COL341 - Machine Learning

Assignment 1 - Linear Regression

In this assignment, we use the "SPARCS Hospital dataset" to predict the "Total Costs" incurred by a patient. We perform feature engineering and feature selection using Lasso Regression, in which the objective is to minimise the following loss function:

$$\mathcal{L} = \frac{1}{n}\|y - X\beta\|^2 + \lambda \sum_{i=1}^{n} |\beta_i|$$

- The dataset (train_large.csv) is of the dimension $1.6 \times 10^6 \times 30$.

Feature Engineering

To improve the accuracy of our regression model, it is crucial that we create and select the best possible features that affect our target, "Total Costs". Looking at the dataset, we make the following observations:

1.) **Birth Weight**: More than 1.4 million data points are zero. This possibly means that the data corresponding to this column is incomplete (as birth weight = 0 is absurd), hence this column is dropped. (See Fig. 1.)

2.) **Identical Columns**: The following pairs of columns are identical:

a) 'CCS Diagnosis Code' and 'CCS Diagnosis Description'
b) 'CCS Procedure Code' and 'CCS Procedure Description'
c) 'APR DRG Code' and 'APR DRG Description'
d) 'APR MDC Code' and 'APR MDC Description'

Hence, we drop the following columns: 'CCS Diagnosis Code', 'CCS Procedure Code', 'APR DRG Code', 'APR MDC Code'.

Other Experiments:

1.) **Weak Correlation with Target**: Features with correlation value lesser than a threshold (chosen as 0.01 here) have been dropped. (See fig. 2.) This led to deterioration in model performance.

2.) **One-Hot Encoding**: Categorical features having less than 10 categories were encoded using the 'One-Hot Encoding' technique. This also led to deterioration in model performance.

```
X['Birth Weight'].value_counts()
```

```
0       1440871
3200      13118
3300      12856
3400      12758
3100      12156
          ...
8600          1
7600          1
7700          1
7900          1
8700          1
Name: Birth Weight, Length: 73, dtype: int64
```

Fig.1: Birth Weight

```
correlation['Total Costs']
```

Health Service Area	0.046579
Hospital County	-0.008275
Operating Certificate Number	0.081673
Facility Id	0.038324
Facility Name	0.007473
Age Group	0.119026
Zip Code - 3 digits	-0.050828
Gender	0.044238
Race	-0.023202
Ethnicity	-0.020007
Length of Stay	0.687518
Type of Admission	-0.021247
Patient Disposition	0.121898
CCS Diagnosis Code	-0.078249
CCS Diagnosis Description	-0.020052
CCS Procedure Code	0.032062
CCS Procedure Description	0.044778
APR DRG Code	-0.086296
APR DRG Description	-0.005759
APR MDC Code	-0.061414
APR MDC Description	-0.077908
APR Severity of Illness Code	0.292057
APR Severity of Illness Description	-0.248220
APR Risk of Mortality	-0.202012
APR Medical Surgical Description	0.236112
Payment Typology 1	0.022999
Payment Typology 2	-0.031003
Payment Typology 3	-0.006134
Birth Weight	-0.112692
Emergency Department Indicator	-0.005839
Total Costs	1.000000

Fig.2: Correlation with Target

After dropping the above columns, we use sklearn's 'PolynomialFeatures' to create polynomial features of degree = 2. This is to increase the flexibility of the model.

Feature Selection using Lasso Regression

After performing the aforementioned feature engineering, we ended up with a matrix of size $1.6 \times 10^6 \times 377$. We use sklearn's LassoLars to train the model. The regularisation parameter, λ, is found using cross-validation. This gives $\lambda = 0.001$ (r-score used as metric)

57 features (out of 377) are chosen by the model. These features are:

```
['APR MDC Description', 'Ones Ethnicity', 'Health Service Area Length of Stay',
'Health Service Area APR MDC Description', 'Health Service Area APR Medical Surgical Description', 'Hospital County^2',
'Hospital County Zip Code - 3 digits', 'Hospital County Length of Stay', 'Hospital County APR MDC Description', 'Operating
Certificate Number Zip Code - 3 digits', 'Operating Certificate Number Length of Stay',
'Operating Certificate Number APR MDC Description', 'Operating Certificate Number APR Medical Surgical Description',
'Facility Id APR MDC Description', 'Facility Id APR Severity of Illness Description', 'Facility Name Length of Stay',
'Facility Name APR MDC Description', 'Age Group Length of Stay', 'Age Group APR Severity of Illness Description',
'Age Group APR Risk of Mortality', 'Zip Code - 3 digits APR MDC Description', 'Gender APR MDC Description',
'Gender APR Severity of Illness Code', 'Gender APR Medical Surgical Description', 'Race APR MDC Description', 'Race APR
Medical Surgical Description', 'Ethnicity APR MDC Description', 'Length of Stay^2', 'Length of Stay Type of Admission',
'Length of Stay CCS Diagnosis Description', 'Length of Stay APR MDC Description',
'Length of Stay APR Severity of Illness Code', 'Length of Stay APR Severity of Illness Description',
'Length of Stay APR Risk of Mortality', 'Length of Stay APR Medical Surgical Description',
'Length of Stay Emergency Department Indicator', 'Type of Admission APR MDC Description',
'Patient Disposition APR MDC Description', 'Patient Disposition APR Severity of Illness Description',
'CCS Procedure Description APR MDC Description', 'APR DRG Description APR Medical Surgical Description',
'APR MDC Description^2', 'APR MDC Description APR Severity of Illness Code', 'APR MDC Description APR Risk of Mortality',
'APR MDC Description APR Medical Surgical Description', 'APR MDC Description Payment Typology 1',
'APR MDC Description Payment Typology 2', 'APR MDC Description Payment Typology 3',
'APR MDC Description Emergency Department Indicator', 'APR Severity of Illness Code APR Medical Surgical Description', 'APR
Severity of Illness Code Payment Typology 2', 'APR Risk of Mortality^2', 'APR Medical Surgical Description^2',
'APR Medical Surgical Description Payment Typology 1', 'APR Medical Surgical Description Payment Typology 2',
'APR Medical Surgical Description Payment Typology 3', 'APR Medical Surgical Description Emergency Department Indicator']
```

OBSERVATIONS

- 14 out of 57 features have 'Length of Stay' as one of the contributing variables. This is because 'Length of Stay' has a strong correlation with the target variable, 'Total Costs' (See Fig 2 for exact value). Hence, this is an important predictor.

- The average r2-score of the model over 25 iterations comes out to be 0.6589.