Published in Towards Data Science

Andrew Cole Follow

May 8, 2020 · 7 min read · Listen

Predicting Customer Churn With Classification Modeling

Part 1: Exploratory Data Analysis

In todays commercial world competition is high and every customer is valuable. Understanding the customer is of utmost importance, including being able to understand the behavior patterns of that customer. **Customer Churn** is the rate at which a commercial (very prevalent in SaaS platforms) customer leaves the commercial business and takes their money elsewhere. Understanding customer churn is vital to the success of a company and a churn analysis is the first step to understanding the customer.

Photo by Clay Banks on Unsplash

I decided to perform a churn analysis from a Kaggle data set which gives the customer information data of a telecommunications company (Telcom) trying to better understand their customer churn likelihood. While we will eventually build a classification model to predict likelihood of customer churn, we must first take a deep dive into the Exploratory Data Analysis (EDA) process to get a better understanding of our data. Github Repository with code and notebooks can be found here.

The Data

As mentioned above, the data is sourced from Kaggle. In our dataset, we have 7043 rows (each representing a unique customer) with 21 columns: 19 features, 1 target feature (Churn). The data is composed of both numerical and categorical features, so we will need to address each of the datatypes respectively.

Target:

- Churn — Whether the customer churned or not (Yes, No)

Numeric Features:

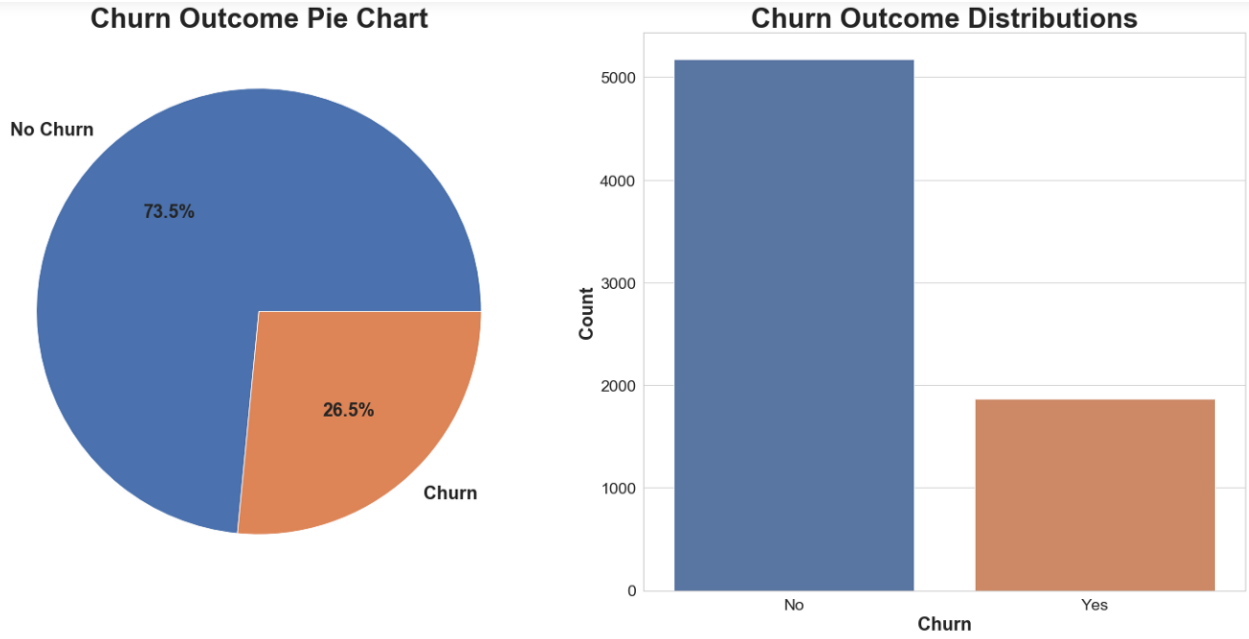
- Tenure — Number of months the customer has been with the company
- MonthlyCharges — The monthly amount charged to the customer
- TotalCharges — The total amount charged to the customer

Categorical Features:

- CustomerID
- Gender — M/F
- SeniorCitizen — Whether the customer is a senior citizen or not (1, 0)
- Partner — Whether customer has a partner or not (Yes, No)
- Dependents — Whether customer has dependents or not (Yes, No)

- MulitpleLines — Whether the customer has multiple lines or not (Yes, No, No Phone Service)
- InternetService — Customer's internet service type (DSL, Fiber Optic, None)
- OnlineSecurity — Whether the customer has Online Security add-on (Yes, No, No Internet Service)
- OnlineBackup — Whether the customer has Online Backup add-on (Yes, No, No Internet Service)
- DeviceProtection — Whether the customer has Device Protection add-on (Yes, No, No Internet Service)
- TechSupport — Whether the customer has Tech Support add-on (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 — Term of the customer's contract (Monthly, 1-Year, 2-Year)
- PaperlessBilling — Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod — The customer's payment method (E-Check, Mailed Check, Bank Transfer (Auto), Credit Card (Auto))

