Vinh Vu

Telco customer analytics

Springboard

12/17/2018

I. Business problem:

Through this project, I want to identify the root causes for customer churn and predict it by machine learning.

II. Client:

My client, Telco Company, is a telephone and internet service provider with over 5000 customers. In order to grow and maintain profitability, it's essential that they learn how to maintain a dedicated customer base and reduce churn. Based on my analysis, Telco can:

1. Identify customers that are likely to churn and reach out to them to try to stop them from churning via special offers targeted to their needs
2. Focus marketing on customers that are more likely to be long term customers
3. Modify their services to improve the likelihood customers will stay longer term

III. Data collection and wrangling summary:

1. Obtain data :

I am going to use Telco Customer Churn dataset which is provided by Kaggle. I will use the panda library to import csv file.

Attribute	Description
CustomerID	Customer ID
Gender	Customer gender (female, male)
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
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
Churn	Whether the customer churned or not (Yes or No)

2. Data Wrangling·

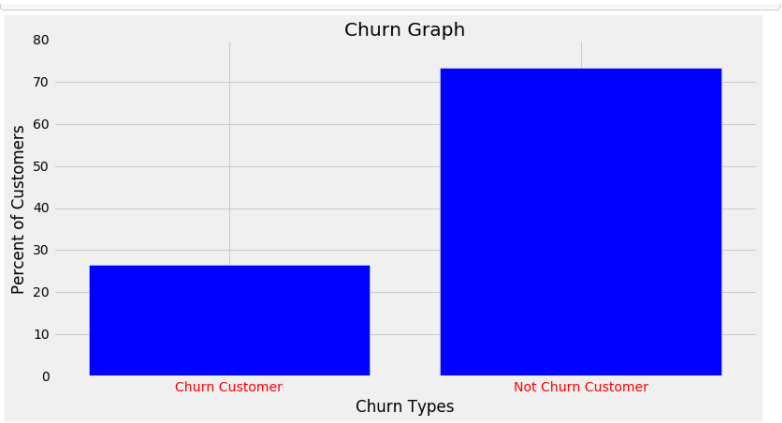
This dataset has no missing values.

3. Constructing the Data:

- Convert columns with yes/ no to 1/0
- When a column has multiple values, I use get_dummy turning a column to multiple columns

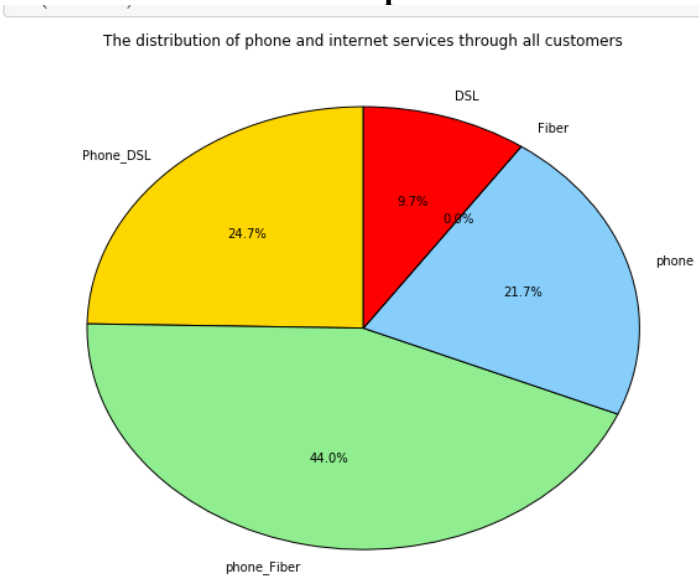
IV. Exploring the Data (Data storytelling)

1. The percentage of churn customers in total customers:



The dataset is imbalanced, the percentage of churn customer is 28.5% total customer, and the percentage of non-churn customer is 71.5% total customer.

