Springboard DSC Capstone Project 1:
Milestone Report

Albert Chiu
October 2018

Problem Statement:
For most businesses, new customer acquisition is an expensive and labor-intensive effort. While this is critical for the early stages of a business, once a customer base is built, there should be increased effort in customer retention. Happy, current customers are vital: they reduce the expenditure on marketing, they provide free word of mouth advertising, they are more likely to provide valued feedback and are more likely to pay for premium features/products.

Churn is the metric used to measure customer retention typically measured as rate at which customers stop subscribing to a service.

For my Capstone Project 1, I will build models to estimate the probability of a customer will stop subscribing to a service. Using these models, I plan to infer what the profile of a churn customer is. This project will help subscription-based businesses identify customers that are prone to churn and make more informed decisions on how they may retain these individuals.

Data Set:
The data set for this project was obtained from Kaggle in csv file:
https://www.kaggle.com/blastchar/telco-customer-churn/home.

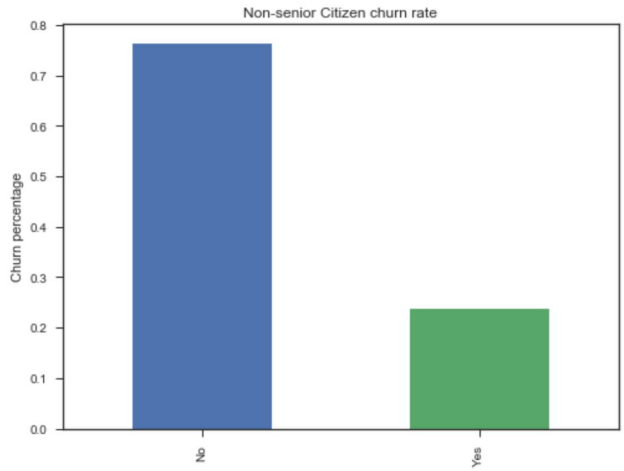
It contains 21 features and 7043 observations. Most of the features were categorical and described either the services that the customer was signed up for or the demographics of the customer. From exploring the csv file, the data set was fairly clean and required minimal modifications:

- 11 of the observations contained null entries, which were dropped
- The TotalCharges feature column was converted from object to float

