Assignment 1.1 - COL 341

Feature Engineering

Methods used:

1. **Feature Extraction -** Redundant features were removed from the given list of features. Out of the given 30 features, I removed 8 redundant features. Features like Facility Name and Facility ID are equivalent features, so I removed Facility Name. Other redundant features include CSS Diagnosis Description, CSS Procedure Description, etc. These features are equivalent to their code, so they were removed. I also removed payment typologies as the mode of payment is not going to play a great role in deciding the cost. After feature creation, I also used **PCA (principal component analysis)** for finding out the principal 17 features.

2. **Feature Creation -** Since some features like the **length of stay, hospital county, and hospital Facility ID** will play a greater role in deciding the cost. So, all their inter combination and their squares and higher powers were added to the features. To create more features, I used **Polynomial Features** from the sklearn library of python.

3. **Target Encoding -** Since most of the features belonged to classified features, so I used target encoding using the mean value for feature engineering. Features like emergency indicator, CSS code, APR DRG code, gender, race, etc. belong to particular classes, so encoding them using their priority is more beneficial for better results. I also tried **hot encoding** but it didn't give as good results as target encoding.

4. **Lasso Regression** - After doing all the above steps, **231 features** were created. From those features, 39 were found out to be inactive features using lasso regression as their coefficient came out to be zero. These inactive features were then removed from the list of features.

Selected Features

1. **Length of the Stay:** This is one of the most important features for deciding the cost as living for a greater number of days means the patient is using more resources and so, will be charged more.

2. **Facility ID:** Some of the hospitals charge more in comparison to the others, so this will play a good role in deciding the cost charged by the hospital.

3. **Hospital Area, Hospital County, Zip Code:** Hospitals that are located in metro cities charge more in comparison to the small-town hospitals, thus these are some of the important features.

4. **CSS Diagnosis Code, CSS Procedure Code, APR DRG Code, APR MDC Code:** These codes decide the kind the illness and the kind of treatment the person is being given. So, they are going to play a good role in deciding the cost charges by the hospital.

5. **Emergency Indicator, Severity of Illness, Risk of Mortality:** If the condition is not proper then, he is likely to be in intensive care, so he will be charged more in comparison to others.

6. **Race and Ethnicity:** Depending upon the class, government hospitals may charge differently. Also, foreigners are charged more in comparison to the Indians. Thus, these features will have some role in deciding the cost of treatment.