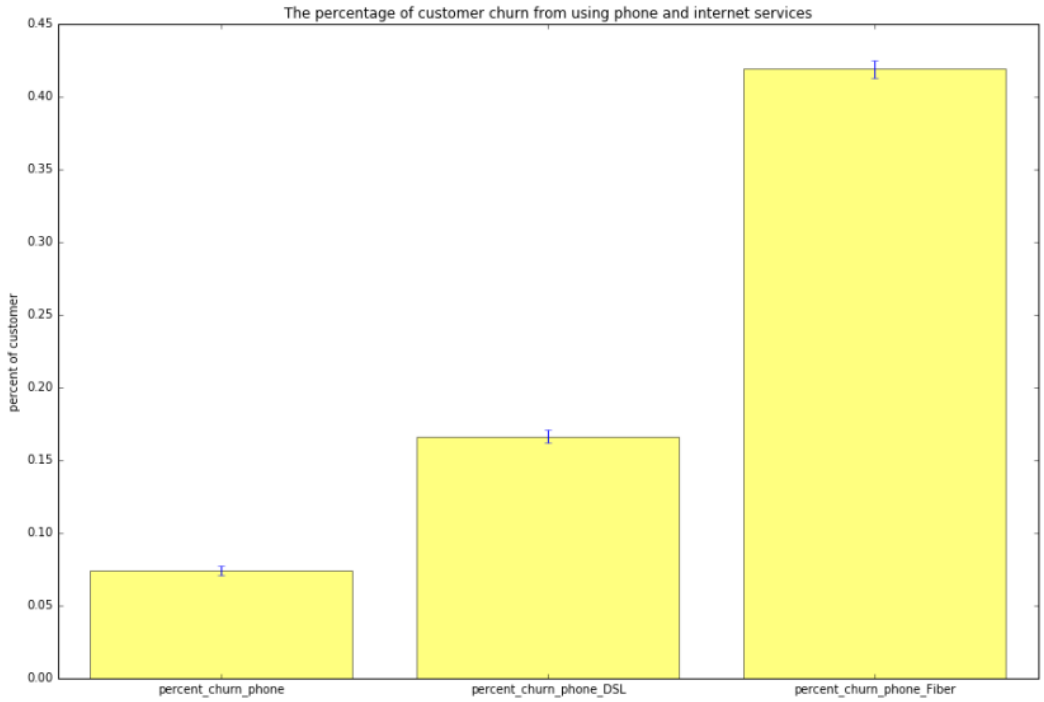
2. The distribution of phone and internet service through all customers¶



Customers who buy both phone and Fiber packages make up the majority with 44% of total customers. The percentage of customer using phone and DSL, is the second place in chart with 24.7% of total customers. The percentage of customer from using phone only are 21.7 % of total customers. The percentage of DSl only is 9.7% of total customers and fiber is <0.01%.

Takeaway: the customers who use both phone and fiber are more likely to churn than other service groups.

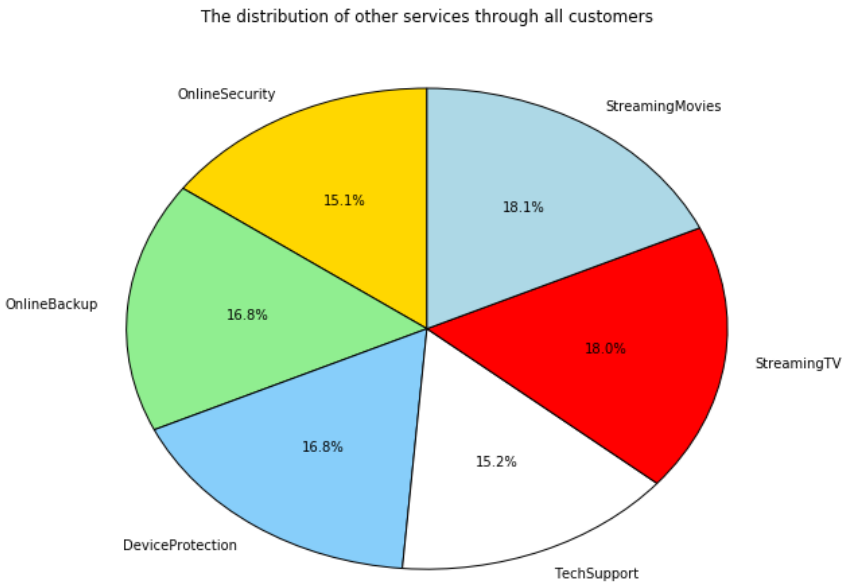
3. The percentage of customer churn from using phone and internet services:



The bar chart shows the percentage of customer churn in three distinct service groups of phone and DSL, phone and Fiber, and phone only. The percentage of customer churn from using phone and fiber is majority with 41% of total customers. The percentage of customer churn from using phone and DSL, is the second place in chart with 16% of total customers. Finally, the percentage of customer churn from using phone only are 7 % of total customers.