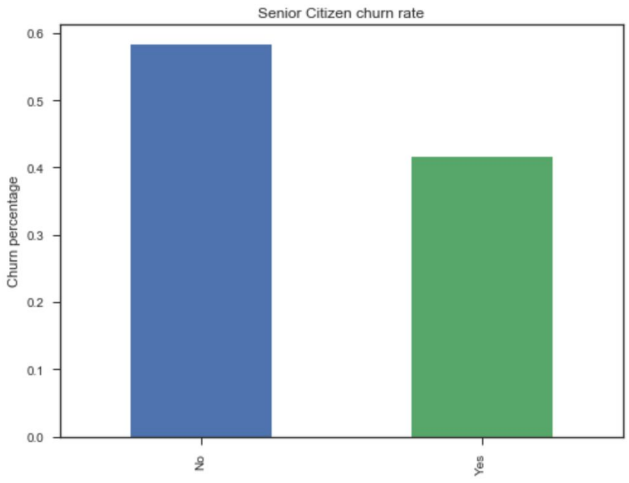
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7032 entries, 0 to 7042
Data columns (total 21 columns):
customerID          7032 non-null object
gender              7032 non-null object
SeniorCitizen       7032 non-null int64
Partner             7032 non-null object
Dependents          7032 non-null object
tenure              7032 non-null int64
PhoneService        7032 non-null object
MultipleLines       7032 non-null object
InternetService     7032 non-null object
OnlineSecurity      7032 non-null object
OnlineBackup        7032 non-null object
DeviceProtection    7032 non-null object
TechSupport         7032 non-null object
StreamingTV         7032 non-null object
StreamingMovies     7032 non-null object
Contract            7032 non-null object
PaperlessBilling    7032 non-null object
PaymentMethod       7032 non-null object
MonthlyCharges      7032 non-null float64
TotalCharges        7032 non-null float64
Churn               7032 non-null object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.2+ MB
```

Exploratory Data Analysis:

Using data visualization and statistical inference techniques, I gained some insights into the behavior of the customers in this data set. For example, I discovered that senior citizens surprisingly had a much higher churn rate than non-senior citizens:

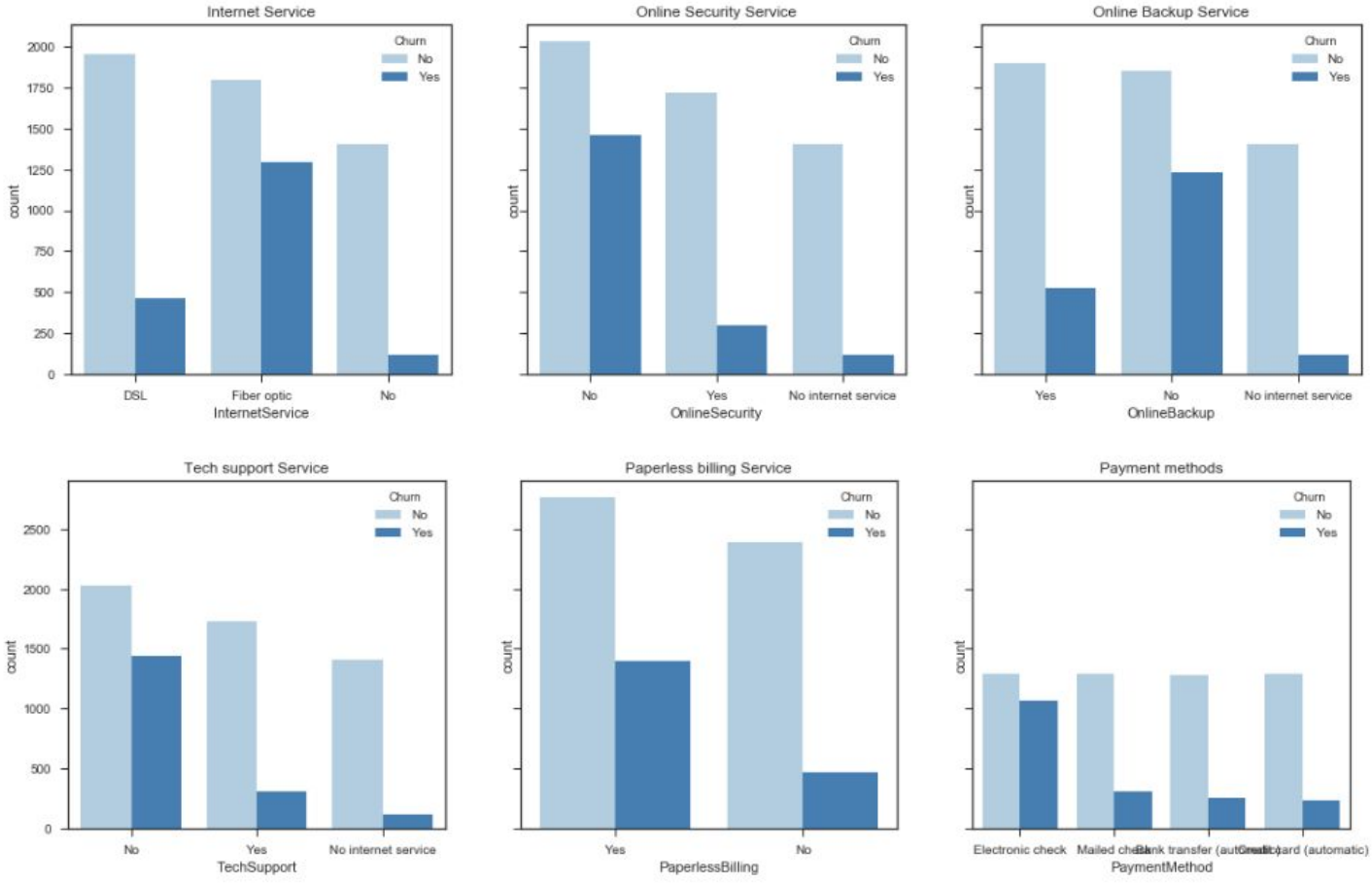



```
Non-senior citizen churn rates:
No     0.763497
Yes    0.236503
Name: Churn, dtype: float64
```




