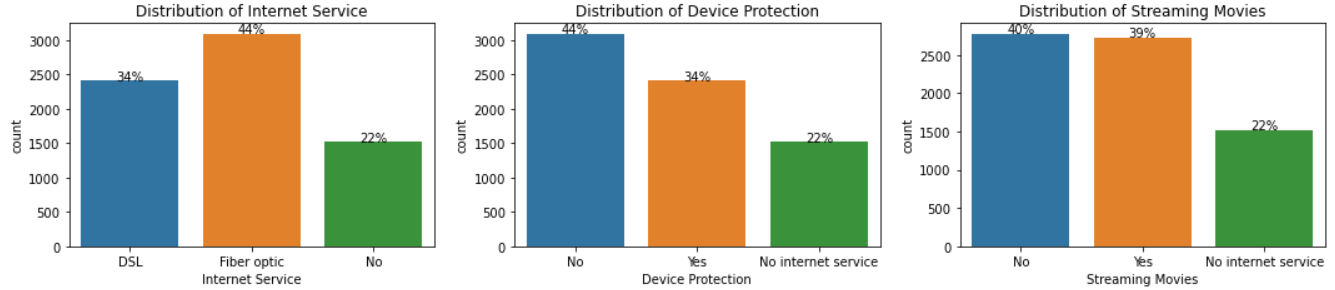
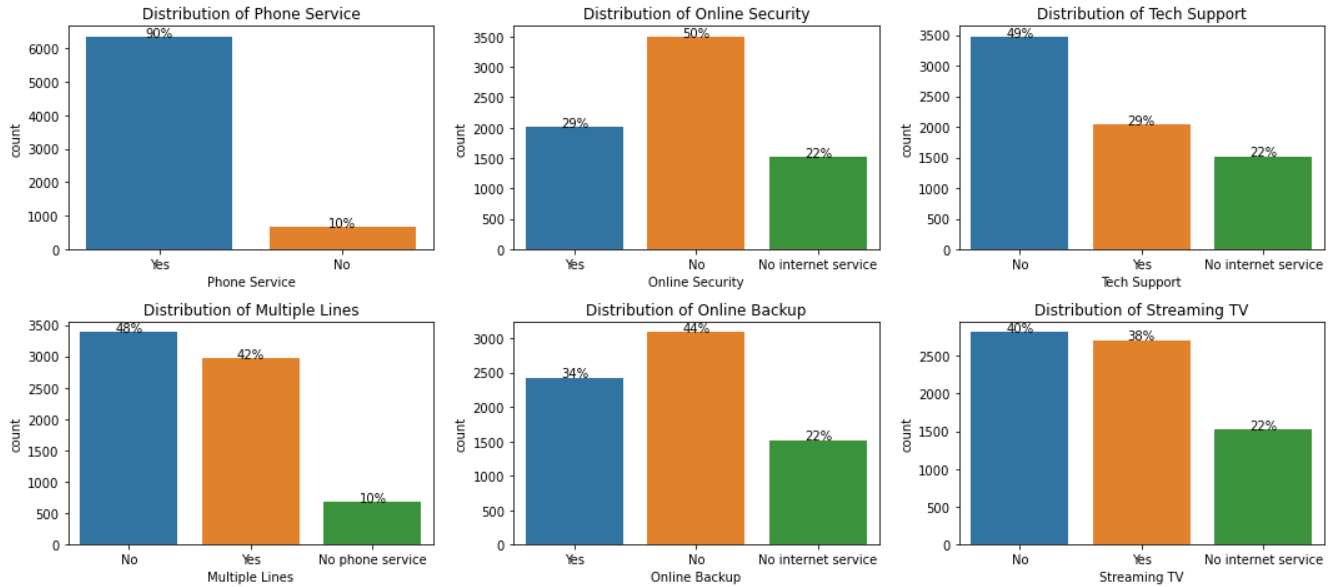


Equal distribution in Gender, and almost equal in Partner

Most of our customers are not Senior Citizen, only 1/6 are Senior Citizen

2/3 of our customers are Economic independents and 1/3 of our customers are enconomic dependent.

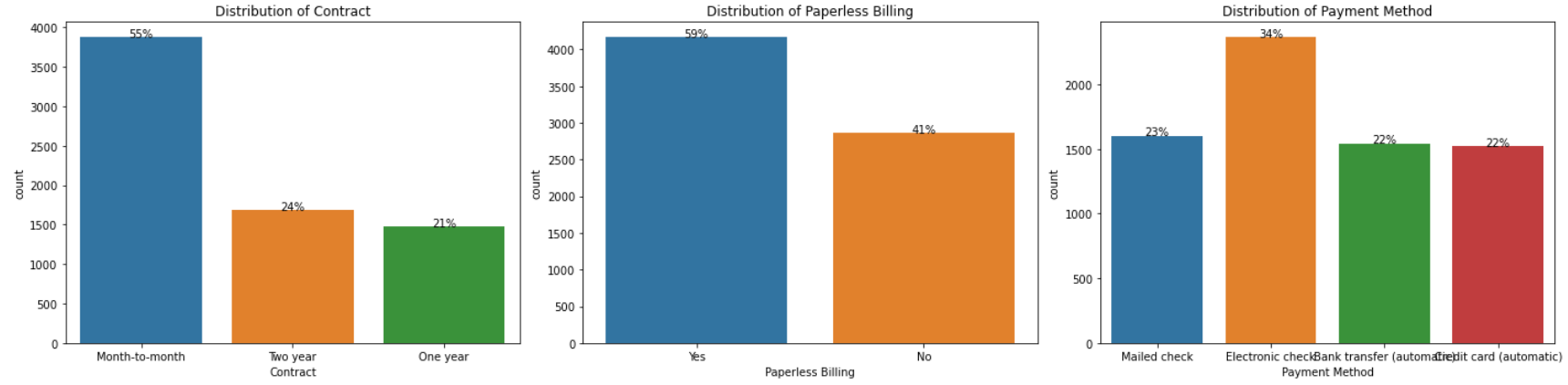
***Services\_group***



Most of our customers have used our Phone Services, among these customers, around 1/2 have Multiple Lines.