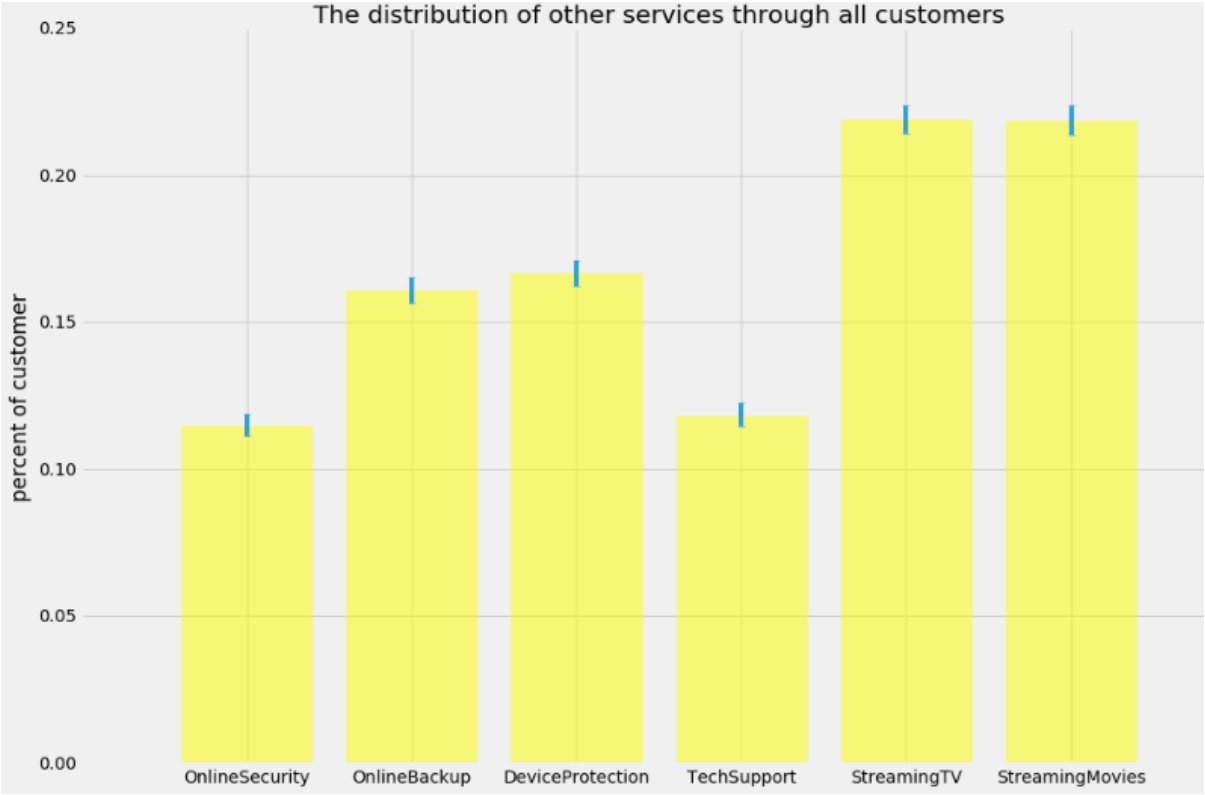
Takeaway: the group of customers who use both phone and fiber are significantly more likely to churn than other service groups.

4. The distribution of other services through all customers



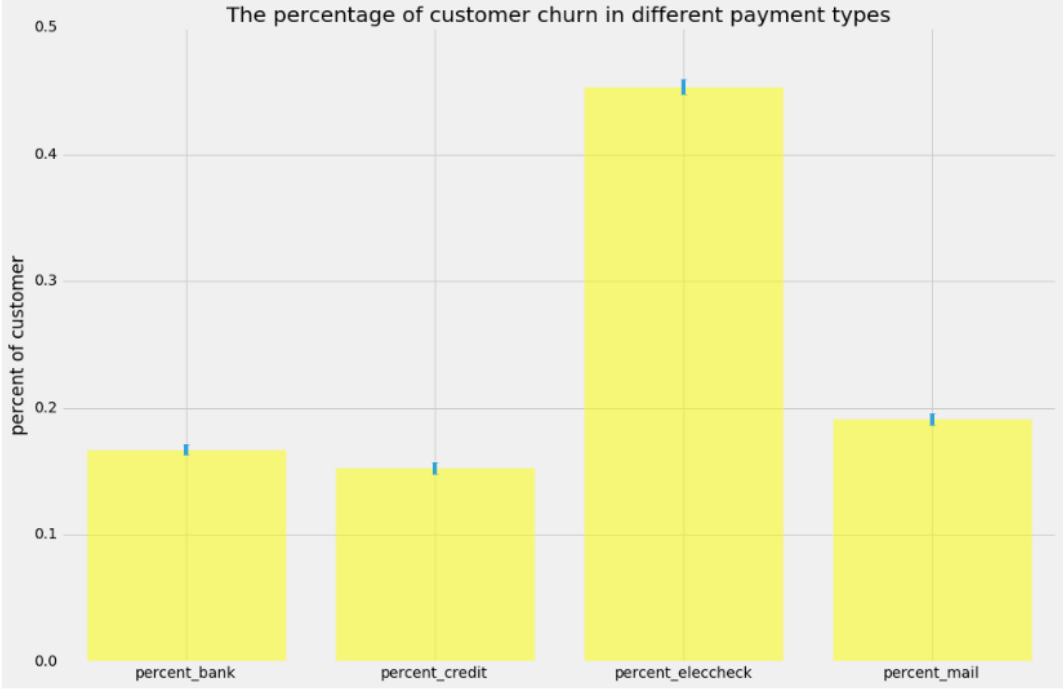
From The pie chart, the percentage of customer from using Stream Tv and Stream Movies are majority 18.1% and 18% of total customers. The percentage of customer from using Device Protection and Online Backup are the second place, 16.8% of total customers. The last two are Online Security and Tech Support, 15.1% and 15.2% of total customers.

5. The percentage of customer churn from other services



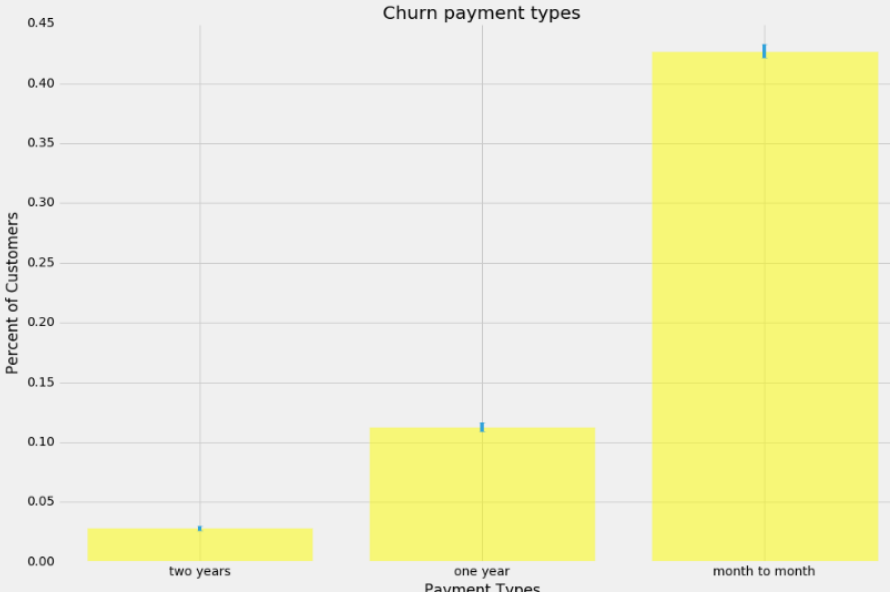
From The bar chart, the percentage of customer churn from using Stream Tv and Stream Movies are majority, 21% of total customers. The percentage of customer churn from using Device Protection is the second place, 16.6% of total customers. The percentage of customer churn from using Online Backup is 16% of total customers. The last two are Online Security and Tech Support, 11.5% and 11.8% of total customers.

6. The percentage of customer churn in different payment types



The bar chart illustrates the percentage of customer churn in four different payment types: Credit card (automatic), Bank transfer (automatic), Mailed check, Electronic check. The Electronic check is the most popular payment type with 42% of total customers. The second highest payment type is Mailed check, 19% of total customers. The Credit card (automatic) and Bank transfer (automatic) are mostly the same with 16% and 15% customers.

7. The percentage of churn customers in different payment plans.

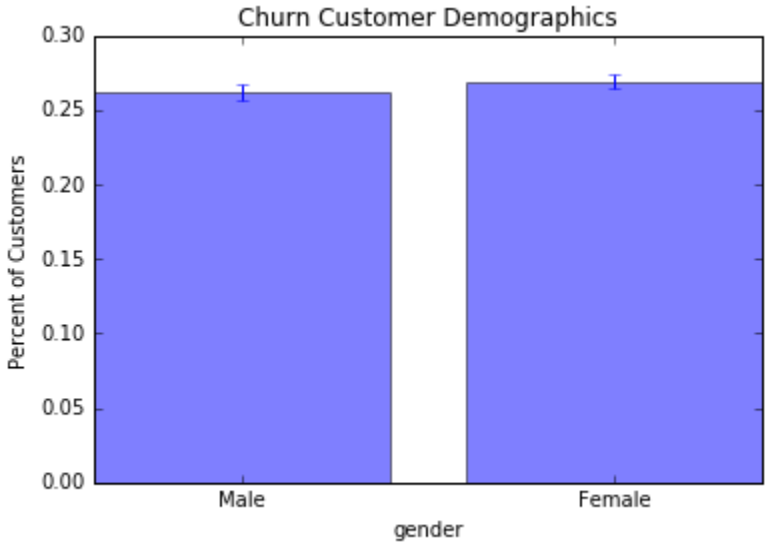


The bar chart illustrates the percentage of churn customers in three different payment plans: month to month, one year and two years. The month to month plan has the highest percent of churn customer, 42% compare to the others. The one year plan is 2% and 11% for two years.

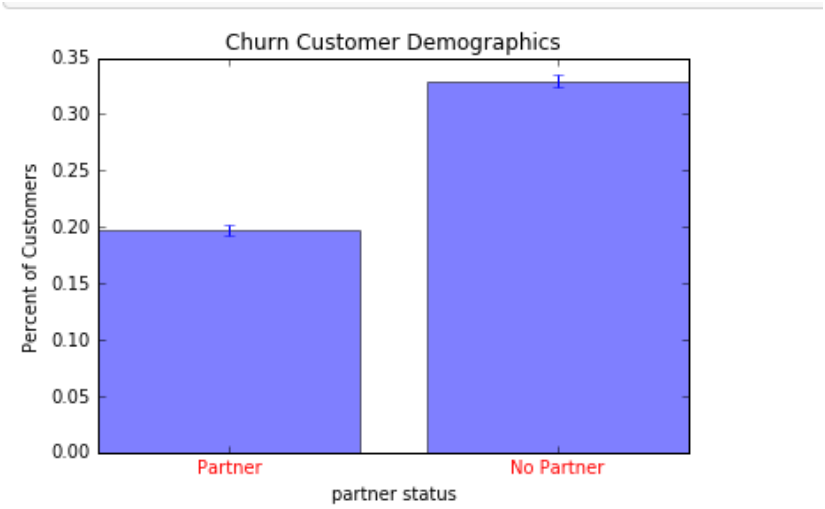
Take away: customers are likely to churn when they sign up for month to month payment plan.