```
Senior Citizen churn rate:
No     0.583187
Yes    0.416813
Name: Churn, dtype: float64
```

Additionally, the services that non-churn and churn customers are enrolled in are very different. Non-churn customers are much more likely to be signed for tech support, online security and backup service. Churn customers heavily prefer fiber optic internet service and electronic payment/billing.



4. Using alpha = 0.05, is there a significant difference in the percentage of churn customers and non-churn customers in choosing fiber optic internet service?

H0: There is no significant difference in the percentage of churn customers vs non-churn customers that use fiber optic for internet service.

Ha: There is a significant difference in the percentage of churn customers vs non-churn customers that use fiber optic for internet service.

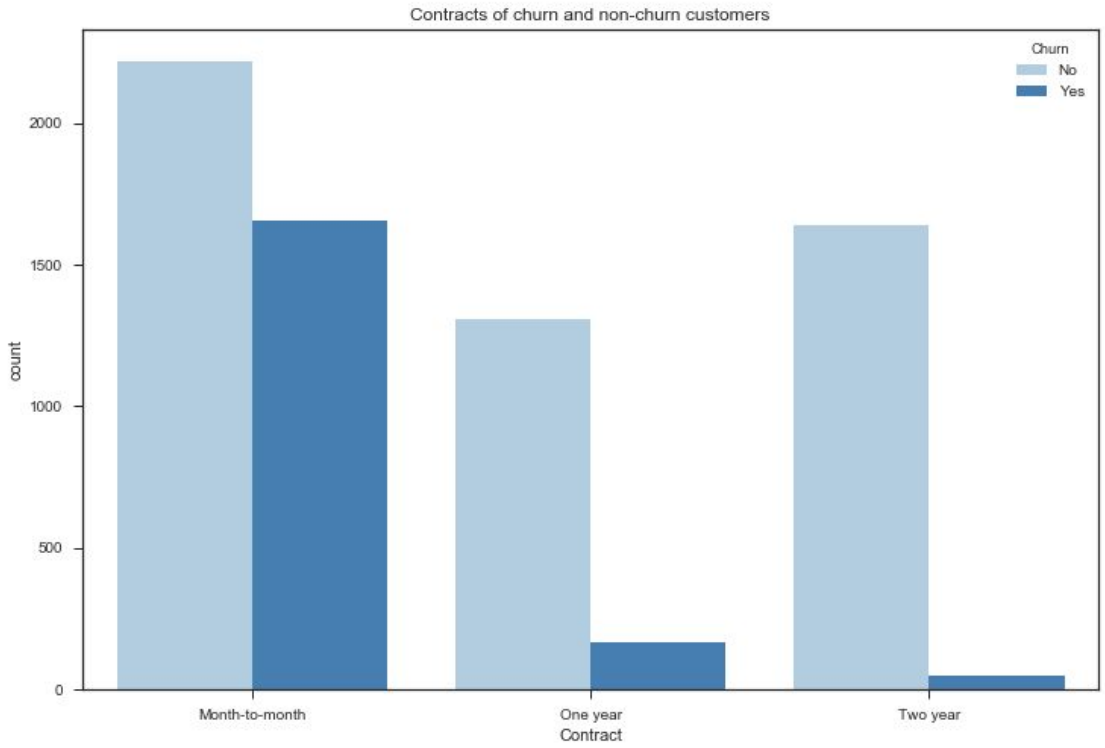
```
# number of churn and non-churn customers using fiber optic internet service
churn_fiber_optic = churn_df.InternetService[churn_df.InternetService == 'Fiber optic'].count()
nochurn_fiber_optic = nochurn_df.InternetService[nochurn_df.InternetService == 'Fiber optic'].count()

# calculating p-value
stat, pval = pz(np.array([churn_fiber_optic, nochurn_fiber_optic]), np.array([len(churn_df),len(nochurn_df)]), value = 
print('P-value:', pval)
```

```
P-value: 1.3792417311863573e-146
```

```
The p-value is less than alpha = 0.05.  Therefore, we can reject the null hypothesis and accept the alternative
hypothesis that percentage of
```

Also, as expected, churn customers overwhelmingly prefer month to month contracts than a long term term contract when compared to non-churn customers.




```
# percentage of each contract for churning customers
churn_df['Contract'][churn_df['Churn']=='Yes'].value_counts('Yes')
```

```
Month-to-month    0.885500
One year          0.088818
Two year          0.025682
Name: Contract, dtype: float64
```

```
# percentage of each contract for non-churning customers
churn_df['Contract'][churn_df['Churn']=='No'].value_counts('Yes')
```

```
Month-to-month    0.429983
Two year          0.317064
One year          0.252954
Name: Contract, dtype: float64
```

Month-to-month contracts account for over 88% of the contract type for churning customers compared to 43% for non-churning customers. When applying statistical methods, we can see that this is a statistical significant difference:

6. Using alpha = 0.05, is there a significant difference in the percentage of churn and non-churn customers that are on month-to-month contracts?

H0: There is no significant difference in the percentage of churn and non-churn customers that are on month-to-month contracts.

Ha: There is a significant difference in the percentage of churn and non-churn customers that are on month-to-month contracts.

```
# number of churn and non-churn customers that are on month-to-month contracts
mo_to_mo_churn = churn_df.Contract[churn_df.Contract == 'Month-to-month'].count()
mo_to_mo_nochurn = nochurn_df.Contract[nochurn_df.Contract == 'Month-to-month'].count()

# calculating p-value
stat, pval = pz(np.array([mo_to_mo_churn, mo_to_mo_nochurn]), np.array([len(churn_df), len(nochurn_df)]), value = 0)
print('P-value:', pval)
```

```
P-value: 2.7960092525873125e-252
```

The p-value is less than alpha = 0.05. Therefore, we can reject the null hypothesis and accept the alternative hypothesis that there is significant difference in the percentage of churn customers and non-churn customers that are on month-to-month contracts.

Summary:

After exploring the data set in detail, it is clear that there are some statistically significant differences between churn and non-churn customers in the services they use, the payment methods they select and their age group. In the next step of the project, I will attempt to use machine learning techniques to develop algorithms to classify churn and non-churn customers based on these features.