4/5 customers are using Internet Service, among these customers, around 1/3 have Online Security, Tech Support, around 1/2 have Online Backup, Device Protection, Streaming TV and Streaming Movies.

***Customer\_account\_group***



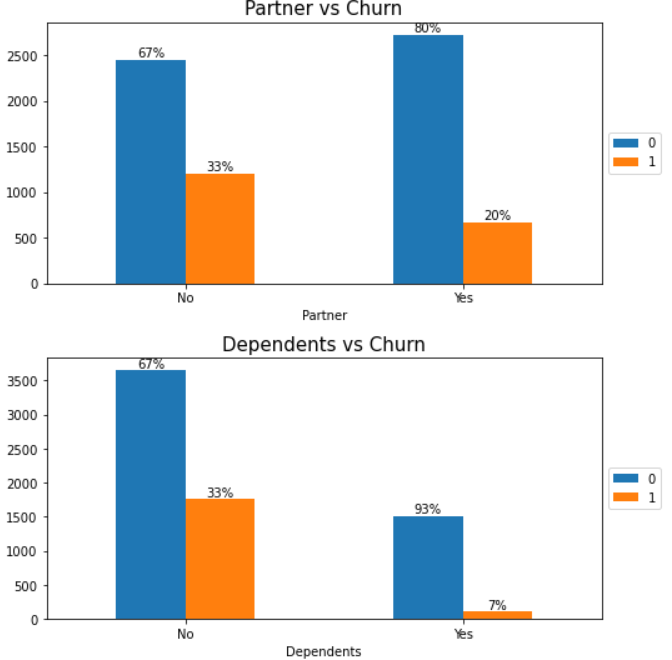
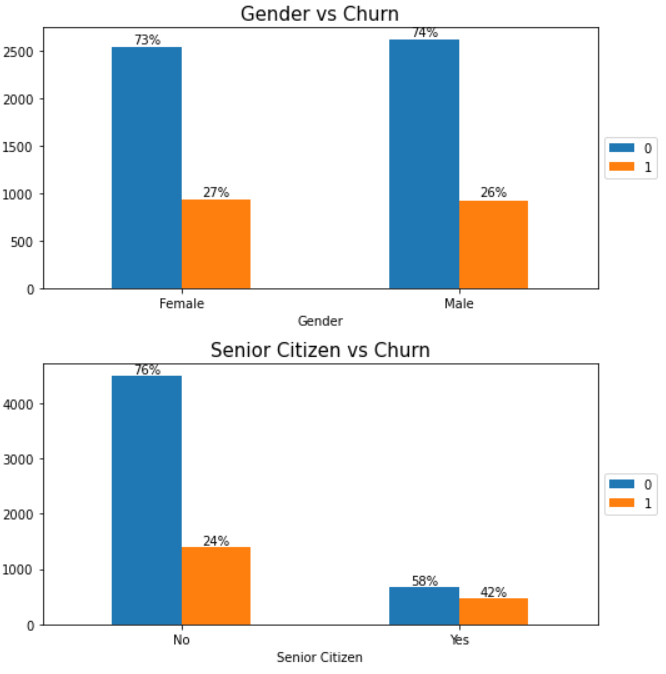
Around 1/2 customers signed month-to-month Contract with us.

More customers are using Paperless Billing.

More customers are going to pay with Electronic check.

**Relationship between Categorical Data and Churn**

***Demographic info***

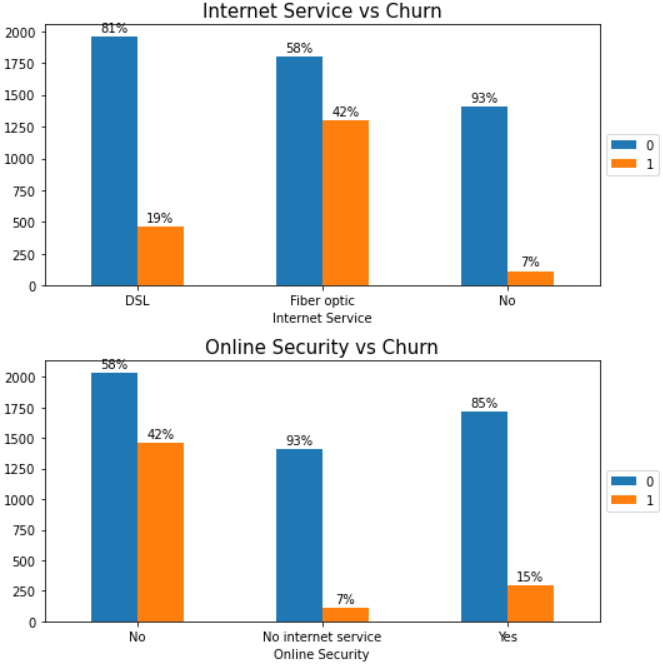
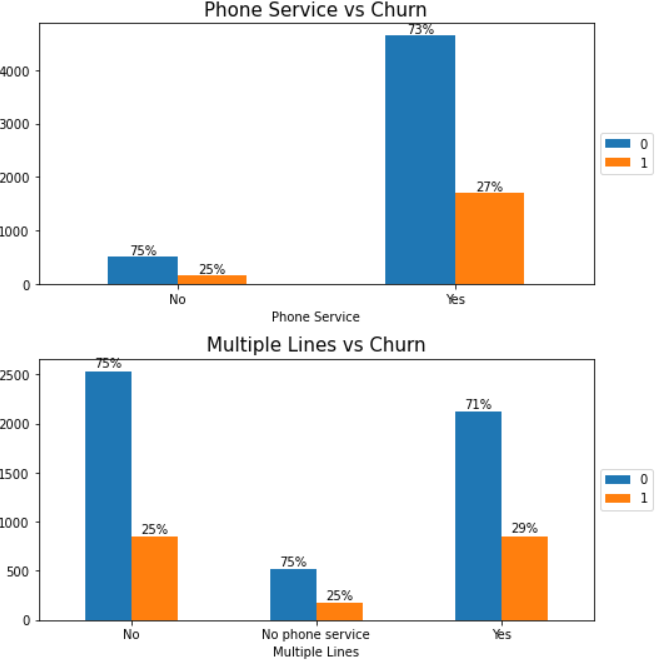


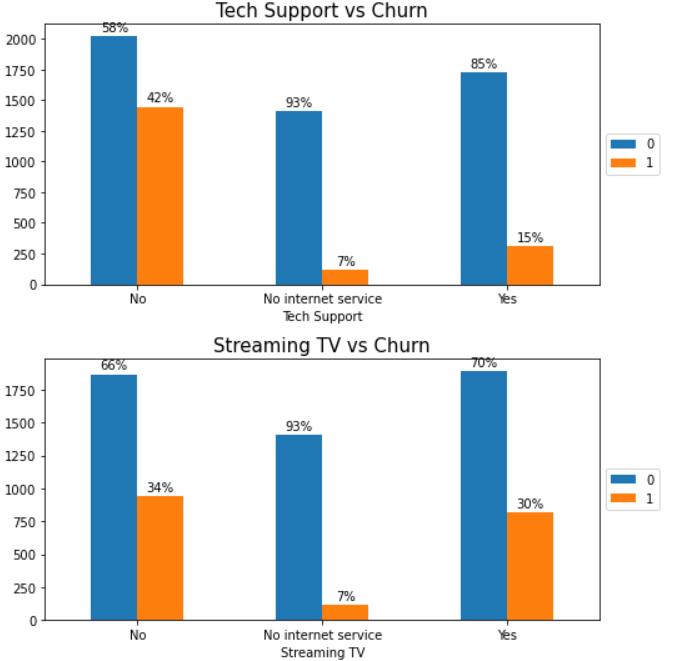
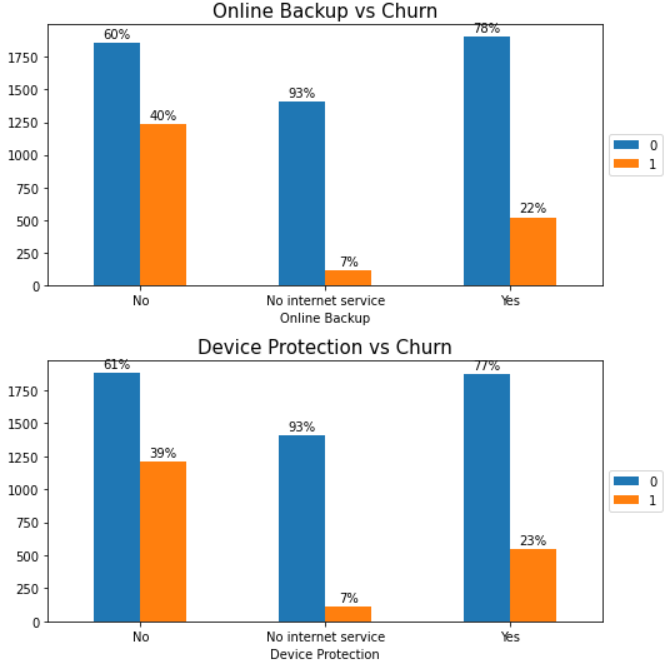
Gender is quite balanced in either churn or not. Thus Gender might not be a feature to affect the churn directly.

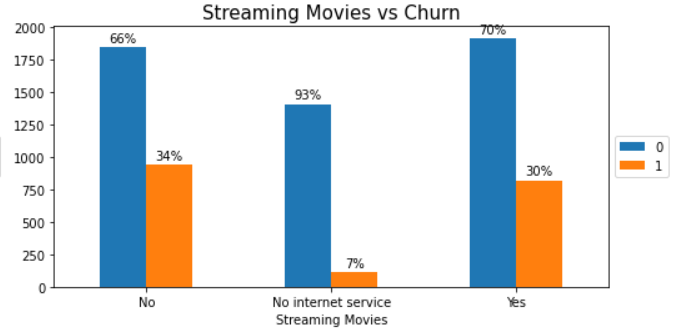
Senior Citizen have a much higher churn rate

Customers without Partner or Dependents are more likely to leave.