8. The percentage of customer churn in Demographics (gender/partner/dependents)

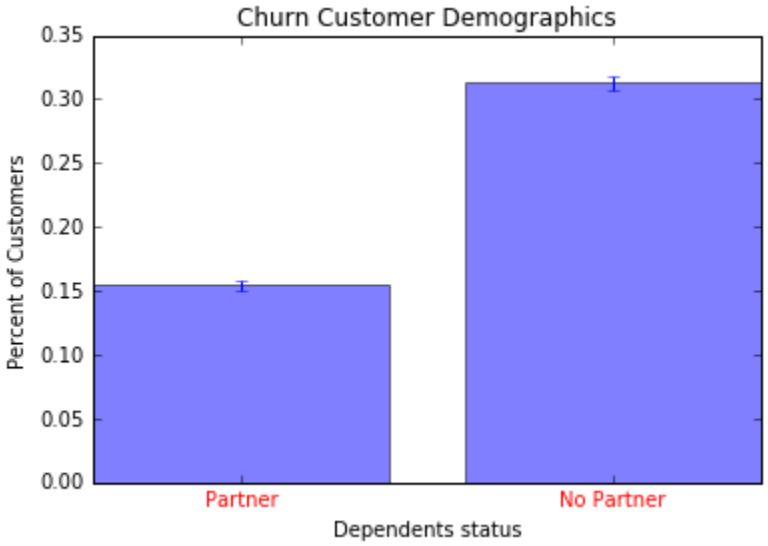
Gender



Partner



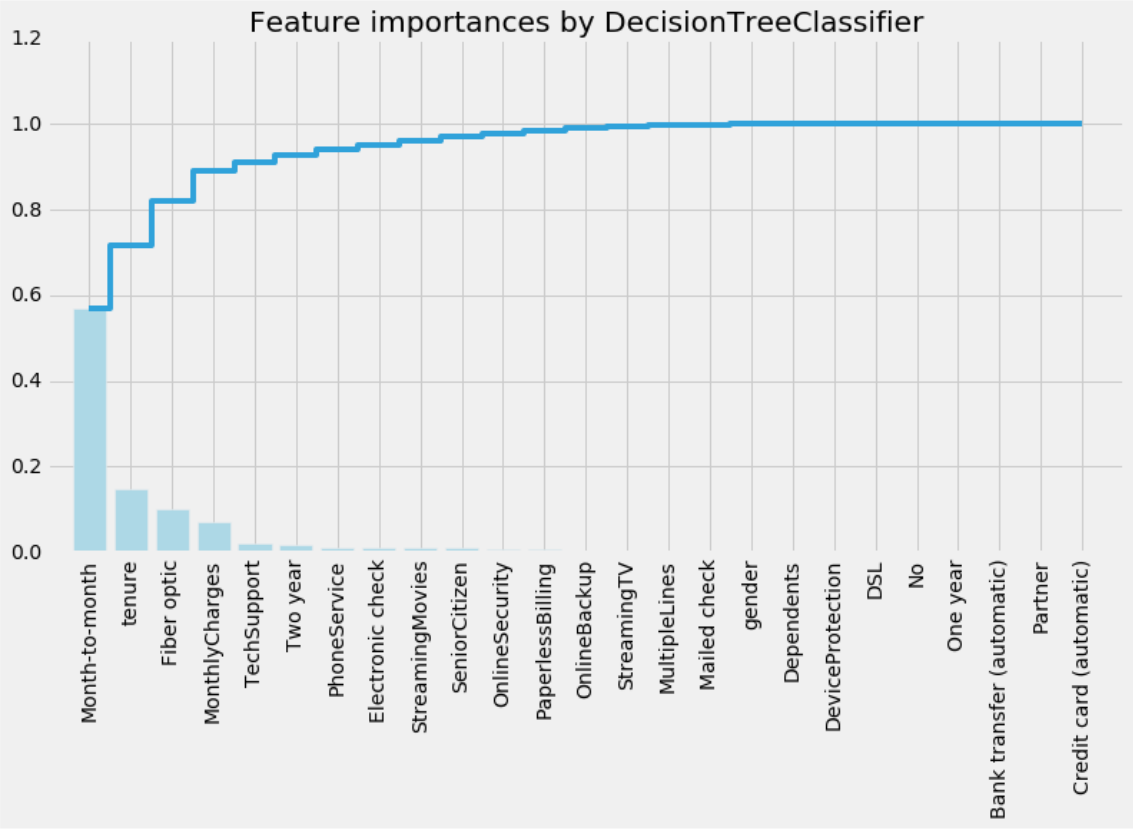
Dependents



The bar chart shows the percent of churn customers compares in three different demographics' categories gender, partner, and dependent. In gender category, the percent of customer churn in female and male are mostly the same, male is 26.9% of total customers and female is 26.1% of total customers. In partner category, the percent of churn customer with partner is 19 % total customers, and the percent with no partner is 32% of total customers. In dependent category, it is mostly the same with partner category. The percent of no dependent is higher than dependents, the no dependent is 31%of total customers, and the dependent is 15% of total customers.

Take away: customers has no partner and no dependent is highly percent to churn than other.

9. Features importance:



Feature Importance

1. Month to month
2. Tenure
3. Fiber optic

V. Inferential Statistics

1. T-TEST

The T-Test is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups.

One sample test:

The One Sample T-Test determines whether the sample mean is statistically different from a known or hypothesized population mean.

· Null Hypothesis: there is no difference in tenure between churn customers and the customer population.

· Alternate Hypothesis: there is a significant difference in tenure between churn customers and the customer population.

At 95% confidence level to test the hypothesis. Stats.ttest_1samp () function is used to conduct one sample T-Test.

Ttest_1sampResult(statistic=-31.856572712421674, p value=3.0614037111362083e-178)

The result the t value is -31.85, and p value < 0.05. In order to proof that null hypothesis is rejected surely, we calculate degrees of freedom whether the t-statistic is outside of the quantiles of the t-distribution.
The t-distribution left quartile range is: -1.96123406594
The t-distribution right quartile range is: 1.96123406594
(17.093332889308954, 18.86493356333952)
The t-statistic is outside of the quantiles of the t-distribution and p value is less than 0.05. Therefore null hypothesis is rejected.

2. Two Sample test:
---Gender:

Null hypothesis: there is no difference between male and female in customer churn.

Hypothesis: there is a difference between male and female in customer churn.
Ttest_indResult(statistic=-0.72261049878576156, pvalue=0.46994323541735661)

P value<0.05 , so null hypothesis is rejected
---Dependent status:

Null hypothesis: there is no difference between dependent and no dependent in customer churn.

Hypothesis: there is a difference between dependent and no dependent in customer churn.
Ttest_indResult(statistic=-15.409078802902004, pvalue=2.1775286391572522e-52)

P value<0.05 , so null hypothesis is rejected
---Partner status:

Null hypothesis: there is no difference between partner and no partner in customer churn.

Hypothesis: there is a difference between partner and no partner in customer churn.
Ttest_indResult(statistic=-12.841725043203832, pvalue=2.529114349220257e-37)

P value<0.05 , so null hypothesis is rejected
---senior discount:

Null hypothesis: there is no difference between senior and no senior in customer churn.

Hypothesis: there is a difference between senior and no senior in customer churn.
Ttest_indResult(statistic=11.580732091336619, pvalue=9.3643915616853527e-30)

P value<0.05, so null hypothesis is rejected

Summary:

Based on the statistical analysis, statistical significance and practical significance are significantly difference. The correlation coefficient for the tenure and churn is -0.35, which shows that the tenure and the churn has strong inverse correlation. Moreover, p value is less than 0.05, there shows significant correlation coefficient between the tenure and the churn.

