Feature Engineering

Feature engineering is an important tool to increase the performance of a model. In this project, we are using linear regression and so, we want to create features so that they are linearly proportional to the target variable. In this project, our target variable is total costs. I have calculated correlation between different features and Total Costs. Among them the highest correlation was of Length of Stay (LOS) which was 0.69. This seems practical as LOS and Total Costs should have high correlation as if a patient stay for more days, it is more likely patient would have to pay more and if patient stay for less days, then patient would have to pay less.

I thought that LOS and Total Costs are linearly proportional, but if we think for a particular hospital, then LOS and Total Costs would have good correlation as each hospital has its room charges, infrastructure, capacity, etc. and each hospital charged differently. Each hospital's type is denoted by its Operating Certificate Number. So, I have done one hot encoding of Operating Certificate Number and each one hot encoded column denotes a particular hospital type and so, multiply each one hot encoded Operating Certificate Number column by Length of Stay column.

Similar encoding style I have done for features:

- Ethnicity
- Type of Admission
- Age Group
- Payment Typology 1
- APR Severity of Illness Description
- APR Risk of Mortality
- APR Medical Surgical Description
- APR MDC Description
- CCS Procedure Description
- CCS Diagnosis Description
- APR DRG Description

Some features have high number of unique value and so, one hot encode only top 20 most frequent values in the feature.

Then I have done mixed features one hot encoding. In this I have taken 2 features: Health Service Area and Emergency Department Indicator. Then each combination of Health Service Area and Emergency Department Indicator denotes a one hot encoded column. Again, multiply these columns by LOS column. The reason being each Health Service Area in emergency will have different Total Costs/LOS ratio.

Similar encoding scheme, I have done for Age Group and Emergency Department Indicator. The reason being different age group people in emergency will have total costs proportional to LOS.

After this I have drop the unimportant columns whose one hot encoding we have done. Features: Race, Gender and Facility ID were also drop.

Outliers Removal: To improve the performance of a model, we should remove outliers. In this case, I have classified points as outliers which have Total Costs greater than 2,00,000. I have considered non-outliers points while training.

Correlation: In linear regression, we want each feature should have good correlation with target feature. So, after feature engineering is done, for each feature I have tried to find the power of that feature that have high correlation with Total Costs and updated the feature. The power range is [0.1,1.5] with 0.1 step size.

Feature Importance: To find the importance of the feature, I have used LassoLars. In this, each feature is assigned a weight based on its importance. The feature with high weight is more important. But the problem here is we want the feature to be normalized. So, I have multiply weight of each feature by its mean and this product is the importance of the feature.