Capstone Project 2: Milestone Report

Telecom Churn

Problem

The problem I want to solve is being able to predict when a customer is going to churn so that before the customer even thinks of leaving the subscription, our client can take action.

Client

My client is a Telecom company that w/ the given predictive model and information, will be able to provide promotions, offers, etc. to customers who are predicted to churn given the information.

Data Wrangling Scope

I am using a dataset on Kaggle that will need to be imported, combined, manipulated, and prepared to better understand the dataset.

Rough Outline

Start w/ exploring the data. There are numerous columns within this data set so we'll be able to delve into looking at potential significant predictor variables. We will use heat maps, histograms, bar charts, etc. to get a good sense of the data. Then implement some probabilistic tests to dive into this deeper. Lastly, train, validate and test multiple algorithms to predict if customers will churn.

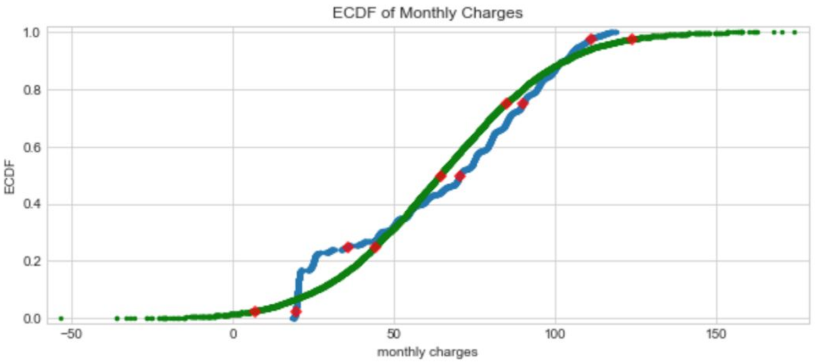
Deliverables

Jupyter Notebook, PowerPoint Slides.

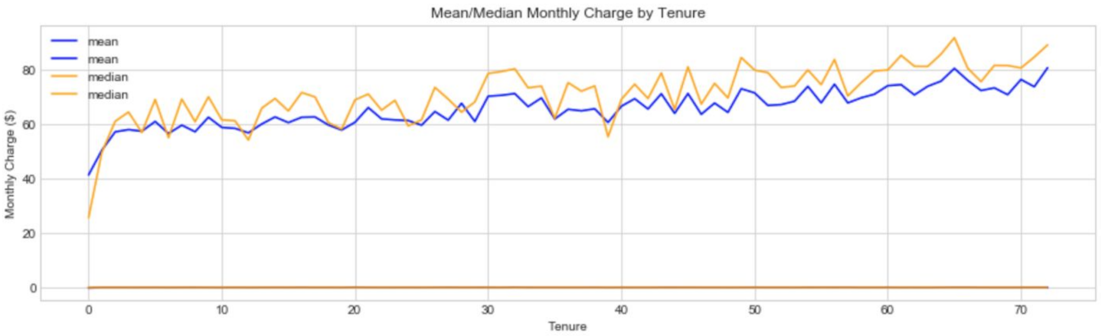
Data Wrangling

This dataset is actually extremely clean which is to no surprise because this was curated from IBM. I performed the necessary procedures to make sure of this and validate.

