COL774 Assignment - 1.1

Divam Manchanda - 2022ME21336

Yash Bansal - 2022CS51133

Step 1: Dropping Similar Features

Features representing similar or the same things were grouped, more specific ones of these features were kept, and the remaining were dropped from the model as follows:

Cluster 1:

Hospital Service Area, Hospital County, Facility Name, Permanent Facility ID:

- All represent, in some sense, the Location of the hospital
- Facility Name / Permanent Facility ID - the most specific indicator, encompasses any effect the Hospital Service Area and Hospital County may have.
- Facility Name arbitrarily selected in place of Permanent Facility ID. (Won't make a difference as either one would be encoded later)
- Therefore, Facility Name was kept as a feature, and the remaining features from the cluster were dropped

Cluster 2, 3, 4, 5, and 6:

2) CCSR Diagnosis Code, CCSR Diagnosis Description

3) CCSR Procedure Code, CCSR Procedure Description

4) APR DRG Code, APR DRG Description

5) APR MDC Code, APR MDC Description

6) APR Severity of Illness Code, APR Severity of Illness Description

- In each case, the Code and Description are one-to-one mapped to each other and represent the same thing. The description is kept as a feature and is encoded, and the code is dropped in each case.

Other Dropped Features:

- Operating Certificate Number
- Zip Code - 3 digits

The model was trained once with and once without both of the above features, and a near 0 change was observed in the predictions post-training.

Thus, it was inferred that both the above features have a very low correlation with the target variable cost, and hence, both these features were dropped.

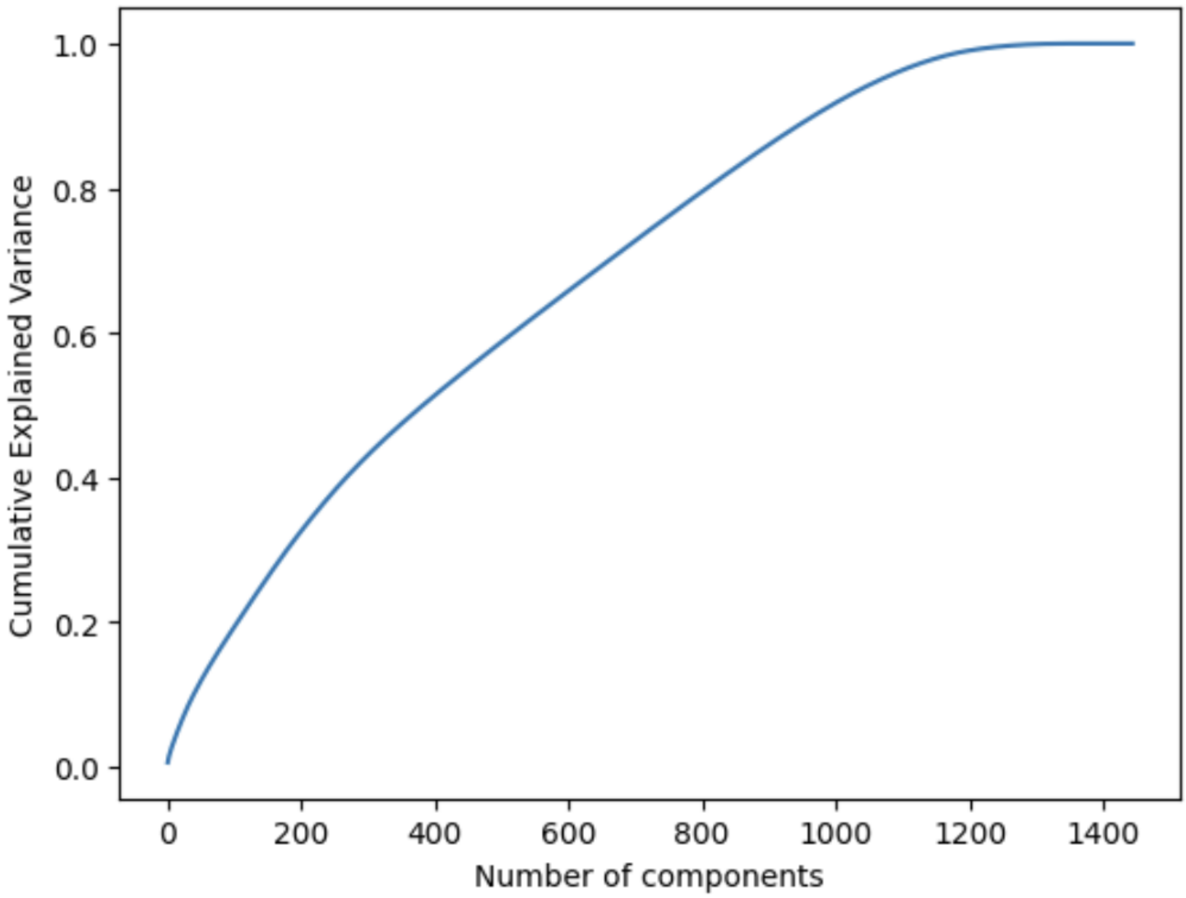
Remaining features:

- Facility Name
- CCSR diagnosis description
- CCSR procedure description
- APR DRG description
- APR MDC description
- APR severity of illness description
- Age Group
- Gender
- Race
- Ethnicity
- Type of Admission
- Patient Disposition
- APR Risk of Mortality
- APR Medical Surgical Description
- Payment Typology 1
- Payment Typology 2
- Payment Typology 3
- Birth Weight
- Emergency Department Indicator

<u>For the remainder of this report, any feature that splits the data into categories (such as Gender/Ethnicity/APR descriptions, etc) is referred to as a "Classifier Feature" with "n number of categories"</u>

One hot Encoding creates more features out of a classifier than target encoding, and it allows you to assign one weight to each category (unlike a single weight for the entire classifier as in target encoding). Thus, one-hot encoding allows each category classifier to individually control the algorithm's decision, and hence allows for better fine-tuning.

So, we initially one-hot encoded all our features, yielding a total of 1445 features in our algorithm. We ran PCA on these 1445 features and computed the number of Principal Components required to maintain 95% of the Variance in the data



For 98% and 95% variance conservation, 1153 and 1068 principal components were required, respectively.

With 300 PCs, we were only able to capture about 43% of the variance in our data

Since the required number of PCs was a substantially bigger number than the 300 feature cap we were supposed to maintain, we tried to do away with one hot encoding for classifiers with a large number of categories.

So, we directly implemented one-hot encoding on any classifiers with less number of categories.

For features with a large number of categories (in the hundreds), we trained our model multiple times, with each feature once target encoded and once one-hot encoded, so as to see if we lose a significant amount of information from the training data in case we target encode the classifier.

We also added a dummy variable (b) - the constant multiplied by $X_0 = 1$ in the gradient descent equation to account for any intercept in the data.

Post all this:

What we did to each of the features

- Facility Name - Target Encoded → 1 feature
- Patient Disposition - Target Encoded → 1 feature
- CCSR Diagnosis Description - Target Encoded → 1 feature
- CCSR Procedure Description - Target Encoded → 1 feature
- APR DRG Description - target encoded → 1 feature
- APR MDC Description - target encoded → 1 feature
- Birth Weight - left as it is (continuous variable, not classifier hance no need to encode)
- