Capstone Project 1

Exploratory Data Analysis

By creating a correlation heatmap with all the features in the dataset, I found out that some of the features are strongly correlated. Specifically, OnlineSecurity_No internet service, OnlineBackup_No internet service, DeviceProtection_No internet service, TechSupport_No internet service, StreamingTV_No internet service, and StreamingMovies_No internet service are all strongly correlated to InternetService_No. The correlation coefficient is 1. This is quite self-explanatory because if a customer does not subscribe internet service, he/she would not have online security, online backup, device protection, tech support, streaming TV, streaming movies consequentially. To avoid these collinearities, I removed all the above features (only keep InternetService_No).

Moreover, the feature of PhoneService is negatively correlated with MultipleLines_No phone service This is also quite self explanatory: a customer who does not subscribe phone service would not have multiple lines consequentially. To solve this problem, I removed the feature of MultipleLines_No phone service.

Furthermore, the feature of TotalCharges is strongly correlated to the feature of Tenure. This is also logical since the customers who stayed longer with the company would pay more in total. Similar to above solution, I removed the feature of TotalCharges. In addition, the feature of InternetService_Fiber optic is strongly correlated with the feature of MonthlyCharges, which suggests that whether the Internet is fiber or not would have strong impact on monthly charges. In this case, I removed the feature of InternetService_Fiber to avoid collinearity.

Upon finishing above actions, I re-plot the correlation heatmap and now it looks that there is no strong correlation between any two features.