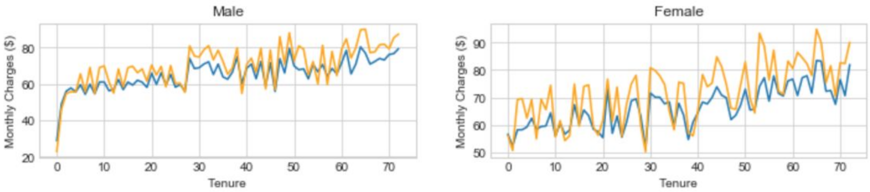
Exploratory Data Analysis



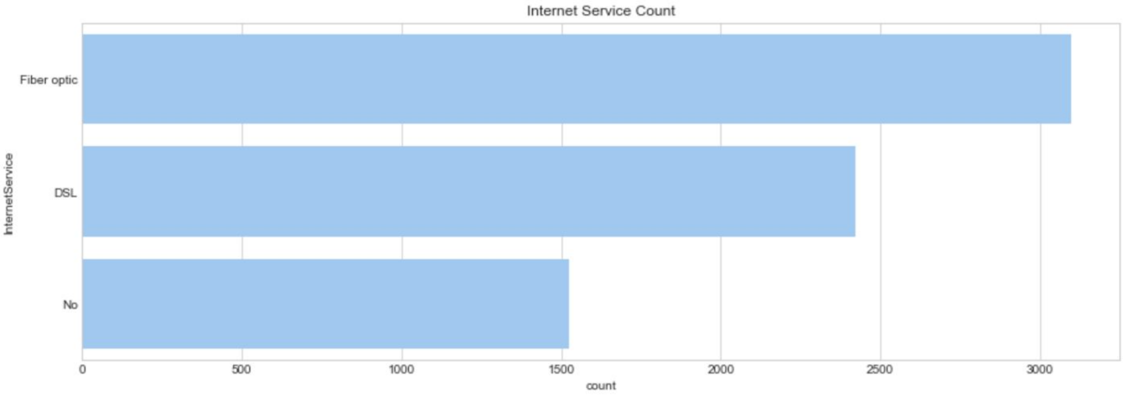
Looks very interesting and doesn't neccessarily follow a normal distribution w/ respect to Monthly Charges. Let's take a deeper dive.

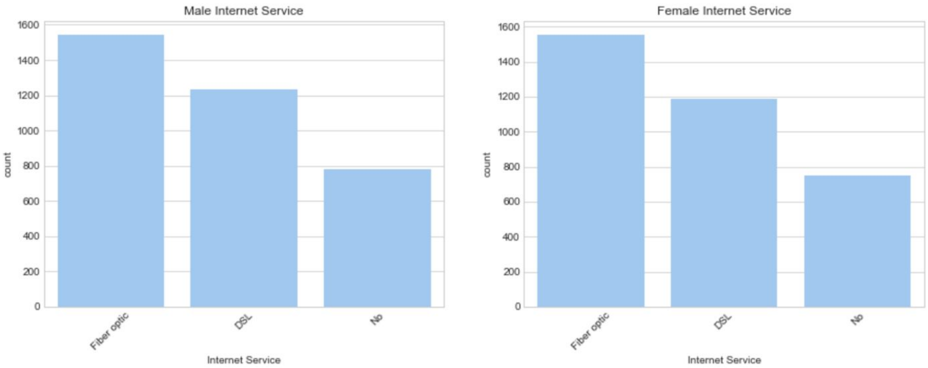


There looks like a upward trend w/ monthly charges for both mean and median values which does intuitively make sense due to loyalty and the investment throughout the years