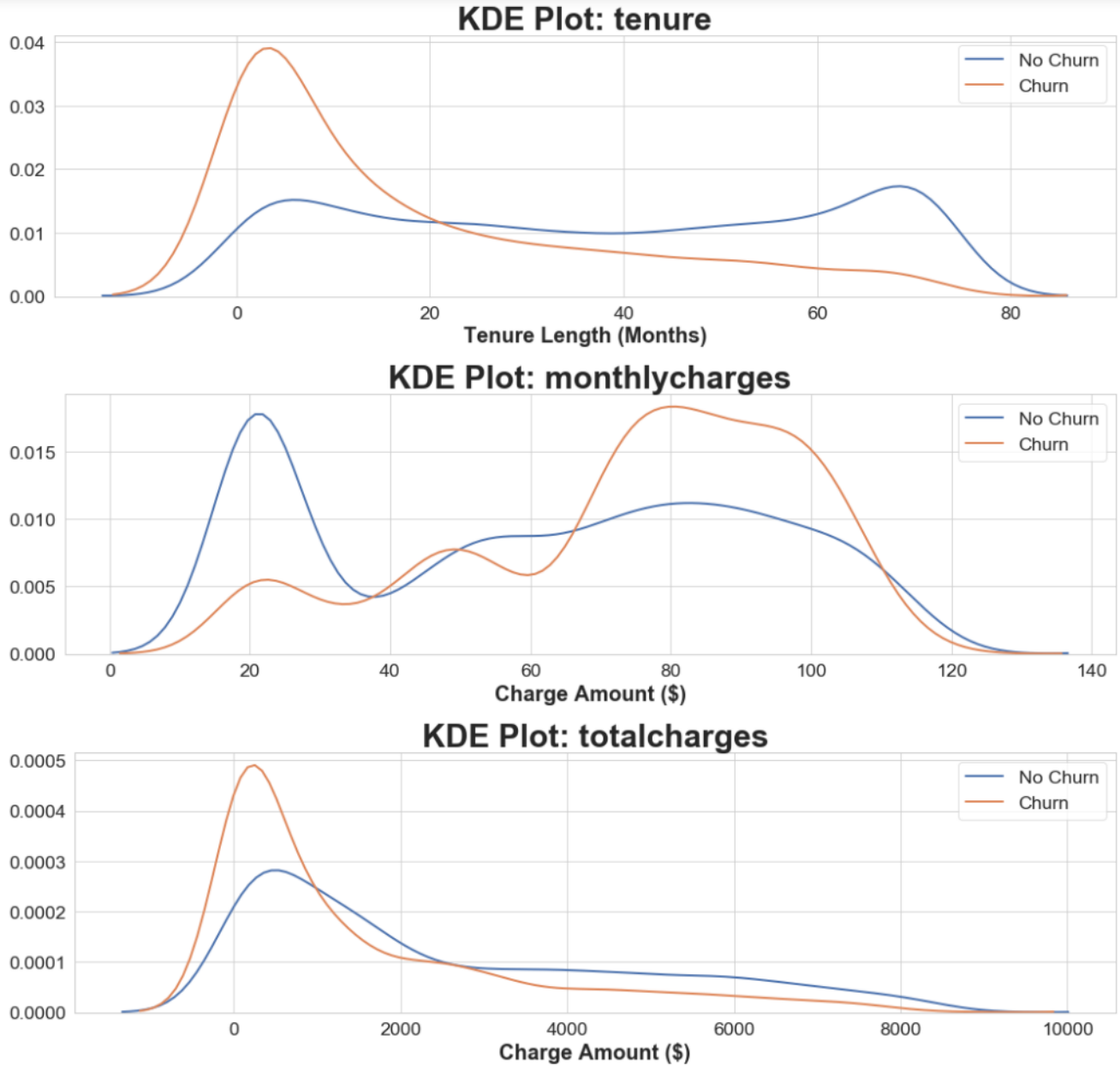
Target

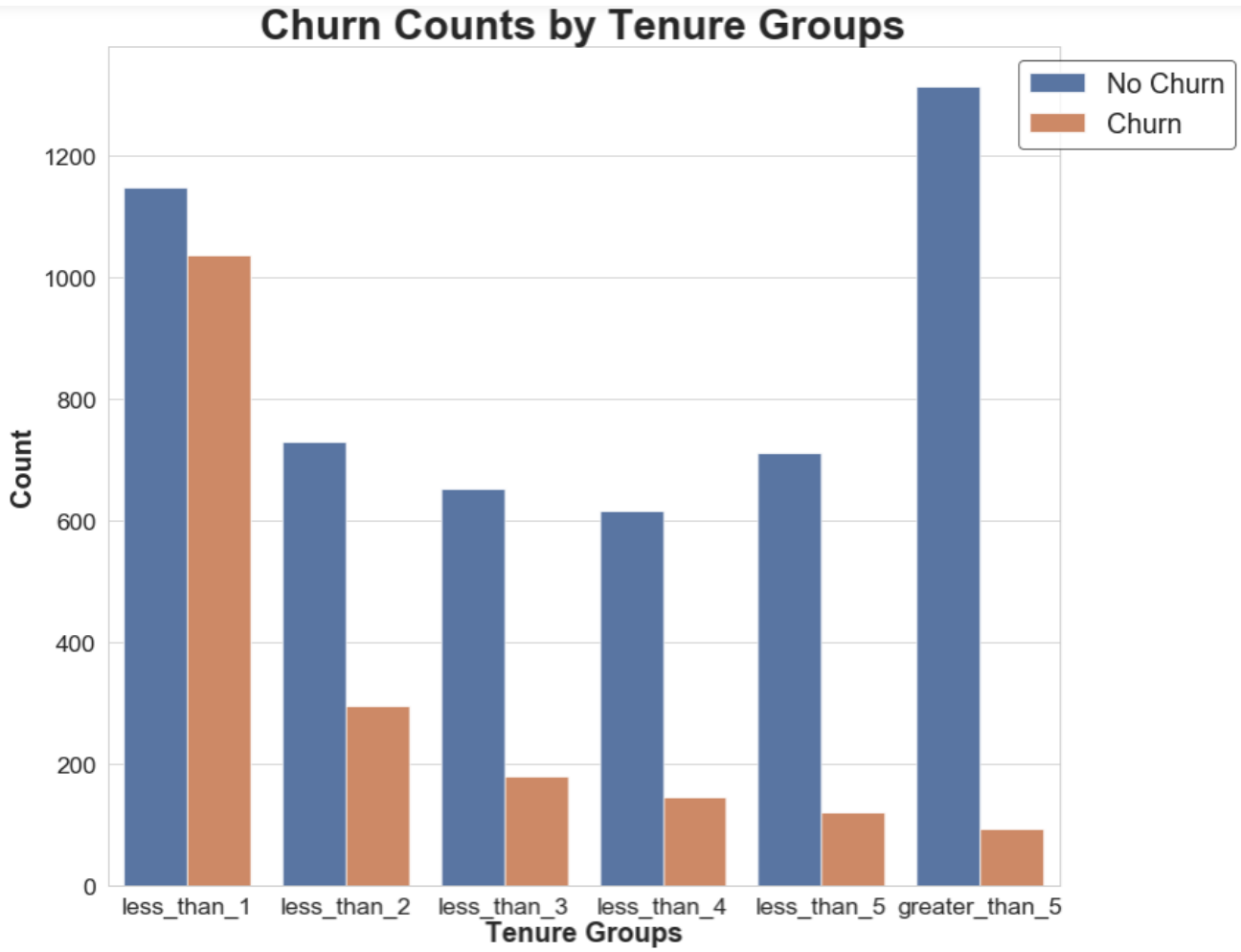
We can see from the pie chart on the left, about 27% of the Telcom customers from our dataset end up churning. This does seem like a rather high amount, but since I do not work for Telcom and do not have prior telecommunications experience for domain knowledge, I will just take this value for what it is and not read too much into it yet. As this is our target variable, we will use Churn as an element in most of our variables' EDA.

