EDA Inferential Statistics

December 28th, 2018

Regarding:

Predicting Customer Churn

Springboard Course Work
Capstone Project #1

Cliff Robbins
cliff@gearforgesoftware.com
M 612.701.2998

Proposal

My project will focus on a problem that 28 million business face each day of operation, customer churn.

Definition

Customer churn, also known as customer attrition, customer turnover or customer defection is the loss of clients or customers. Many companies include customer churn rate as part of their monitoring metrics because the cost of retaining current customers compared to acquiring new customers is much less.
Within customer churn there is the concept of voluntary and involuntary churn with voluntary being a customer leaves on their own choice while involuntary could be attributed to customer relocation to a long term care facility, death or customer relocation in a different state/geography. In most analytical models, involuntary churn is excluded from the metric.

Formulation of a Question

When a company first starts up, the founding members can typically handle all of the various customer concerns. As the company continues to grow, the founders can no longer service all of the various clients with support handled by a customer service team. The customer service team focuses on current issues and a proactive approach is lost.
As the company grows, the company still cares about its clients; however, due to the large customer base they can no longer address each and every customer. This is a real problem for companies. How does a company proactively predict if a customer is happy or unhappy? How does a company know if a customer is so unhappy that they are willing to leave? If a company knew if a customer was getting ready to leave, could they reach out to the customer and mend the relationship?

Hypothesis

I believe past customer data can predict future customer churn.

Prediction

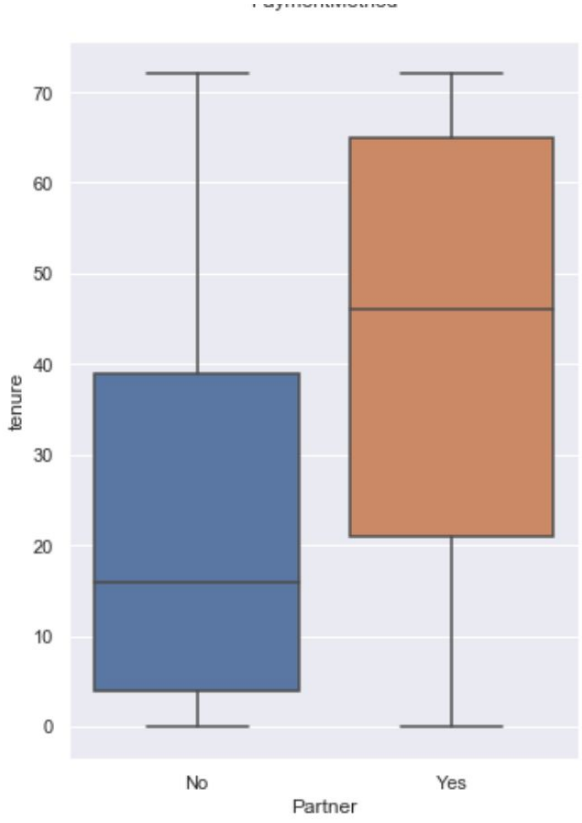
If I had past customer data that showed various features and whether they stayed or churned we could use that data to predict future outcomes of current customers.

Goal of EDA Inferential Statistics:

The goal is to use inferential statistics to identify strong correlations between pairs of independent variables along with practicing using different tests to analyse relationships between variables.

Investigation:

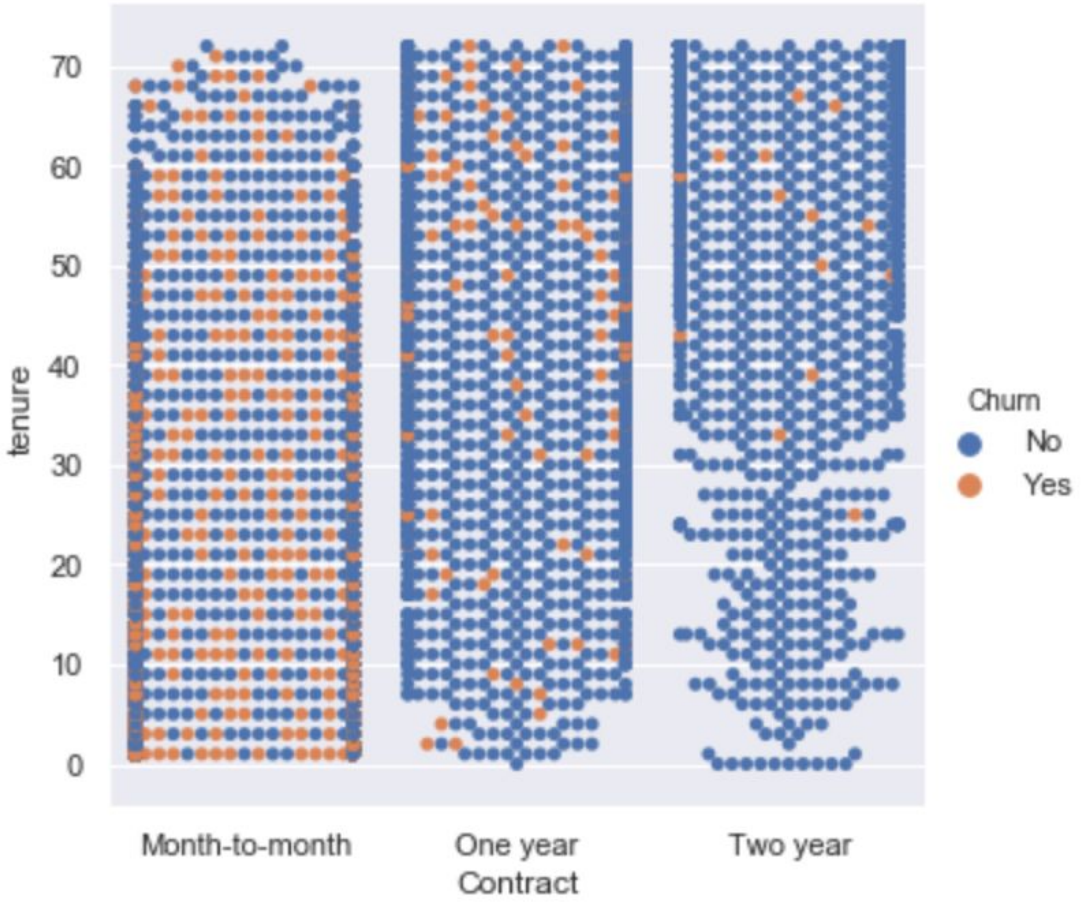
The first step of EDA was to visualize the data to understand the relationship between the various features and the predictor. I initially used tenure to understand how long a customer stayed hypothesizing that the longer a customer stayed the less likely they are to churn.
Because most of my data is categorical, I started with boxplots to understand the categories within each feature against tenure. What I noticed was that some of the categories within each feature had higher levels of tenure than their counterparts. Here is an example of Partner vs Tenure that shows those with a partner have a higher level of tenure.



The next step was to take the following features and do a catplot of the categories against tenure along with churn.

1. Phone Service
2. Multiple Lines
3. Internet Service
4. Contract Length
5. Paperless Billing
6. Payment Method
7. Dependents
8. Senior Citizen
9. Partner

Amongst those features, I could see distinct patterns that indicate some categories are prone to churn more than others. In the below graphic, you can see that month-to-month billing has more frequencies of churn than the other 2 contract types.



Leveraging Inferential Statistics:

The features that provide visual correlations between the categories and churn next need to be checked for correlation strength.

I set Alpha equal to 0.05 or 5%.
My hypothesis is no relationship between categories and churn.

We will leverage p-value, Pearson Chi-Square and Cramer's phi.

Note: Cramer's phi will measure how strong the relationship between the 2 variables, the closer to 1 the strong the relationship.

I tested the following categorical features against churn:

- gender
- SeniorCitizen
- Partner
- Dependents
- PhoneService
- MultipleLines
- InternetService
- Contract
- PaperlessBilling
- PaymentMethod

Here are the results for all of the categorical features:

- gender
    - No Relationship (fail to reject H0)
- SeniorCitizen
    - Relationship (reject H0)
- Partner
    - Relationship (reject H0)
- Dependents
    - Relationship (reject H0)
- PhoneService
    - No Relationship (fail to reject H0)
- MultipleLines

    - Relationship (reject H0)
- InternetService
    - Relationship (reject H0)
- Contract
    - Relationship (reject H0)
- PaperlessBilling
    - Relationship (reject H0)
- PaymentMethod
    - Relationship (reject H0)

Conclusion:
Based on my visual EDA I had anticipated that all of the following features had correlation or relationships between the categorical feature and churn; however, I was wrong.

- gender
- SeniorCitizen
- Partner
- Dependents
- PhoneService
- MultipleLines
- InternetService
- Contract
- PaperlessBilling
- PaymentMethod

What I found was that the gender and phone service did not have a strong relationship with churn which is contrary to what I expected.

The dataset has 18 features and 1 target used to determine if the customer churned or not. Of the 18 features, 8 features have a strong relationship with churn.