# Telco customer analytics

# Business problem:

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

# 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

# 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) |

1. **Data Wrangling·**

This dataset has no missing values.

customer.isnull().any()

Out[55]:

customerID False

gender False

SeniorCitizen False

Partner False

Dependents False

tenure False

PhoneService False

MultipleLines False

InternetService False

OnlineSecurity False

OnlineBackup False

DeviceProtection False

TechSupport False

StreamingTV False

StreamingMovies False

Contract False

PaperlessBilling False

PaymentMethod False

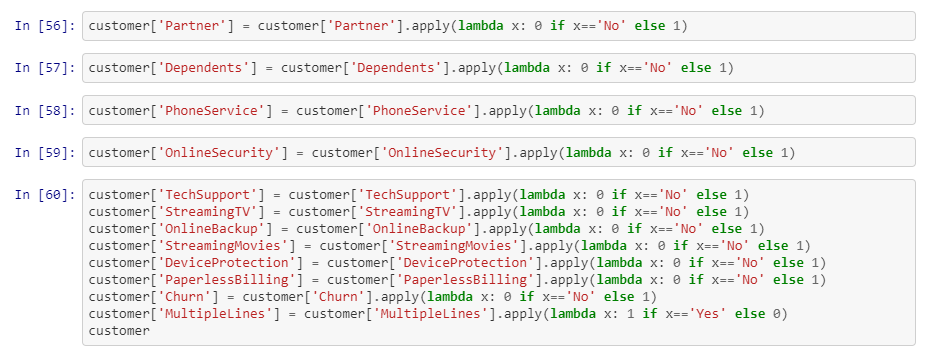
MonthlyCharges False

TotalCharges False

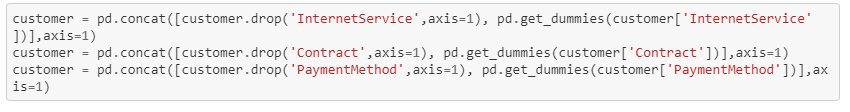
Churn False

dtype: bool

1. **Constructing the Data:**
   * Convert columns with yes/ no to 0/1



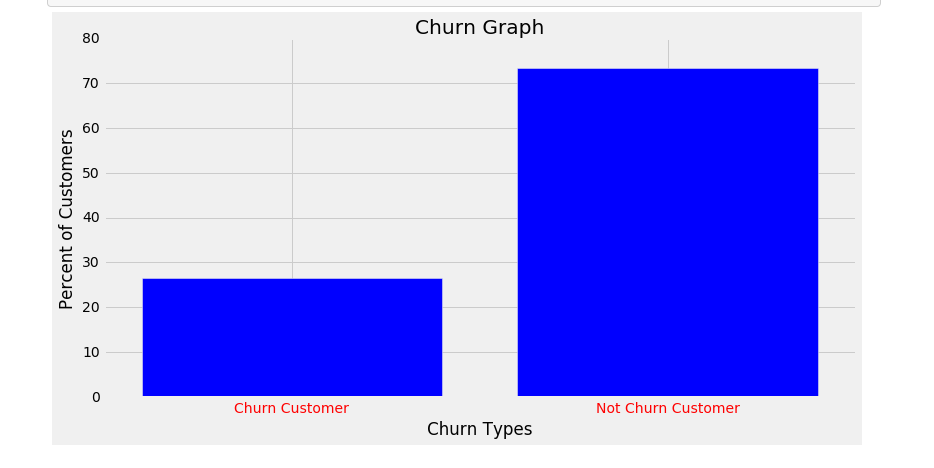
* + When a column has multiple values, I use get\_dummy turning a column to multiple columns



**IV.** **Exploring the Data (Data storytelling)**

1. **The percentage of churn customers in total customers:**

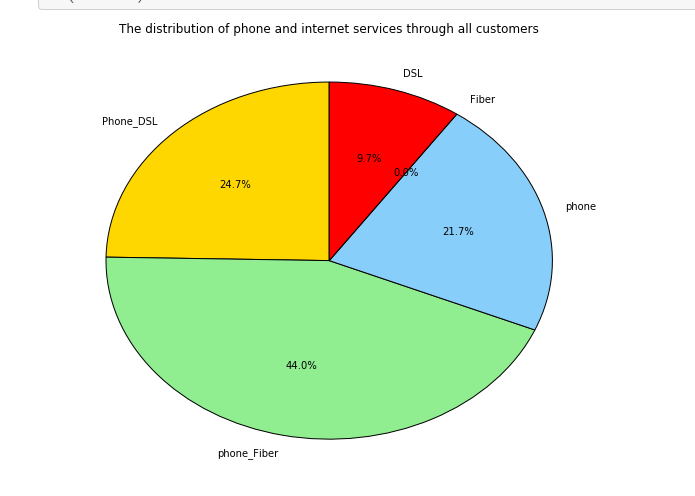
This graph shows the result of report in which customers are churn or not churn. From the graph, it is clear that the majority of customers are not churn with just 43 % different to churn customers.



1. **The distribution of phone and internet service through all customers**[**¶**](https://render.githubusercontent.com/view/ipynb?commit=b910163ec2157d3b3eb98dd8c6e0f79369b3f47b&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f76696e687076752f63617073746f6e652d312f623931303136336563323135376433623365623938646438633665306637393336396233663437622f556e7469746c65642e6970796e62&nwo=vinhpvu%2Fcapstone-1&path=Untitled.ipynb&repository_id=145277473&repository_type=Repository#The-distribution-of-phone-and-internet-service-through-all-customers)

From the pie chart, the percentage of Fiber and phone is majority with 44% of total customers. The percentage of customer from 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% total customer and 0% of fiber only.

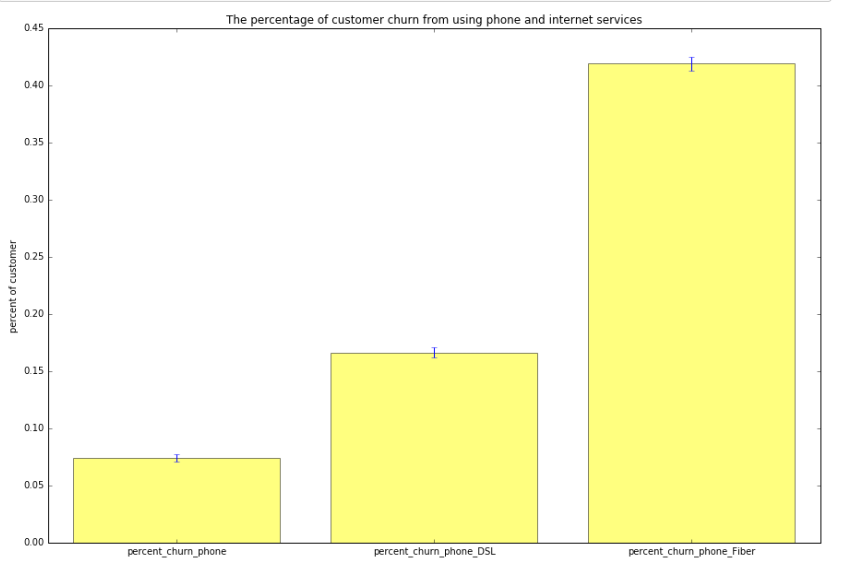
**Hypothesis:** the customers use phone and fiber are likely to churn than other service groups.



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

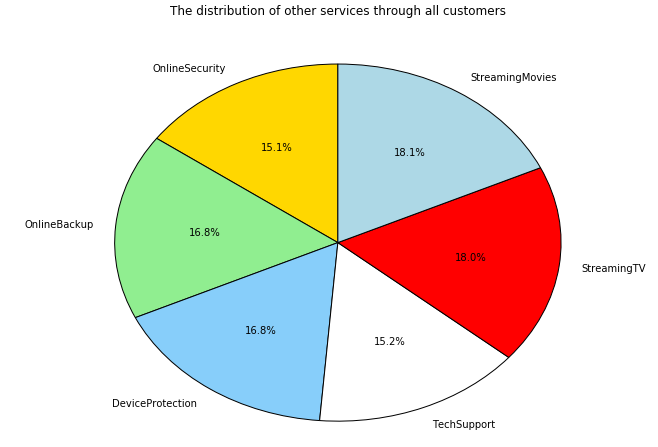
The bar chart shows the percentage of customer churn in three distinct service groups of phone and DSL, phone and Fiber, and phone only. The percentage of customer churn from using phone and fiber is majority with 41% of total customers. The percentage of customer churn from using phone and DSL, is the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

**Hypothesis:** it is clearly that the customers use phone and fiber are likely to churn than other service groups.



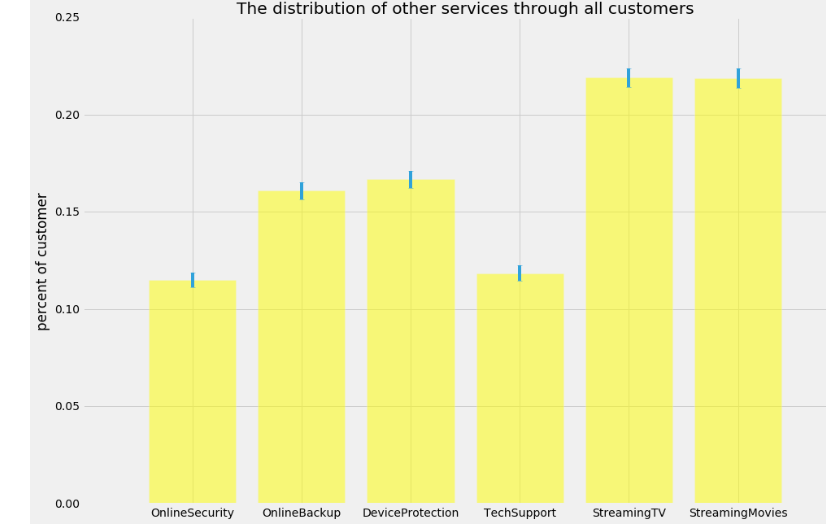
1. **The distribution of other services through all customers**

From The pie chart, the percentage of customer from using streamTv and streammovies are majority 18.1% and 18% of total customers. The percentage of customer from using DeviceProtection and OnlineBackup are the second place, 16.8% of total customers. The last two are OnlineSecurity and TechSupport, 15.1% and 15.2% of total customers.



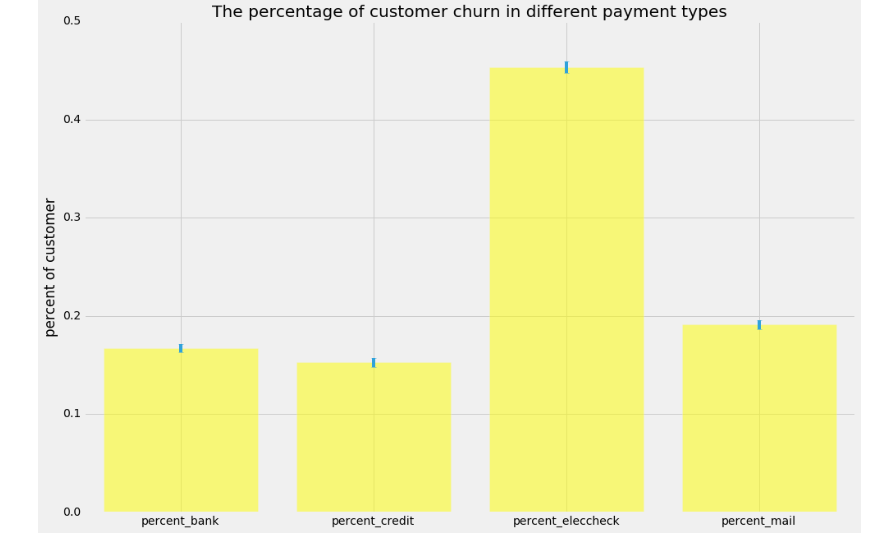
# The percentage of customer churn from other services

From The bar chart, the percentage of customer churn from using streamTv and streamMovies are majority, 21% of total customers. The percentage of customer churn from using DeviceProtection is the second place, 16.6% of total customers. The percentage of customer churn from using OnlineBackup is 16% of total customers. The last two are OnlineSecurity and TechSupport, 11.5% and 11.8% of total customers.



# The percentage of customer churn in different payment types

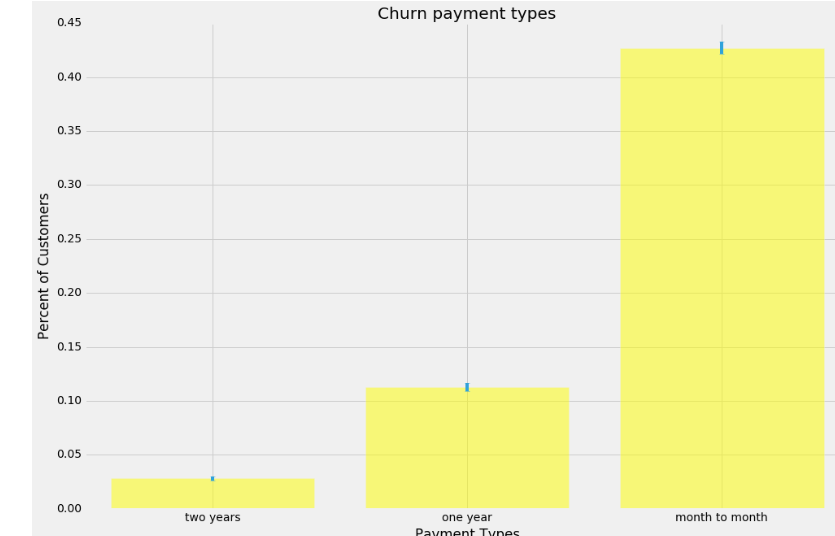
The bar chart illustrates the percentage of customer churn in four different payment types: Credit card (automatic), Bank transfer (automatic), Mailed check, Electronic check. The Electronic check is the most popular payment type with 42% of total customers. The second highest payment type is Mailed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



# 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.

Hypothesis: customers are likely to churn when they sign up for month to month payment plan



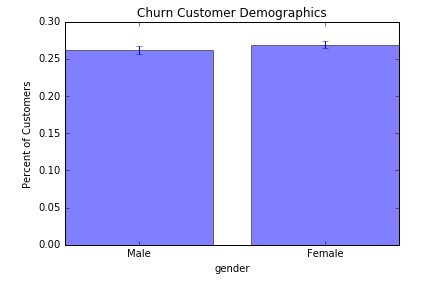
1. **The percentage of customer churn in Demographics (gender/partner/dependents)**

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 customers have no partner likely to churn than partner. The percent of churn customer with partner is 19 % total customer, and the percent with no partner is 32% of total customer. 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 customer, and the dependent is 15% of total customers.

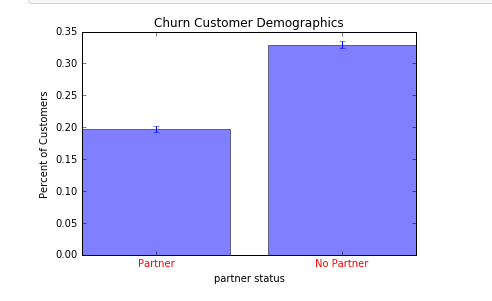
Hypothesis: the customer has no partner and no dependent is highly percent to churn than other.

In [73]:

**Gender**

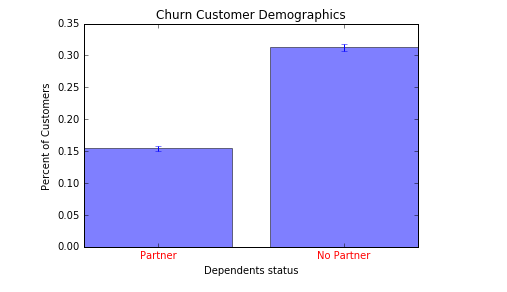


**Partner**



**Dependents:**

In [76]:



# IV. 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 custom

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, pvalue=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 different between male and female in customer churn.

Hypothesis: there is a different 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 different between dependent and no dependent in customer churn.

Hypothesis: there is a different 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 different between partner and no partner in customer churn.

Hypothesis: there is a different 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 different between senior and no senior in customer churn.

Hypothesis: there is a different 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. Although Statistical significance, the correlation coefficient for the tenure and churn is -0.35, which shows that the tenure and the churn has no correlation. However, p value is less than 0.05, there are some significant correlation coefficient between the tenure and the churn. Moreover, we may conduct more tests on other variables from data set in order to determine churn customers, such as gender, dependent, partner, and senior citizen.

Demographic factors are the most important factors to analyze customer churn, because of the correlation between demographic and Churn. In gender, there is a different between male and female group. In partner status, there is a different 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 different between no dependent and dependent. Finally, the senior and non-senior also are different in the percentage of churn customers.

1. **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 this case, the conventional model evaluation methods do not accurately measure model performance when faced with imbalanced datasets. We has to know the different errors and concepts.

In this type of evaluation: False Positive and F**alse Negative** errors.

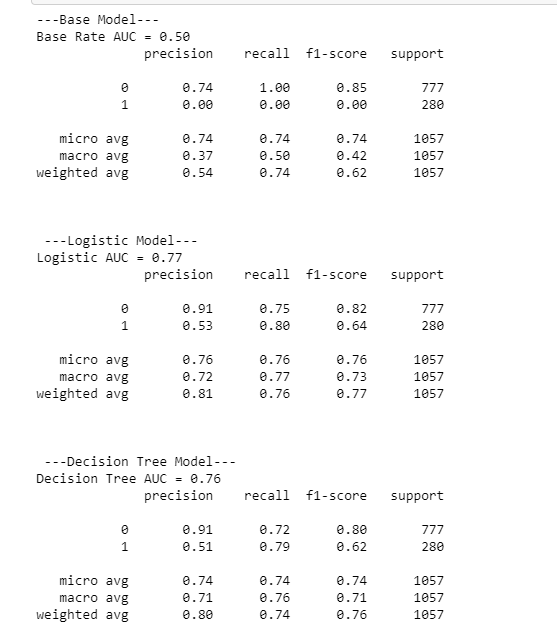
**False Positives (Type I Error)**: You predict that the customers will churn, but do not

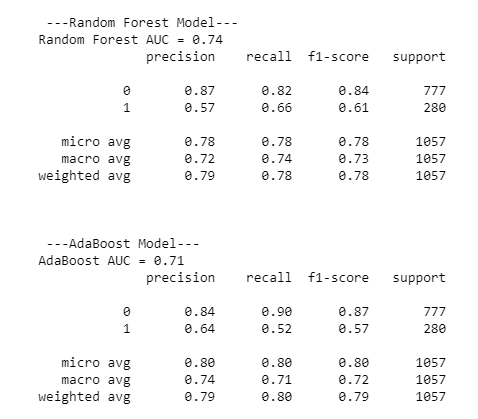
**False Negatives (Type II Error)**: You predict that the customers will not churn, but does churn

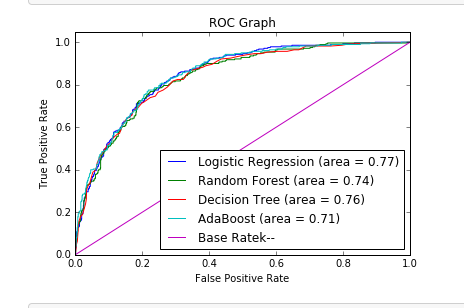
In this problem, what type of errors do we care about more? False Positives or False Negatives?

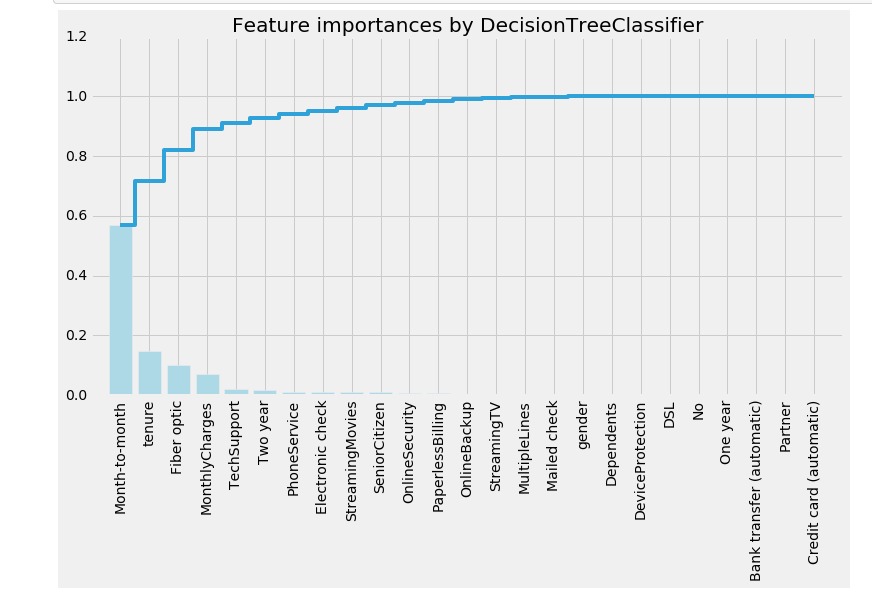
**Different Ways to Evaluate Classification Models:**

* Decision tree models allow you to develop classification systems that predict or classify future observations based on a set of decision rules
* AdaBoost Model: AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem..
* Logistic Regression Model:   Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.
* Random Forest Model: Random Forest grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest)









**Feature Importance**

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

# Interpreting the Data

Summary: With all of this information, this is what we know why customers probably churn:

1. Customer generally left when they are month to month payment type.
2. Customer is likely to churn in first ten months.
3. Customers is likely to leave customer when they use Fiber optic service.
4. But there is 0% customers use fiber only, customers are likely to cancel services when they use phone and fiber.
5. Customers have no partner and no dependent is highly percentage to churn than other.
6. Customer month to month, tenure, and fiber optic were the three biggest factors in determining churn.

## **Potential Solution**

**Binary Classification**: Churn V.S. Non Churn

**Instance Scoring**: Likelihood of customers responding to an offer/incentive to save them from leaving.

**Need for Application**: Save customers from leaving

For this problem, we should look at the probability that customer will leave the company rather than the prediction whether customer left within certain time frame. We can rank the probably of leaving for employees, we can have deal or any benefit to keep them.

Consider customer churn where a customer is given deal by Telco because telco think the customer will leave the company within a month, but the employee actually does not. This is a false positive. This mistake could be expensive and time consuming for both telco and customers, but is a good investment for relational growth.

Compare this with the opposite error, where Telco does not give deal to the customers and they do leave. This is a false negative. This type of error is more unfavorable because the company lost an customer, which could lead to lose money.

Depending on these errors, different deal are weighed based on the type of customer. For example, customer has multiple service purchased, on time payment which is good customer (potential growing) vs customer has one service purchased, not on time payment which is not good customer(better to lose this customer than saved ).