Demographic factors are also important to customer churn, because of the correlation between the various demographic factors and Churn. In gender, there is a difference between male and female group. In partner status, there is a difference between partner and no partner, no partner has higher percentage churn customers than partner. In dependent status, it is similar with partner, there is also a significant difference between no dependent and dependent. Finally, the senior and non-senior also are different in the percentage of churn customers.

VI. Results and In-depth analysis using machine learning

For this project, we use four model such as Decision Tree Model, AdaBoost Model, Logistic Regression Model, and Random Forest Model

- Base rate generally refers to the (base) class probabilities unconditioned on evidence, frequently also known as prior probability. A Base Rate Model is a model that always selects the majority class which compares to other models. In this project, we compare Churn =0 where are customers who don't churn.

- In churn vs not churn customer graph, the percentage of churn customer is 28.5% total customer, and the percentage of non-churn customer is 71.5% total customer. The base rate model would predict every non churn customers and ignore churn customers.

- Example: The base rate accuracy for this data set, when classifying everything as 0's, would be 71.5% because 71.5% of the dataset are labeled as 0's (not churn customer).

Class Imbalance

Due to skewed distribution of customers who churn and don't churn, the data set is an imbalanced problem. Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. In order to face imbalanced class, we measure model performance through different errors and concepts; such as Precision, recall. Due to this project, it is a binary class problem. Precision and recall is very important, because Precision is sort of like accuracy but it looks only at the data you predicted positive (in this example you're

only looking at data where you predict a churn) and the recall is also sort of like accuracy but it looks only at the data that is “relevant” in some way.

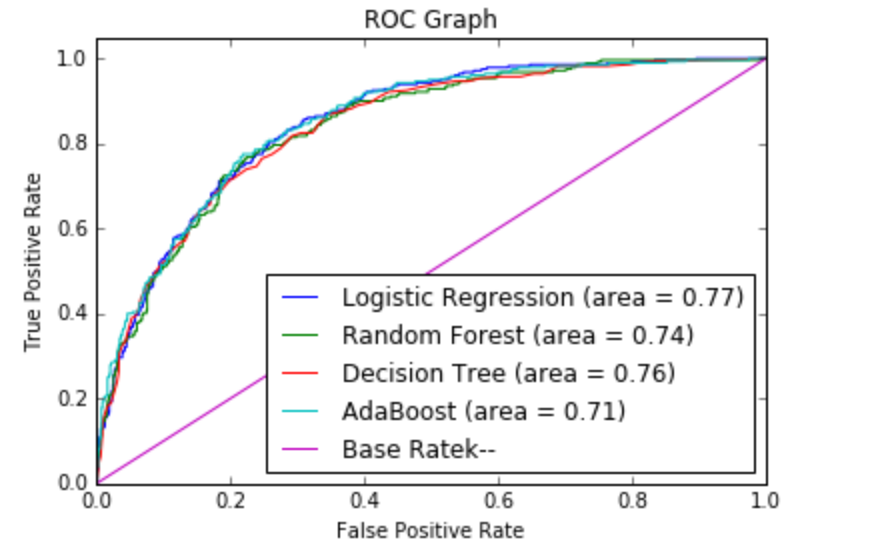
Due this project is classification problem, I will use grid search in order to find the best parameter. Then it is used other models like logistic regression, random forest and decision tree in order to find the best fit model.

Different Ways to Evaluate Classification Models:

```
---Base Model---
Base Rate AUC = 0.50
              precision    recall  f1-score   support

           0       0.74      1.00      0.85       777
           1       0.00      0.00      0.00       280

   micro avg       0.74      0.74      0.74      1057
   macro avg       0.37      0.50      0.42      1057
weighted avg       0.54      0.74      0.62      1057



 ---Logistic Model---
Logistic AUC = 0.77
              precision    recall  f1-score   support

           0       0.91      0.75      0.82       777
           1       0.53      0.80      0.64       280

   micro avg       0.76      0.76      0.76      1057
   macro avg       0.72      0.77      0.73      1057
weighted avg       0.81      0.76      0.77      1057



 ---Decision Tree Model---
Decision Tree AUC = 0.76
              precision    recall  f1-score   support

           0       0.91      0.72      0.80       777
           1       0.51      0.79      0.62       280

   micro avg       0.74      0.74      0.74      1057
   macro avg       0.71      0.76      0.71      1057
weighted avg       0.80      0.74      0.76      1057
```

```
 ---Random Forest Model---
Random Forest AUC = 0.74
              precision    recall  f1-score   support

           0       0.87      0.82      0.84       777
           1       0.57      0.66      0.61       280

   micro avg       0.78      0.78      0.78      1057
   macro avg       0.72      0.74      0.73      1057
weighted avg       0.79      0.78      0.78      1057
```

