**Customer Churn Analysis**

Weekly Report

Ahmedabad University

4rier Series

Prof. Mehul Raval

CSE523 - Machine Learning

**Problem Statement**: Customer Churn Prediction and Classification.

**Data Specifications:** The data sample follows Telco, a made-up telecommunications firm, and was obtained from the IBM Developer Platform. It contains information on customer churn. The data includes a "target label" and other relevant features such as demographics, customer account details, and services they've subscribed to.

The sample covers 7043 clients and encompasses 20 different features.

**Libraries used:**

* Numpy: used for linear algebra (correlation)
* Matplotlib: used for visualization
* Plotly: used for visualization
* Pandas: used for data processing, and CSV file I/O

For Data cleaning, we have checked if there is data duplication. As this data has no duplications, we checked for missing values. Basically, we checked for which data the value shows null. The columns were converted into lower for ease of reading.

For some of the data columns, we have removed the data. For example, in “total charges,” we were not able to change its type to float as we found there were spaces in some of the data. There were 11 such data rows which were removed and then converted to our desirable type.

For data inferences, we have run every feature present and coordinated the same with the churn rate. This gives us the dependence of the churn rate across the various features provided.

From the plots above, we can conclude that:

* We observe that features such as gender and streaming television services do not affect the churn rate or have negligible effects.
* It is visible that features like paperless billing which should ideally not affect the churn rate, display some sort of churn behavior.
* Multiple line subscribers have a slightly higher churn rate
* Churning is observed more in recent customers or customers who have to pay higher monthly subscription bills.
* Senior citizens show a higher churn rate of around 40%. Across demographical parameters, customers without a partner or dependents seem to have a greater likelihood of churning.
* Customers with tech support subscriptions or online backup subscriptions, or device protection subscriptions are more likely to churn.
* Clients without internet service show lower churn rates.
* Customers with fiber are more likely to churn.
* Customers who contract month to month are more likely to churn. Unlike those who contract every two year

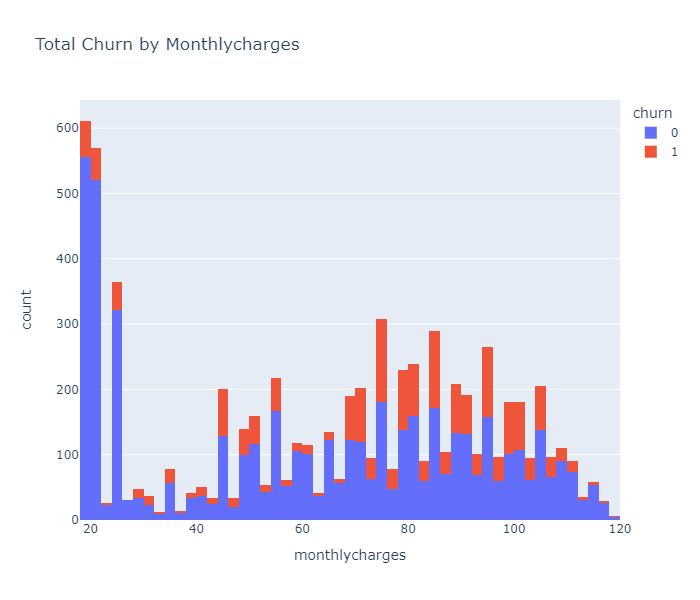


Fig: Total churn by monthly (count vs. monthly charges)

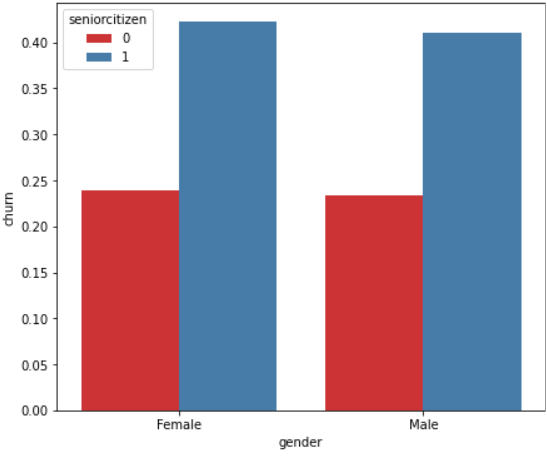


Fig: Percentage of Churn of senior citizen- customers (categorized based on gender)

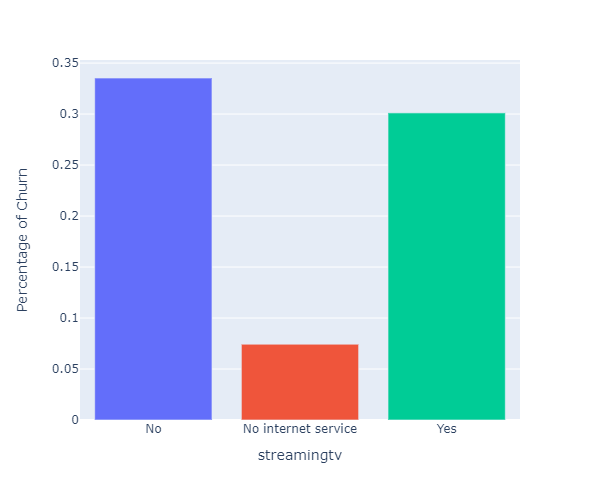
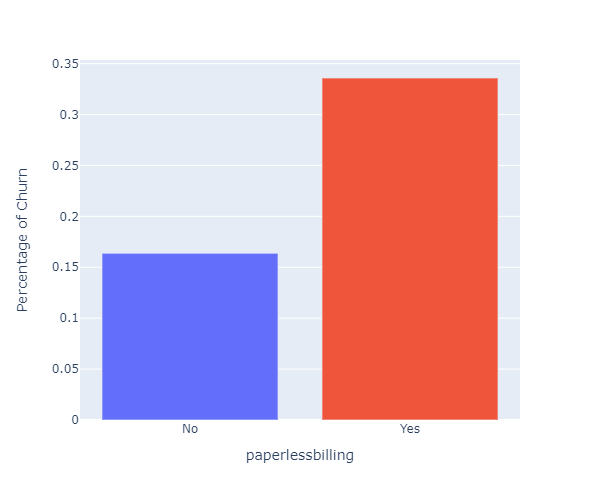
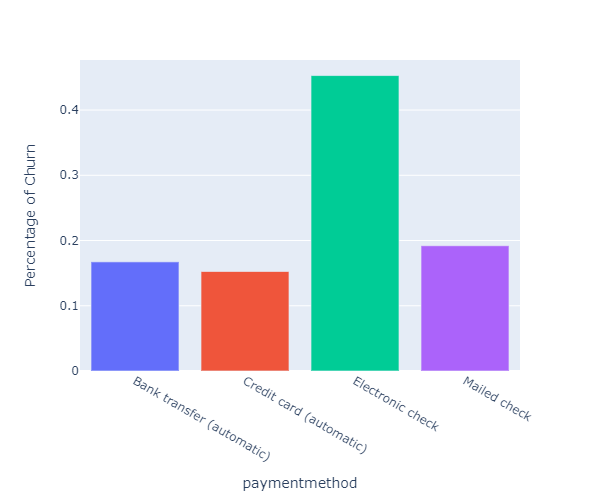
Fig: StreamingTv vs. Percentage of churn

Fig: Percentage of churn vs. paperless billing

**Correlation between features**

When two features are highly correlated, they have a strong linear relationship, and their impact on the dependent variable is nearly the same. Therefore, it's possible to eliminate one of the highly correlated features. In this particular scenario, features such as MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies are highly correlated and can be dropped.