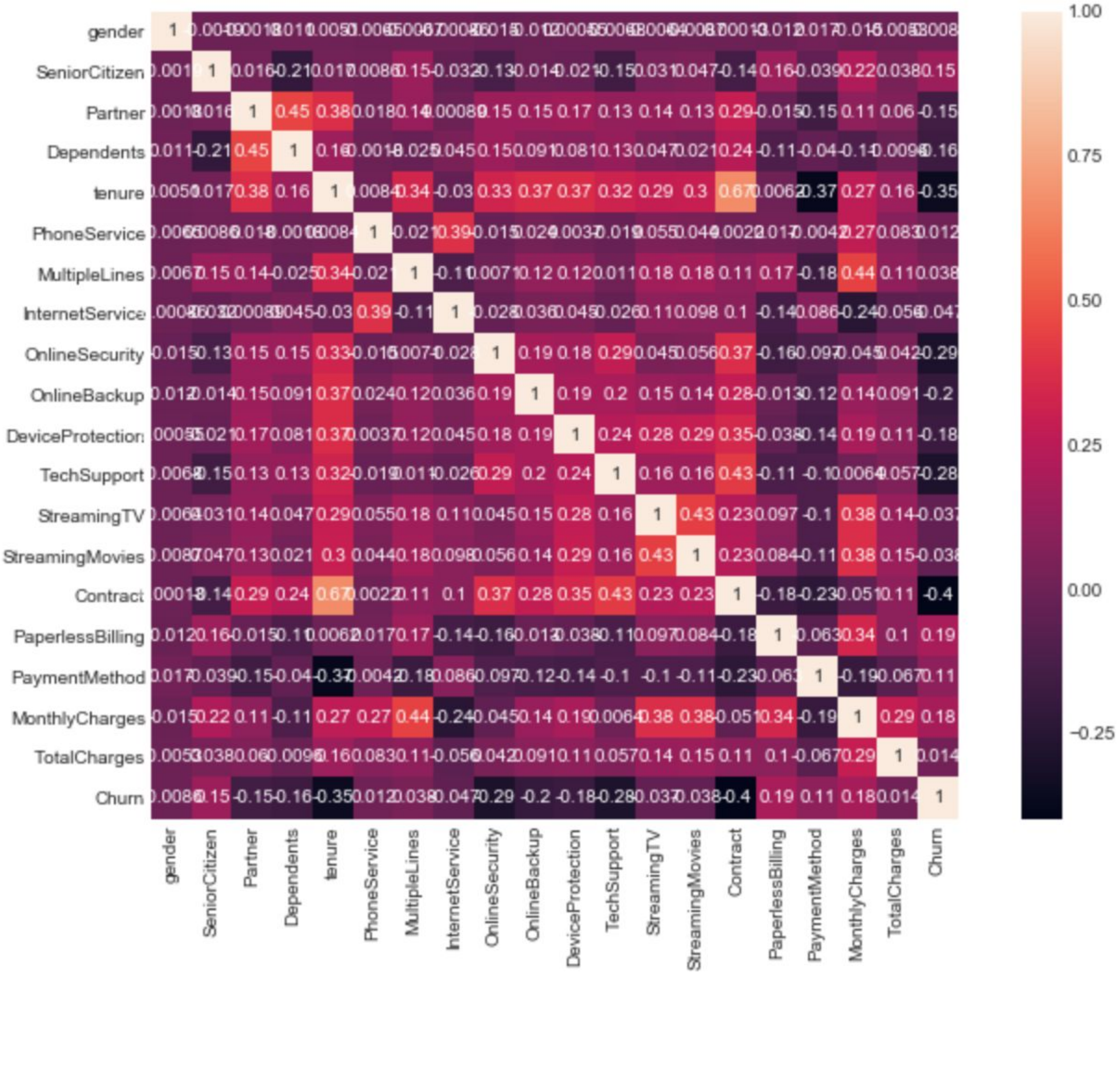


There's a larger variability w/ respect for remales. The general trend of monthly charges is still upwards but the difference in variability is quite interesting but I don't want to make any assumptions in regards to gender as of now.





Males and Females who purchase internet service have relatively the same number of purchases w/ respect to internet service.



Inferential Statistics

The purpose of this analysis is to see if the **difference between the proportion of male and female churn rates** are the same so that we can see if it is statistically **AND** practically significant to take action among the respective gender.

Central Limit Theorem Conditions¶

Random Condition: Each customer is randomly obtained and recorded and thus, our sample **meets** the random condition of the Central Limit Theorem.
Normal Condition: Both sample proportions of male and female churn rates, when multiplied by sample size, are **greater than 10**. Their proportions are both in the **middle of 0 and 1** as well as have a **large number of records**. Thus, this sampling distribution for both sample proportions **meet** the normal condition.
Independence Condition: Both samples have sample sizes that are **less than 10%** of the number of members. Thus, this our sample distribution for both sample proportions **meet** the independent condition.