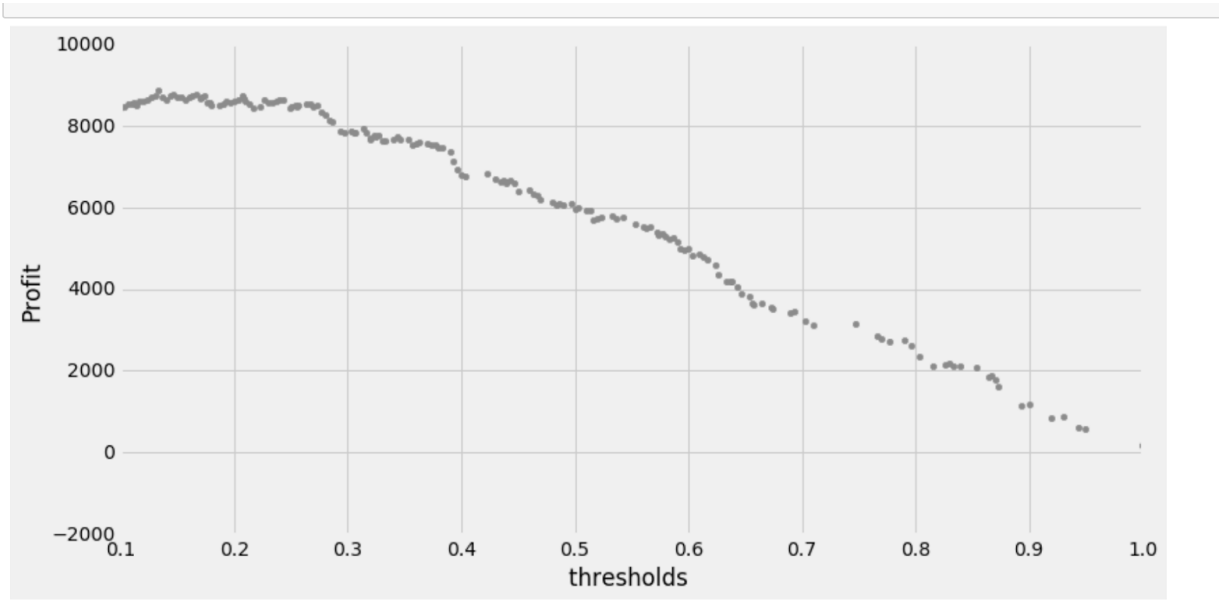
```
 ---AdaBoost Model---
AdaBoost AUC = 0.71
              precision    recall  f1-score   support

           0       0.84      0.90      0.87       777
           1       0.64      0.52      0.57       280

   micro avg       0.80      0.80      0.80      1057
   macro avg       0.74      0.71      0.72      1057
weighted avg       0.79      0.80      0.79      1057
```



According to our ROC graph, our Logistic Regression model is the best predictor for this problem with AUC =0.77.

Choosing a Threshold: Profitability Curve

In order to intelligently choose a threshold for this scenario it would be important to do a profitability analysis. There are 3 main variables that would affect this threshold: cost of service offered to retain customers (ie labor/discount of service etc.), customer lifetime value, probability that customers are retained by service offer. Because of this, the threshold will change for each given combination of these variables. However, to offer an illustration of this, I will proceed here with an example in which the service offered including labor is $20 and the customer lifetime value is $100 and the percentage chance of retaining customers with this offer is 100%.



In the specific example above, the optimal threshold to maximize profitability would be 0.193333. However, as noted above, this threshold would vary based on service and customer and we could use a similar analysis to determine it.

VII. Conclusion:

In summary, this is what we know about why customers churn:

1. Customer are most likely to churn if they are the month to month payment type.
2. Customer is likely to churn in first ten months.
3. Customers is likely to leave when they use Fiber optic service.
4. There is 0% customers use fiber only, customers are likely to cancel services when they use phone and fiber.
5. Customers with no partner or dependent are more likely to churn

6. Month to month, tenure, and fiber optic are the three most significant features in determining churn.

Using the Logistic Regression model built to classify churn, wee can determine the probability each customer will churn. Telco can use this knowledge both to offer an array of products and services more likely to retain customers, and also use this knowledge to allocate resources to provide services specifically aimed at customers likely to churn so they won't leave. One primary weakness in this analysis, is that we cannot predict the exact timeframe in which a customer would leave, a next step in this analysis would be to take this into account using a time series based approach.