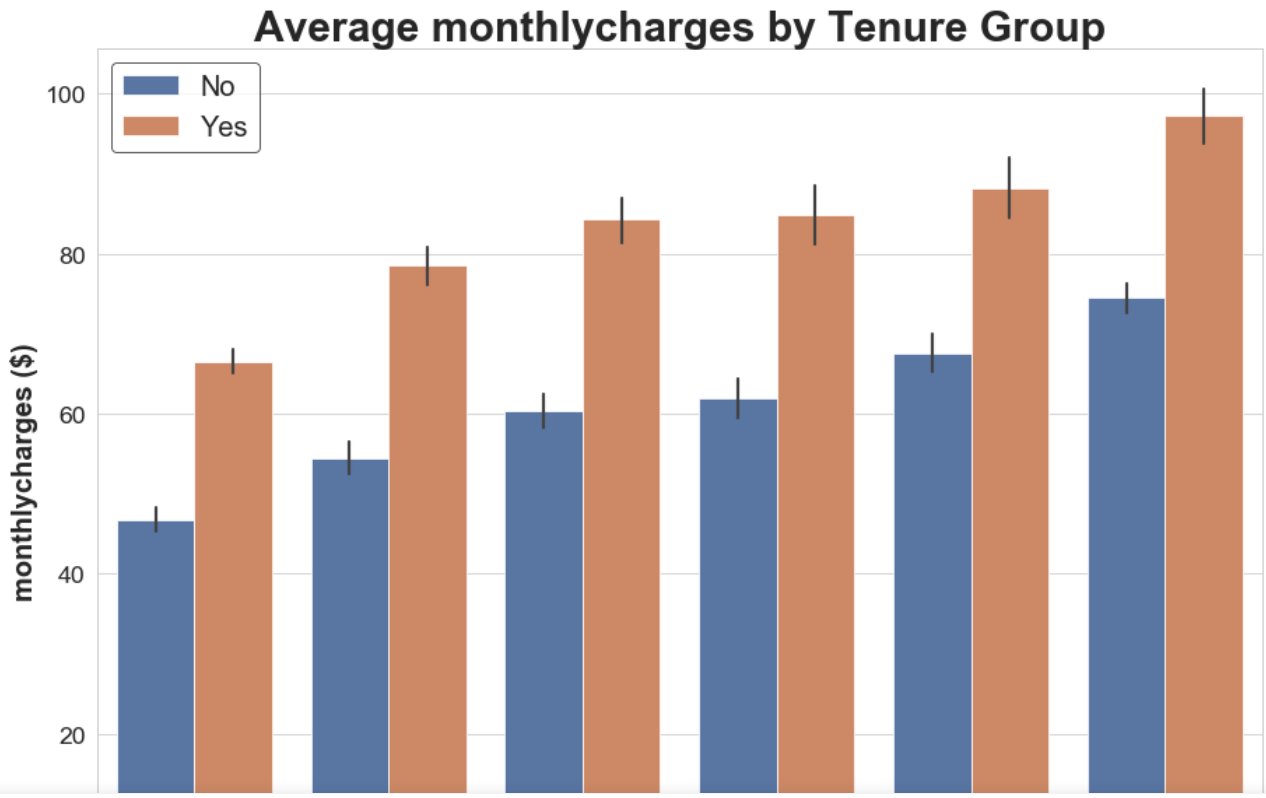
Numerical Features

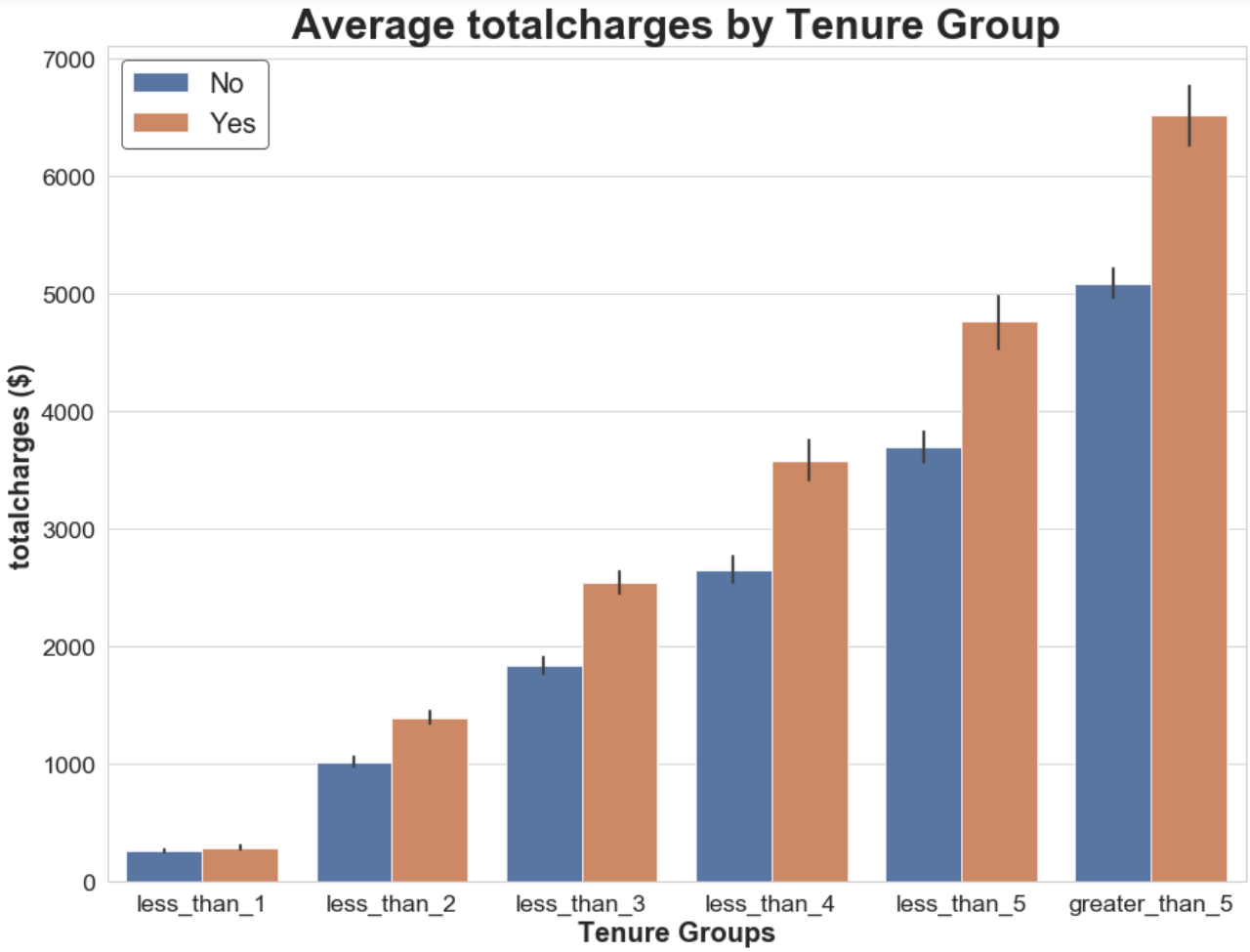
When working with numerical features, one of the most informative statistics we can look at is the distribution of the data. We will use a Kernel-Density-Estimation plot in order to visualize the probability distributions of the relative variables. This graph will show us where there