COL 341

Report – Feature Engineering

While I analyzing the data I found that costs was linearly proportional to the length of stay which it should be. So, I try to create features that are relevant to length of stay. I created one hot encoding for them and then multiply that vector with length of stay column so that I could get data which is linearly proportional to length of stay.

I created the following features:

- One hot encoding of ethnicity merge with health service area. I thought that ethnicity will be correlated to health service area as ethnicity refers to social groups, so they should belong to some area and hence, I merge them with health service area.
- One hot encoding of Type of Admission with APR DGR Code. From definition of APR DGR Code, I thought that doing one hot encoding of type of admission with APR DGR Code will help.
- One hot encoding of Age Group with length of stay. I thought older people normally stay for longer period as compared to young ones. So, cost for older people will depend on length of stay and hence, I did one hot encoding of age group with length of stay.
- One hot encoding of Operating Certificate number with length of stay. From operating certificate number, we can uniquely identify a hospital. So, I thought each hospital has its own cost of stay and so, doing one hot encoding of operating certificate number with length of stay might help.
- One hot encoding of Payment Typology1 with length of stay. I thought people who had stayed for long period would cost more and so there payment method would be different and so, doing

one hot encoding of payment typology1 with length of stay might help.

- One hot encoding of APR Severity of Illness Description, APR Risk of Mortality and APR Medical Surgical Description with length of stay. I thought that if a person is more severe or have a high risk of mortality would stay for long period and so cost will be dependent on length of stay. For, surgery also cost will depend on length of stay.
- One hot encoding of APR MDC Description, CCS Diagnosis Description, CCS Procedure description, APR DRG Description with length of stay. I thought that cost of hospital will depend on length of stay for all of them.
- Merging of Health Service Area and Emergency Department Indicator. I thought that emergency department indicator depends on health service area. For each merging, the cost depends on length of stay. So, did one hot encoding of merge columns of health service area and emergency department indicator with length of stay.
- Merging of Age Group and Emergency Department Indicator. I thought that older people will be taken to emergency more often as compared to young ones. For each merging, the cost depends on length of stay. So, did one hot encoding of merge columns of age group and emergency department indicator with length of stay.
- Now remove columns of Race and gender as they were not relevant for cost of hospital.
- Removed columns whose one hot encoding has been done.
- Also, there was on one to one correspondence between APR DRG Code and APR DRG Description and also of CCS codes and description. So, removed all codes columns as they were not making model better.

- Facility Name and Facility Id have one to one correspondence and I was not getting facility Id column when doing Lasso. So, I drop facility Id.
- I keep rest of features as removing them have no effect on R2 scores.
- Since, number of operating certificate number are 174, so one hot encoding all of them will make my number of features more than 300 and so, I took best 130 features from them based on their frequency on training data and do one hot encoding on them.
- Similarly, for CCS Procedure Description, CCS Diagnosis Description, APR DRG Description, number of unique values were above 200 and so did one hot encoding of their best 20 values based on frequency in training data.
- I also tried doing one hot encoding of remaining features but they were not giving good results.

Name : Raunak Jain

Entry Number : 2019MT10719