COL341_Assignment1.1

Swaraj Gawande

September 2021

1 Removing redundant information

I first removed some features in given training data which convey the same information as in the case of Facility Id and Facility name, CCS Diagnosis Code,CCS Diagnosis Description, APR DRG Code and APR DRG Description, APR MDC Id and APR MDC Description.

Then there were some features which gave information which is subset of information given by another feature as in case of Health Service Area which gives information which is subset of what given by Hospital County which is equivalent to Zip-Code as well.

2 Encoding used

Encoding is required for most of the features in the data because the original labels used for representing the data do not convey the weight in which they affect the target variable. For example Hospital County in which it may be possible that hospital which charges extreme fees could have an average label which means if the weight for that label is to fit its requirements the predictions for other counties may get very off.

I first tried one hot encoding which certainly improved the model performance but was not up to the mark as R^2 score for one hot encoding could at most reach 63. So I then tried target encoding which I implemented using numpy arrays and dictionaries which I further use to encode the test set data as for them target variable is not available. Using only the target encoding for all the remaining features after the deletion according to section above the R^2 score for the model reached 79.

3 Feature Creation

Then after encoding are ready I first tried to use sklearn library to create some features first using PCA to reduce the feature count to 17 and then using all

Polynomial of features of degree 2 doing so did improve the R^2 score a bit but many of the the features had zero weights in lasso regression used which implied there was a waste of computation ability and space and evidently it required very long time to complete the iterations so this method was not used.

Then I used correlation coefficient to determine which feature affects the target variable the most. The results from correlation coefficient hold with the field knowledge that the cost would be be mostly affected by Length of stay(d) followed by Refined Diagnosis(APR DRG, t) and Procedures done (CSS Procedure, t2) and then the Facility used(Facility name, n). So using these 4 features I created more features of higher degree using polynomials such as n^2,d^2,t^2,$t2^2$,$n*d$,$t*d$,$t2*n$,$1000*d+n$,$1000*t+t2$ and $1000*n+t$. Target encoding was also done on the newly added features without which adding features was almost useless. After addition and encoding of new features the R^2 score of the model reached about 85 when training was done on train_large.csv and then tested on train.csv.