***Services***





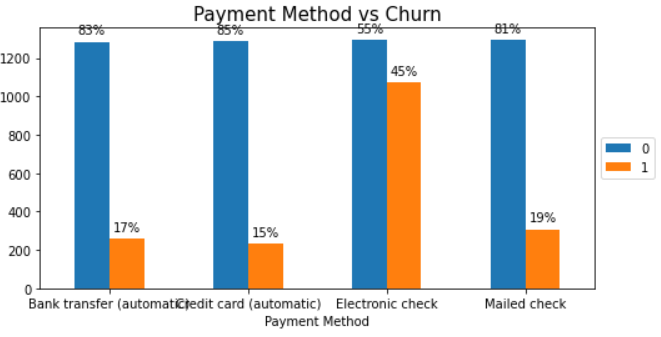
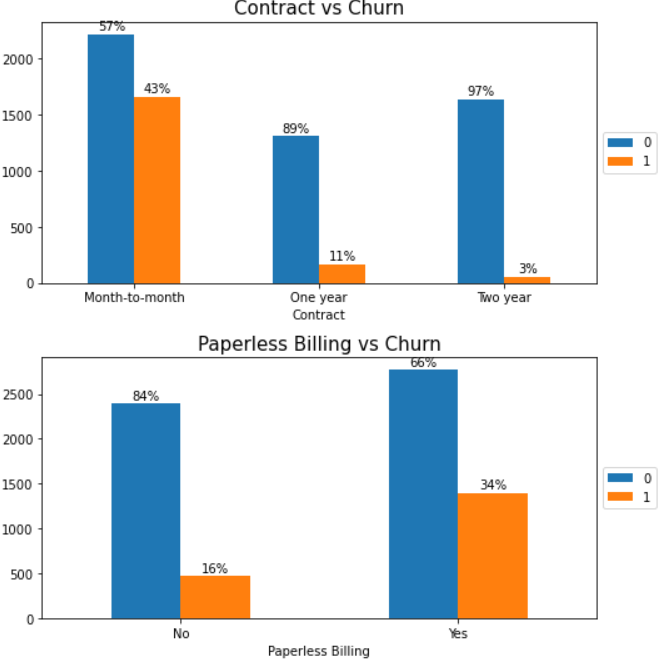


Customers with Fiber optic Internet Service are very likely to churn.

Customers who doesn't sign up for Online Security/Online Backup/Device Protection/Tech Support are very likely to churn.

Streaming TV and Streaming Movies are the most used service, but they're also the service with the highest customer churn rate.

***Customer account info***



Customers with Paperless Billing/using Electronic check Payment Method are very likely to churn.

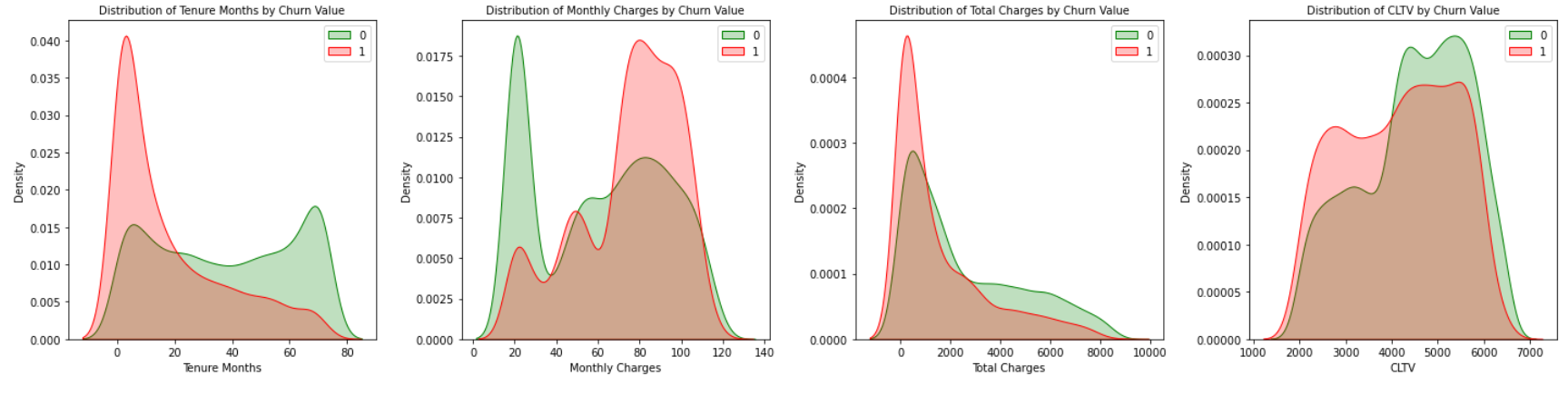
The shorter the Contract, the higher the customer churn rate.

**Research Q1: Which services should company improve to avoid current customer leaving?**

**Solution1**

* Since customers with Fiber optic Internet Service are very likely to churn, the company should improve the Fiber optic technology and provide more DLS Internet Service advertisement currently.
* Since Senior Citizen have a much higher churn rate and there's a lower amount of customers, the company could provide some bonus or discount to those senior citizen.
* The company could encourage customers to sign up for Online Security/Online Backup/Device Protection/Tech Support since those who don't sign up are very likely to churn.
* Since customers without economic dependents are more likely to churn, the company could provide payment by installments program suggestions to those customers.
* Since customers using Electronic check Payment Method are very likely to churn, the company could suggest to customers to use credit card automatic as a default payment method that has lowest churn rate.
* Since month to month contract has much higer churn rate, the company could has a strategy to encourage customer to subscribe with years such as using high discount or adding bonus package.

**Relationship between Numerical Data and Churn**



**Conclusion:**

* *Tenure Months*

Peak - Churn: around 5 months.

There is a trend that the larger the tenure, the smaller the churn rate.

* *Monthly Charges*

Peak - Churn: around 80 dollars ; Not Churn: around 20 dollars.

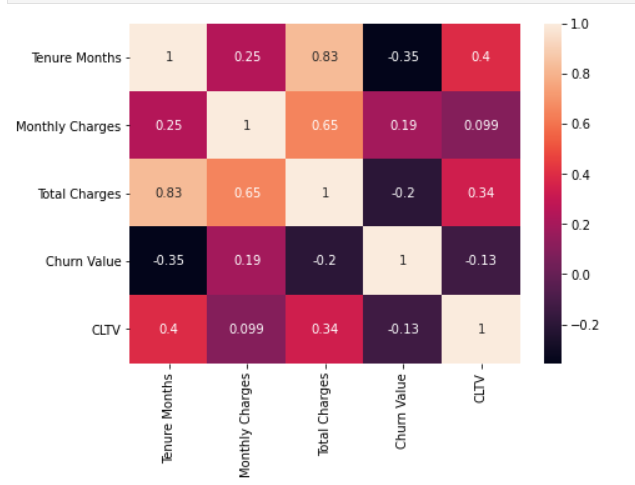
The churn rate and monthly charges are not clearly related to each other.

