Predict Customer Churn
Using Python & Machine Learning Techniques

Jesmeen Fatema

May 2020

Business Problem

- Retaining customers is extremely important for companies as
  - It costs more to acquire new customers than it does to retain existing customers.
  - An increase in customer retention by 5% can create at least a 25% increase in profit.
  - The company can spend less on the operating costs of having to acquire new customers.
- Create a model that can accurately predict / classify if a customer is likely to churn.
- Analyze the data to come up with a possible strategic retention plan.

Data Description

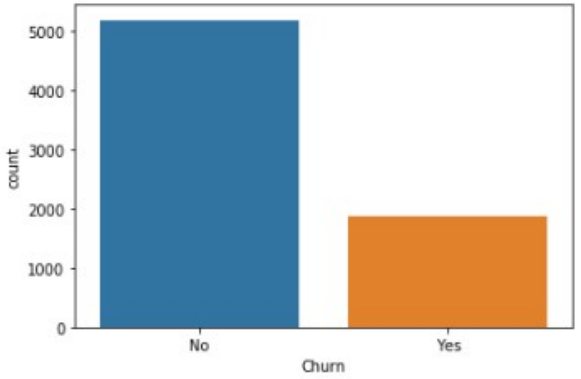
- Sources of Data
  - The data set that will be used in this analysis will come from Kaggle (the Telco company).

Methodology / Data Analysis

- Downloaded (csv file) the data from Kaggle.
- Imported relevant libraries that would be used throughout the program.
- Analyzed the data.
- There are 7,043 customers and 21 data points on each customer
- There were no missing values
- From the statistical analysis, the longest tenure was 72 months, and the monthly maximum charge was $118.75. The minimum monthly charge is about $30.09. The monthly average charge is $64.76.

Methodology / Data Analysis (Contd.)

- Based on the data set 5,174 customers were retained (did not churn) and 1,869 customers churned.



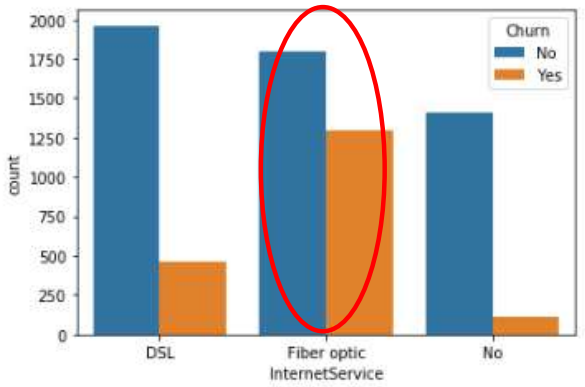
- Gender has virtually no effect on the customer churn.

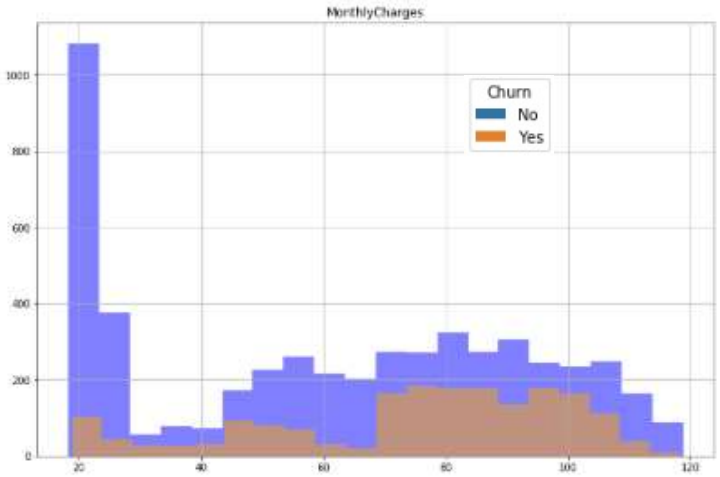
Methodology / Data Analysis (Contd.)

- Customers with Fiber Optic internet service have churned the most

- Most of the loyal customers that stayed with the company had a monthly charge between $20 and $30

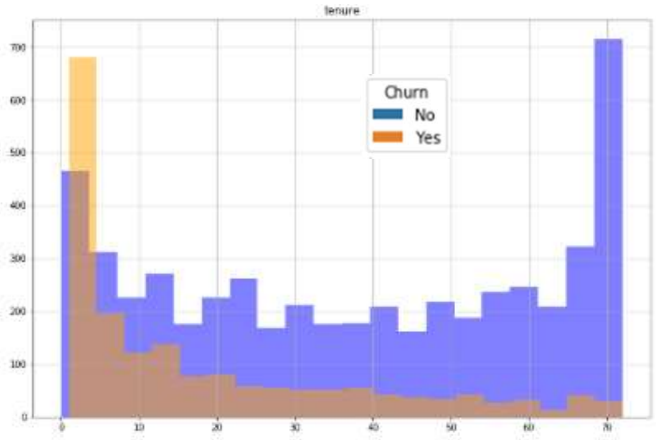
- Most of the customers that churned had a monthly charge of $70 to $100





Methodology / Data Analysis (Contd.)

- Most of the customers that churned stayed between 1 and 9 months with the company
- Most of the retained customers had a tenure between 24 and 72 months



Data Processing & Cleaning

- Dropped unnecessary columns (i.e. customerID)
- Covert all the non-numeric columns to numerical data types by ***Label Encoding***
- Split data into independent / feature (X) and dependent (Y) variables
- Standardizing / scaling the features (X)
- Split the data into training (80%) and testing (20%) data sets

Create Machine Learning Models

- Created a function for different machine learning models to identify which model performs the best :
  - Logistic Regression
  - KNN Classifier
  - Support Vector Classifier
  - Gaussian NB
  - Decision Tree Classifier, &
  - Random Forest Classifier

Evaluate Machine Learning Models

- Precisions, recalls, f-scores, and accuracies of some of models are as follows:

M/L Models	Precision	Recall	f1-score	Accuracy (Test Set)	Accuracy (Training Set)
Log. Regression	0.85	0.91	0.88	0.82	0.80
SVC	0.86	0.90	0.88	0.82	0.80
Random Forest	0.82	0.91	0.86	0.78	0.99
Decision Tree	0.81	0.80	0.81	0.72	0.98

- As expected the Decision Tree and Random Forest Classifiers have the best training accuracies (overfitting tendency)
- However, the accuracies with the testing data set for the Decision Tree and Random Forest Classifiers are not the best.
- Considering all the info in the above table, the Logistic Regression model performs the best.

Conclusions and Recommendations

- The accuracy of the Logistic Regression model was about 82% which is better than the guessing of 73.46%.
- The company may want to lower it's monthly charges at least for new customers for the first 2 years to increase the retention rate.
- The churn rate was fairly high for the customers with Fiber Optic Internet service. The company may want to stop providing that service, which will help reduce the cost. This will be a good strategy to help retain their customers and reduce customer churn.