is the highest likelihood of a new data point falling on our dataset. We will create a KDE for al eatures.



To look at our data in a slightly different way to perhaps get some more information, and because 'tenure' is represented by months which can be a bi noisy, I decided to group the customers based off of their tenure.





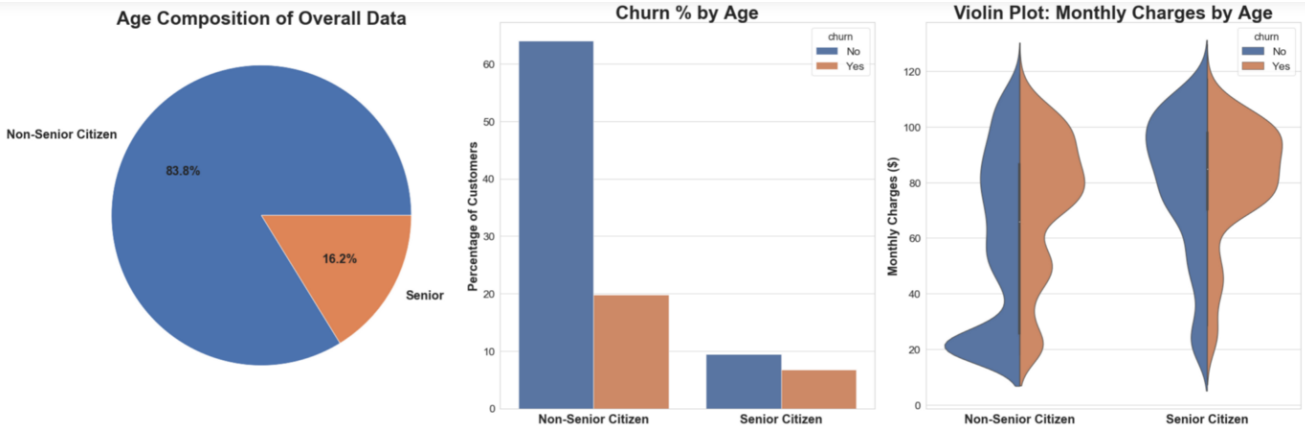


Conclusions:

- Customers who churn have the highest probability of occurring before 20 months of tenure.
- Customers who churn are most likely to have monthly charges greater than $60.
- Generally speaking, the likelihood of a customer churning increases as monthly charges increase
- Distributions for total charges are pretty general, so we will pay most attention to the 'monthlycharges' feature for significance

Categorical Features

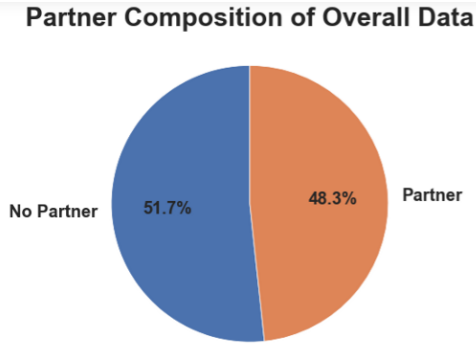
Gender

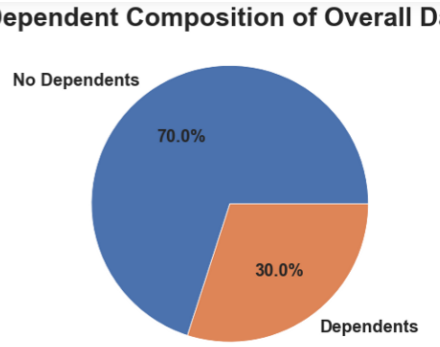


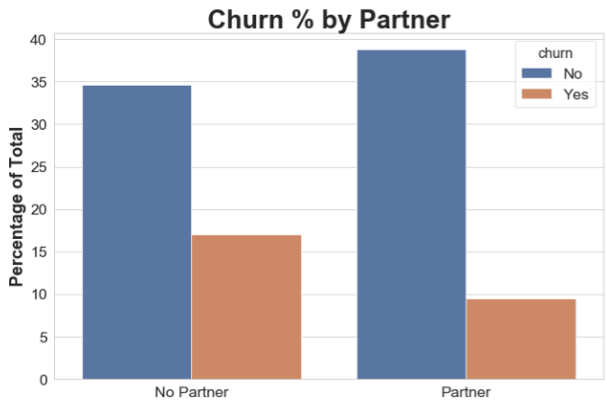
Conclusions:

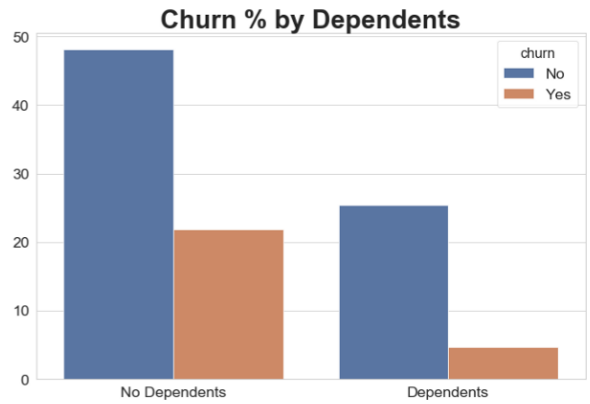
- Our dataset has significantly less senior citizens than non-senior citizens
- *Overall,* more non-senior citizens will churn than senior citizen
- A higher proportion of senior citizens will churn than non-senior citizens
- Senior citizens and non-senior citizens both begin to churn once the monthly charges rise above $60
- Non-senior citizens are most likely to have monthly charges around 20 dollars
- Non-senior citizens will churn are slightly more likely to churn at monthly charges lower than $60 than senior-citizens