Null & Alternative Hypothesis¶

Null Hypothesis: In terms of churn, there is **no difference** for female and male customers.

Alternative Hypothesis: In terms of churn, there is **a difference** for female and male customers.

Significance Level & Power

Significance Level: α = 0.01

Power: We are worried of making a **Type I error** because if there is no difference between the sample proportion of churn for male and female and reject this, telecom companies will be taking action on something that is insignificant --- as a result they will be **wasting time** and **losing money on action items**.

The is a **99% chance** that the true difference between male and female churn percentages is between *-.016 and .013.*

This means we are **99% confident** that there exists a difference between male and female churn percentages.

The probability of getting a Z-score **as extreme or more extreme** than -.22 is .004116%, *assuming the null-hypothesis is true*.

Since our p-value is **greater than** our predetermined significance level of 0.01, we **do not reject** the null hypothesis and assume **no statistical significance** in the difference between proportions of male and female churn rates.

Conclusion

Firstly, our statistical analysis says that we are *confident* that the true difference between winning percentages on male and female is between -.016 and 0.013, 99% of the time. Thus, we are quite confident that the true difference is between -.016 and 0.013.

Thus, gender should not be a predicting variable when looking into churn since we do not reject the null hypothesis. .

Thus, when advising a telecom to take action upon this analysis or not to take action, it is clear to not focus on the gender of the customer although it was a factor that I was curious in delving into.

Inferential Statistics pt. 2

The purpose of this analysis is to see if the **difference between the proportion of monthly and one year plus churn rates** are the same so that we can see if it is statistically **AND** practically significant to take action among the respective gender.

Central Limit Theorem Conditions¶

Random Condition: Each customer is randomly obtained and recorded and thus, our sample **meets** the random condition of the Central Limit Theorem.
Normal Condition: Both sample proportions of month and one year plus churn rates, when multiplied by sample size, are **greater than 10**. Their proportions are both in the **middle of 0 and 1** as well as have a **large number of records**. Thus, this sampling distribution for both sample proportions **meet** the normal condition.
Independence Condition: Both samples have sample sizes that are **less than 10%** of the number of members. Thus, this our sample distribution for both sample proportions **meet** the independent condition.

Null & Alternative Hypothesis¶

Null Hypothesis: In terms of churn, there is **no difference** for customers who have monthly contracts and customers who have one year plus contracts.

Alternative Hypothesis: In terms of churn, there is **a difference** customers who have monthly contracts and customers who have one year plus contracts.

Significance Level & Power

Significance Level: α = 0.01

Power: We are worried of making a **Type I error** because if there is no difference between the sample proportion of churn for monthly and one year plus contracts and reject this, telecom companies will be taking action on something that is insignificant --- as a result they will be **wasting time** and **losing money on action items**.

There is a **99% chance** that the true difference between monthly and one year plus churn percentages is between *.191 and .219.*

This means we are **99% confident** that there exists a difference between monthly contracts and one year plus churn percentages.

The probability of getting a Z-score **as extreme or more extreme** than 35.79 is 9%, *assuming the null-hypothesis is true*.

Since our p-value is **less than** our predetermined significance level of 0.01, we **do reject** the null hypothesis and assume **statistical significance** in the difference between proportions of monthly and one year plus churn rates.

Conclusion

Firstly, our statistical analysis says that we are *confident* that the true difference between churn rates for monthly and one year plus contracts is between .191 and 0.219, 99% of the time. Thus, we are quite confident that the true difference is between .191 and .219.

Thus, contract types will be a good predictor variable to look further into when implementing our model.

Thus, when advising a telecom to take action upon this analysis or not to take action, it is clear to focus on the contract types of the customers which was quite expected beforehand.