* *Total Charges*

There is a trend that the larger the Total Charges, the smaller the churn rate.

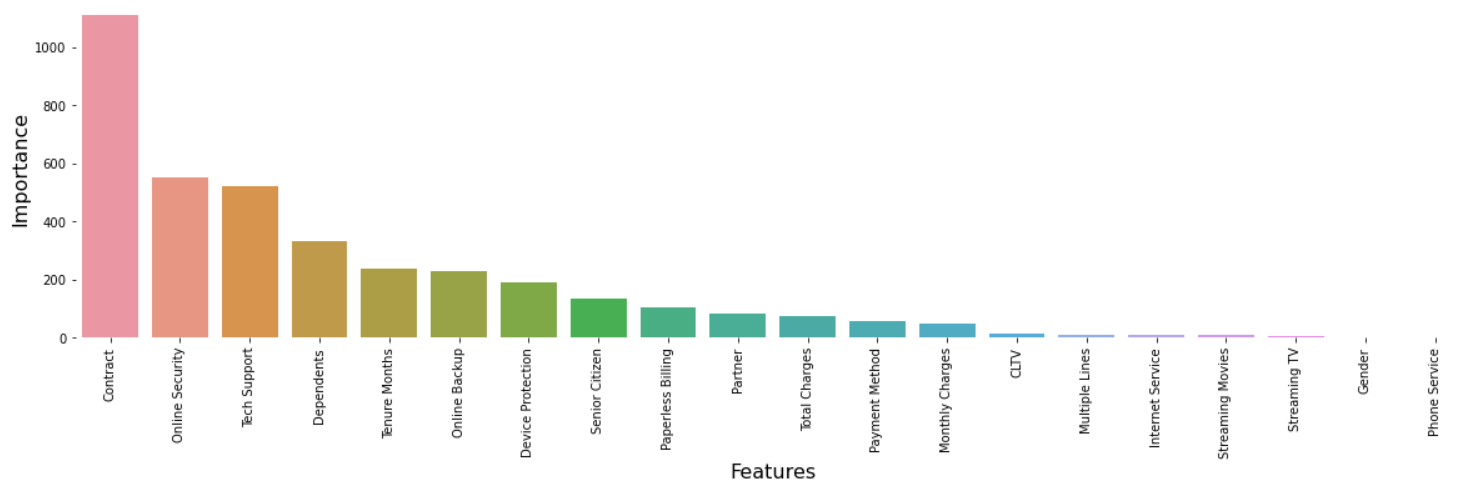
* *CLTV*

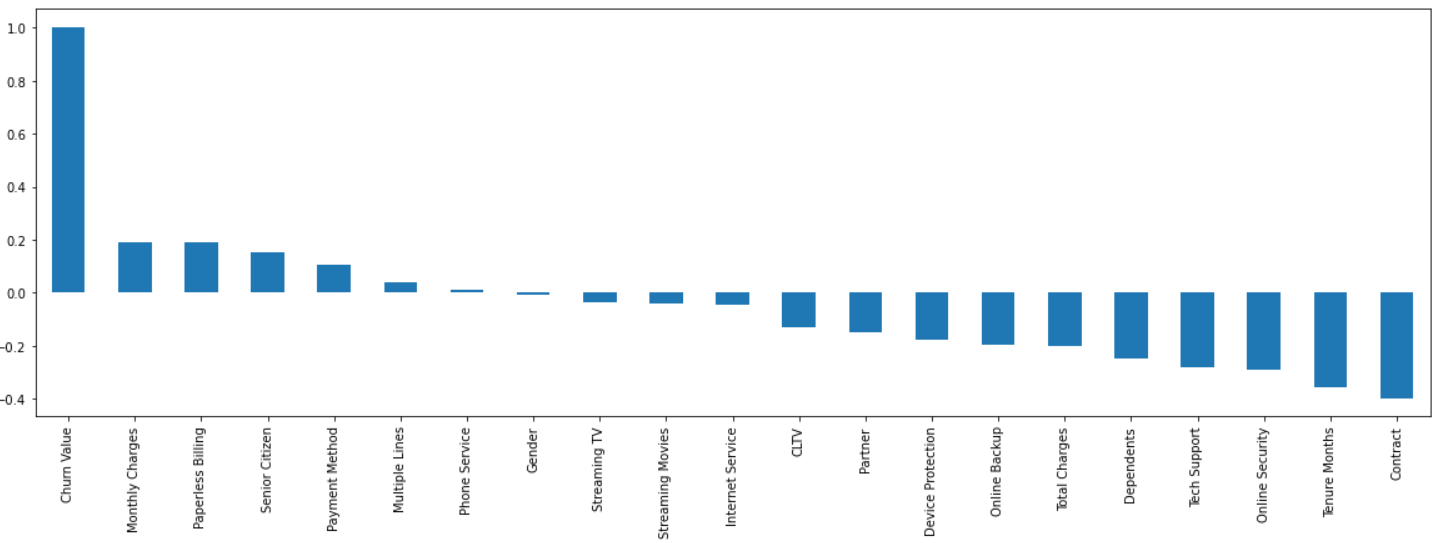
The churn rate and monthly charges are not clearly related to each other.



