COL341 : Machine Learning
Assignment 1.1

Gaurav Jain (2019CS10349)

September 10, 2021

1 Feature Creation and Selection

Feature creation and selection are important part of Machine Learning. The accuracy of prediction on unseen data can be improved by adding non-linear features to the linear regression model. In this report, the first section contains the features created using various techniques and the second section contains how these created features are selected.

1.1 Feature Creation

The following techniques are used for adding new features:

1.1.1 Dropping Non-Essential Features

The 30 features present in the original don't have the same relation to the total cost. Thus to improve the accuracy one should drop non-essential features. The following features are dropped for linear regression:

a. The operational certificate number, Facility ID, and Facility Name represent the same information. They all represent a particular hospital. So only one of these three should be present and the rest should be dropped.

b. Hospital Service Area, Hospital county, and Zip code represent the same information of the location of a hospital. Thus only one is kept and the rest are dropped.

c. Consider for example CCS Procedure Code and CCS Procedure Description. Both of them represent exactly the same information. Hence all such features should be removed from the data.

d. The type of payment should not affect the total bill of the hospital so ideally all three features of Payment Typology should be dropped.

e. Emergency Department Indicator is a feature that is easily determinable using CCS Diagnosis and Procedure Description. So it is not needed in the dataset and thus it is dropped.

1.1.2 Adding Combinations of Features

After removing non-essential features, the accuracy of linear regression on unseen data can be improved by adding combinations of essential features with very high co-variance to the total cost.
Length of Stay and *Facility Name* are one of the most important features. The problem can be solved by estimating the total cost just by using just these two features.
So to capture their importance, these two features are encoded together in a single column in the input matrix. There are other combinations as well which can improve the accuracy. However there is a heavy computation involved in this technique so other combinations are not added.

1.1.3 Target Encoding

Target Encoding is extremely helpful in features where there are only qualitative classes and no precise relation can be found. Target Encoding is applied using the mean of total cost for a particular class of a feature.
Target Encoding is applied on all the remaining features as none of them has precise relationship between their different classes. Target Encoding drastically improved the accuracy of the linear regression. Thus, it is added to the input data.

1.1.4 Polynomial Features

Linear features cannot improve the accuracy of linear regression after a certain limit. Thus, it is essential to add polynomial features to the data. Thus, polynomial features are created for all the features present after applying the first three techniques.
The maximum degree of these polynomial features are limited to 2 as increasing the degree even by 1 can produce huge computation burden on the ML model. Polynomial features can be used in predicting highly accurate values to the problem. Thus, they are added to the input data.

1.2 Feature Selection

After applying the above described techniques of creating features, a subset of features are selected from the set of created features. Selection of features is done using *Lasso Regression* where the optimal value of *Alpha* is determined by using *Cross-Validation* on the training set.
The set of active features present after applying *Lasso Regression* is used to select features from the set of all features. The active columns are saved for prediction task.

1.3 Number of Features

- The number of features present after creating features using the above four techniques are 136.
- The number of features selected after applying *Lasso Regression* are 125.