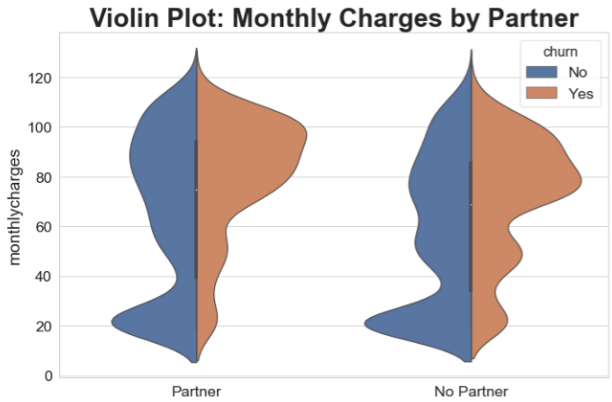
Partner & Dependents

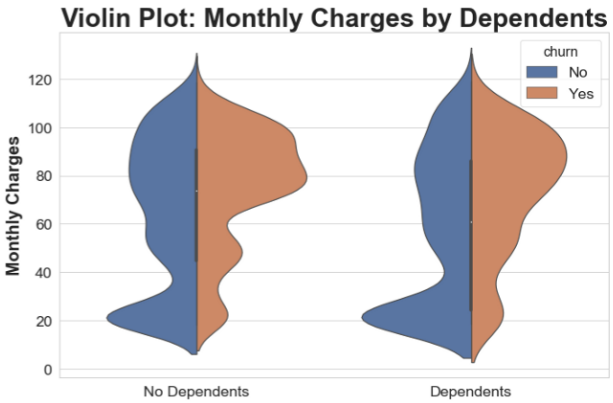








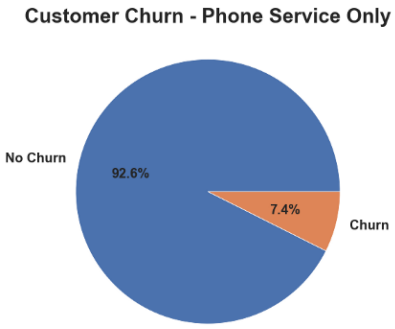


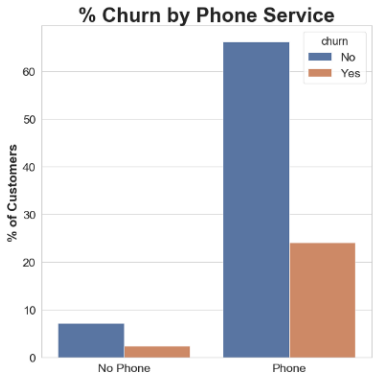


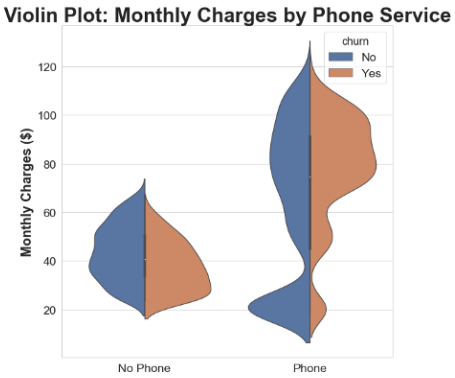
Conclusions:

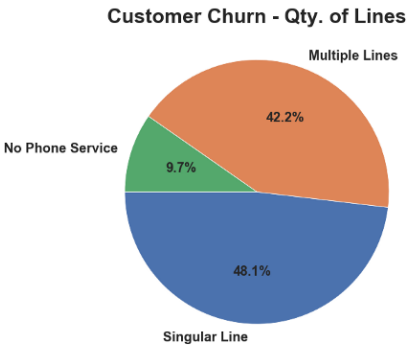
- The dataset is split for customers with partners
- Those without partners churn slightly more than those with partners
- Customers without dependents churn slightly more than those with dependents
- Monthly charges among those who churn and don't churn are pretty similar for both partner values and both dependent values

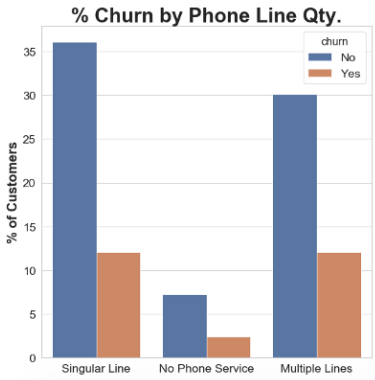
Phone Service & Quantity of Lines

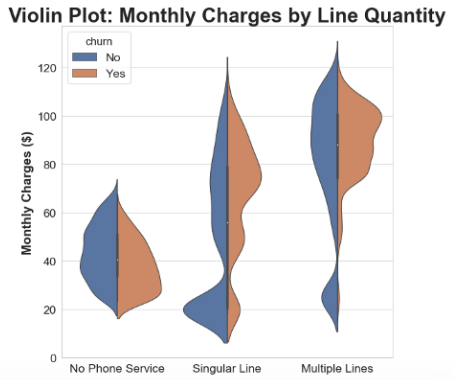








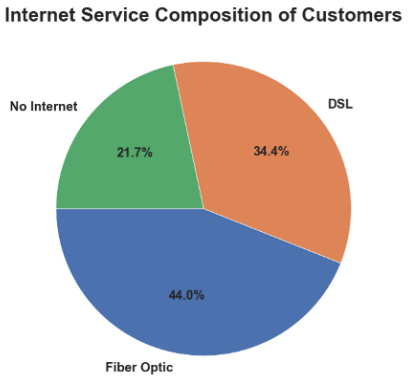


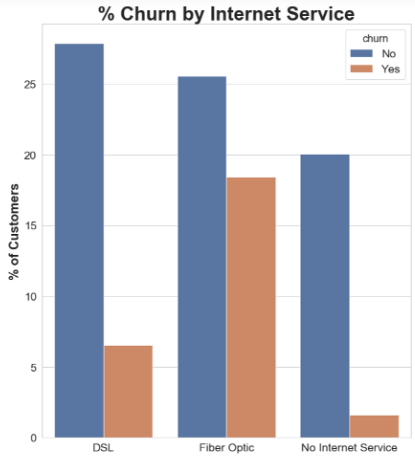


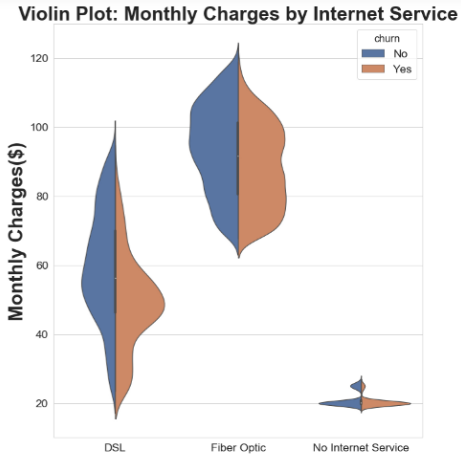
Conclusions:

- Significantly more customers with only phone service will not churn than those other customers
- People with only phone service churn ~25% of the time
- Customers with phone services only pay a higher average monthly charge
- Customers with multiple lines churn at approximately the same rate as those with a singular line
- Customers with multiple lines more frequently pay a higher monthly charge than those with singular phone lines

Internet Service



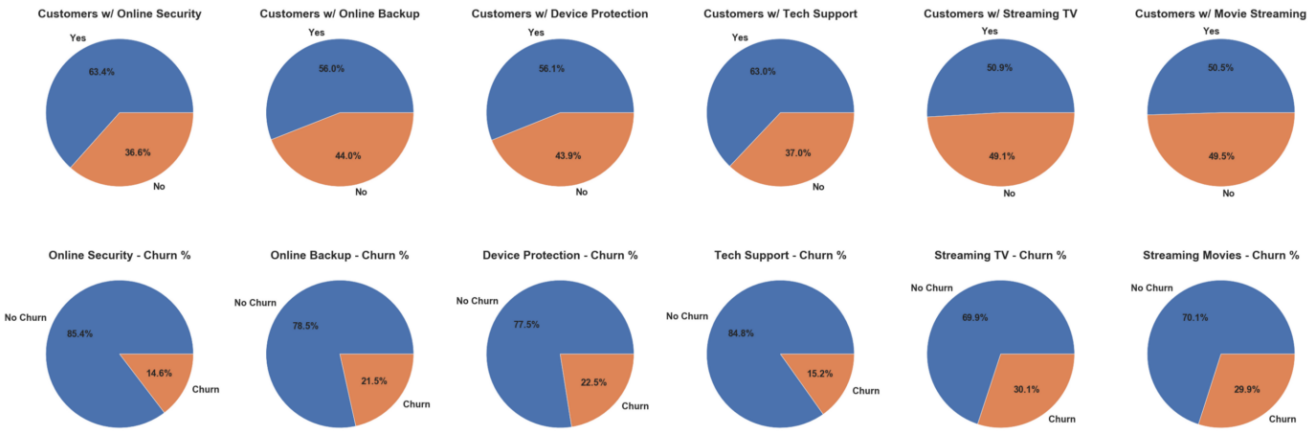


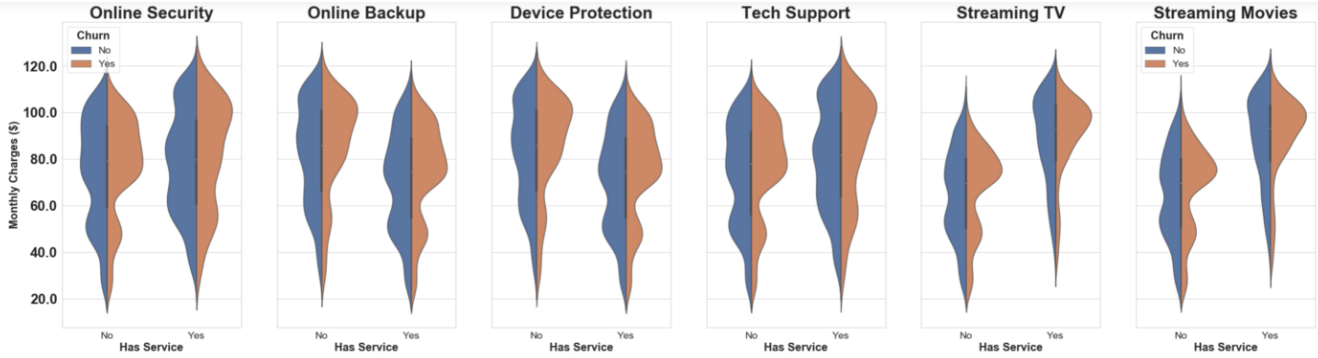


Conclusions:

- Fiber Optic is the most popular internet option
- Fiber Optic Internet Customers churn at significantly proportions than DSL or No Internet customers
- Fiber Optic is a significantly more expensive service, and customers churn slightly more than not when they have this service
- Customers with DSL are most likely to churn when their monthly charges are between $40 and $60.

Add-On Services

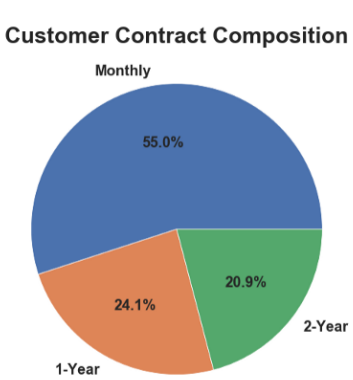


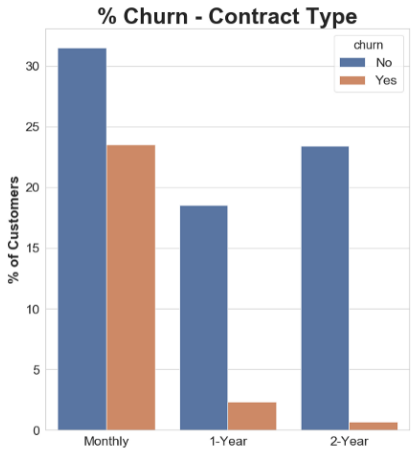


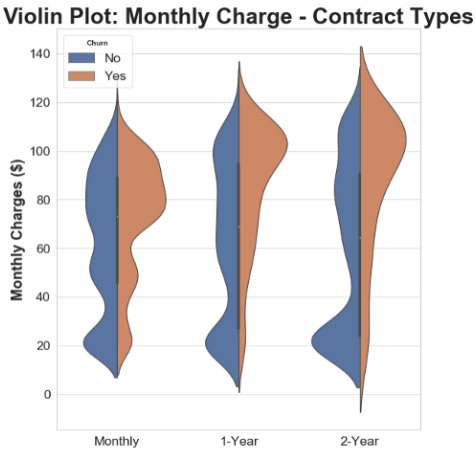
Conclusions:

- Customers with TV streaming and/or Movie Streaming services churn more than all other add-on services
- Churn for customers in most categories will peak around a monthly charge of $100

Contracts

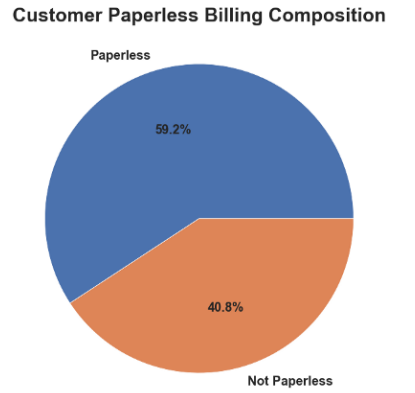


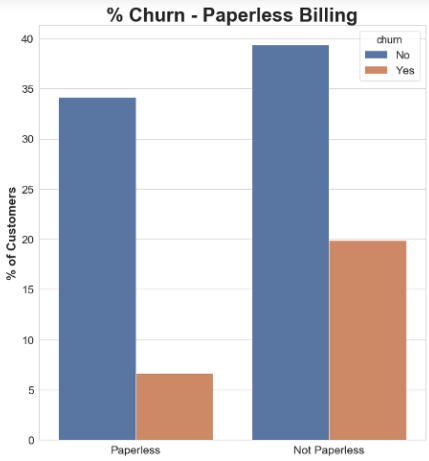


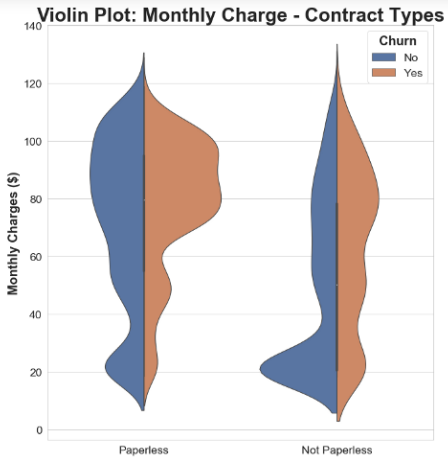


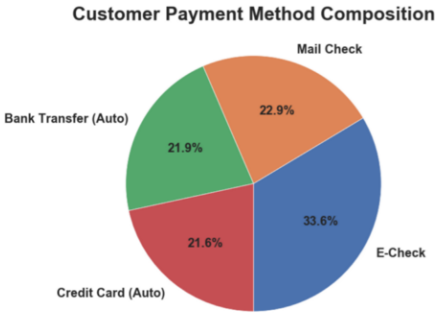
Conclusions:

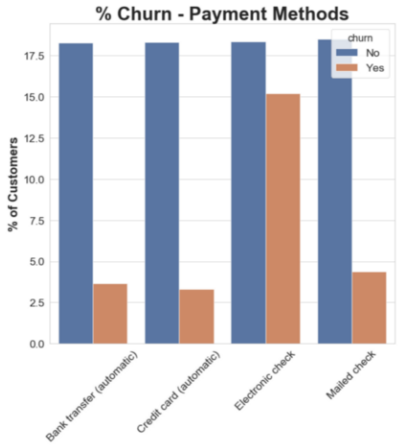
- More than half of customers use a monthly payment option
- Significantly more customers churn on monthly plans
- The longer the plan, the lower the churn rate
- Monthly charges are generally higher the longer the contract is

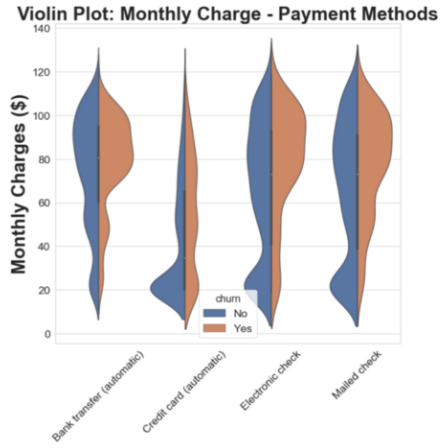












Payments Conclusions:

- Customers with non-paperless billing churn almost 15% more than paperless customers
- Paperless customers churn at similar rates as non-paperless customers when the monthly price is below 60 dollars, once above 60 more paperless customers churn than non-paperless
- Customers who pay with e-check churn more than 10% than customers with all other payment methods
- Customers who pay by credit card have consistent churn rates regardless of monthly charge, whereas customers paying by bank transfer, e-chcek, or mailed check all see an up-tick in churn once monthly charges rise above 60.

project, it is critical to have a concrete foundation of business domain and underlying data understanding. Now that we have taken a look at the key features and the way they interact with each other, we can begin to build our classification models. Take a look at my next blog to see it in action!