The correlation heat map shows Total Charges is correlated to Tenure Months and Monthly Charges but doesn't show these numerical columns have high correlation with Churn value

## **Research Q2: What kind of services are more related to whether a customer leave?**





After the exploration, I did a label encoding method to catecorical data and min max scaler to numerical data. And then apply the Chi-square testing and correlation between features and Churn for selecting importance features. The highly important score and high correlation features show Online security, Tenure Month, Dependent and Tech support are the most related features. The company should pay more attention on those three features first and come up with related strategy to reduce the Churn rate.

1. **Model**

In this part, I used three method:

* ***Single Model for all features with Default Parameters***

I selected LogisticRegression, RandomForest, Naive Bayes, and KNeighbours method with 10 fold cross validation to do the prediction and got corresponding results as below. The default parameters model shows a good result on test and train accuracy but has a bit high standard error on train accuracy. Among those models, logistic regression has a best performance.

* ***Single Model for all features with Tuning Parameters***

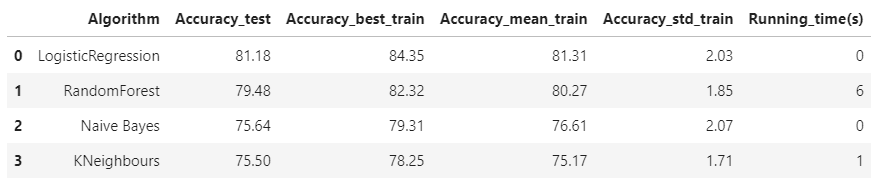
I selected LogisticRegression, RandomForest, Naive Bayes, and KNeighbours method with different parameters into pipeline with min max scaler as the first step and used grid search cross validation with 10 fold to fit the training data and then get the corresponding results.

* ***Hybrid Model combined PCA, Logistic Regression method with Tuning Parameters***

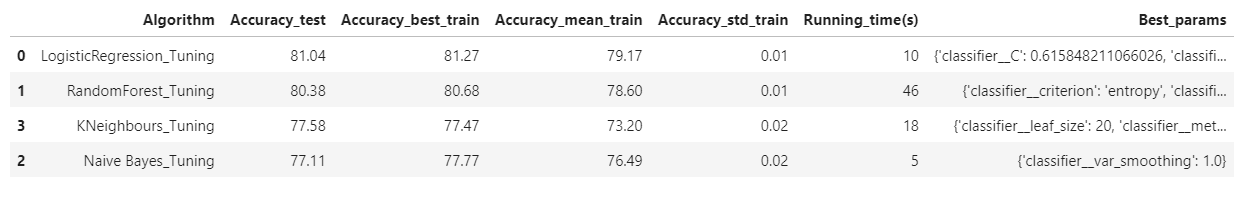
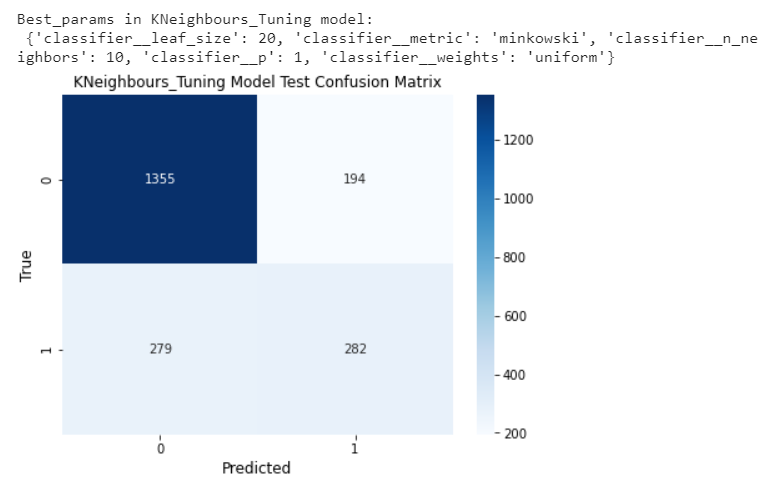
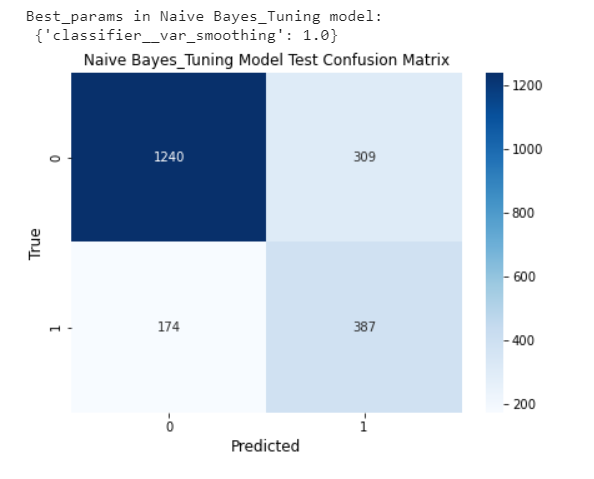
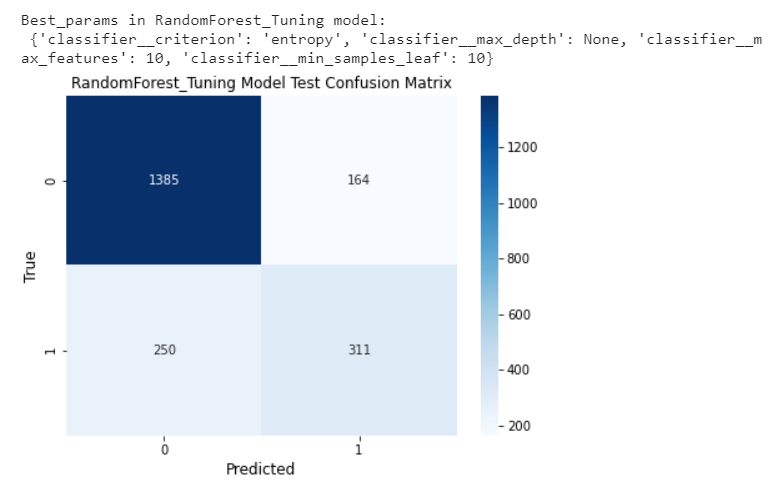
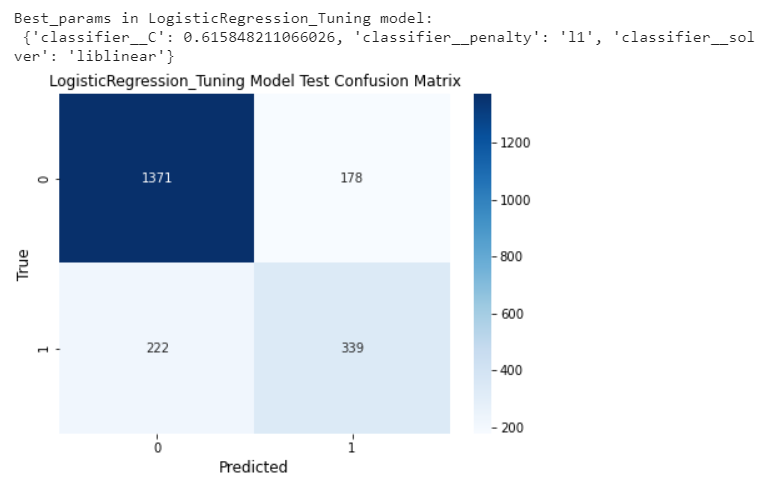
I used pca method to plot the data with different components and get 80% variance ratio which has 11 components and then apply the logistic regression method with tuning parameters in the pipline with standard scaler as the first step.

1. **Experiments and Evaluation**

* ***1. Single Model for all features with Default Parameters***

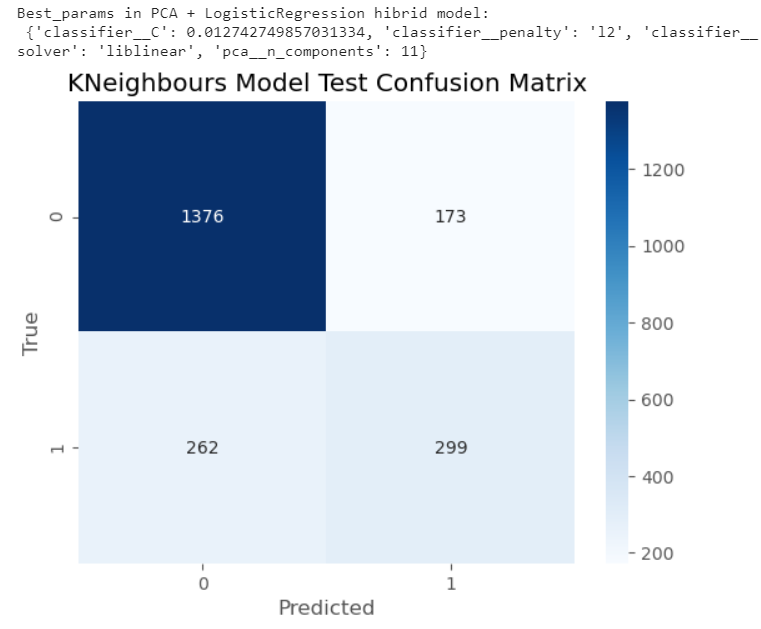
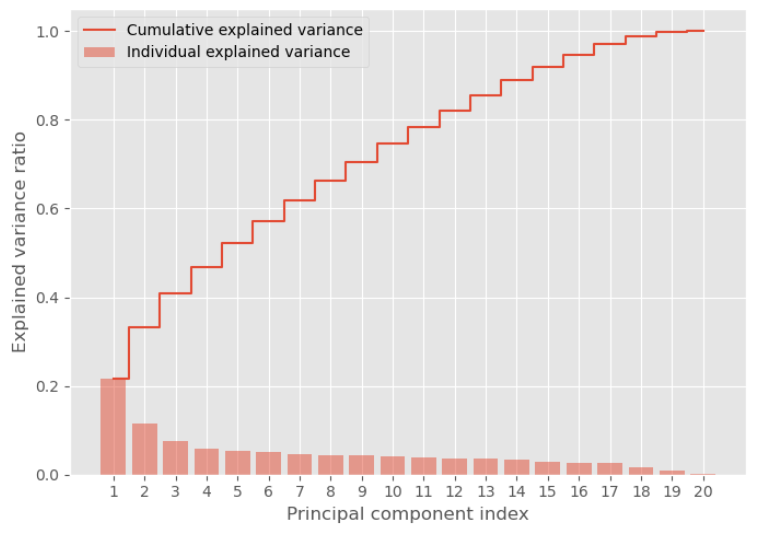
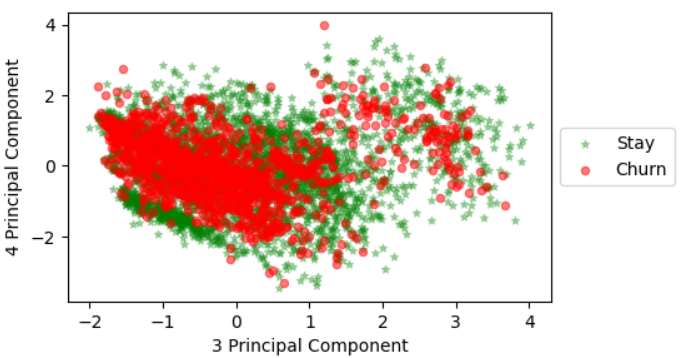
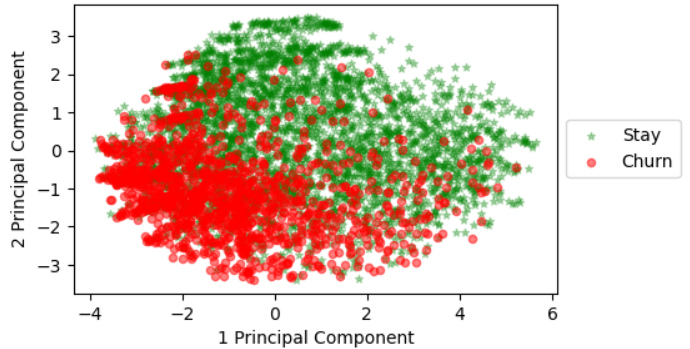


* ***Single Model for all features with Tuning Parameters***

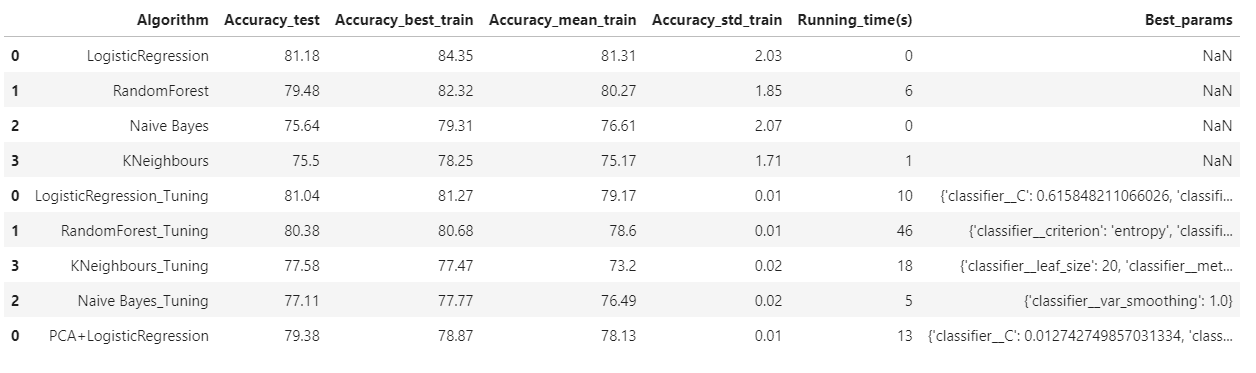


Comparing to the default parameters, tuning parameters to train the data can improve the accuracy score specially for Random Forest and KNeighbours and Naive Bayes in both train and test data

* ***Hybrid Model combined PCA, Logistic Regression method with Tuning Parameters***



Comparing all method results:



**Research Q3: Who are more likely to leave and what the confidence of that predictions are?**

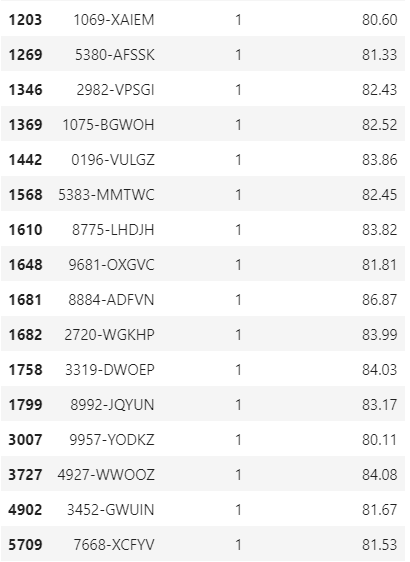
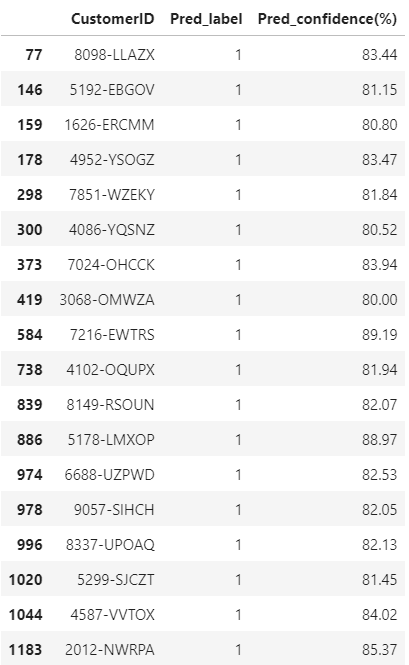
Based on the performance of all models, I'll select the Logistic Regression\_Tuning as the best model to predict the current customers see who are more likely to leave.

Here we suppose current customer dataset are X, and we already split it to train and test data.

There are 2110 customers currently in the test dataset.

There are 517 (24%) customers might be churn,

34 customers among of them have more than 80% confidence that more likely to churn.



1. **Conclusion**

This project has solved the Research Questions about:

* Research Q1: Which services should company improve to avoid current customer leaving?
* Research Q2: What kind of services are more related to whether a customer leave?
* Research Q3: Who are more likely to leave and what the confidence of that predictions are?

Based on the solutions of those questions, the company are now able to make some relevant strategies to improve their services and knowing who are more likely to leave now thus this research can help company reduce the churn rate of their customers to increase the profit.

Reference:

* **<https://www.sciencedirect.com/science/article/abs/pii/S0148296321000308>**
* **<https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs190677>**
* **<https://stackoverflow.com/questions/31749448/how-to-add-percentages-on-top-of-bars-in-seaborn>**
* **https://stackoverflow.com/questions/45969390/difference-between-stratifiedkfold-and-stratif**
* **iedshufflesplit-in-sklearn**
* **<https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee>**
* **<https://joannaoyzl.github.io/2019/07/28/Unsupervised-Learning-in-Python/>**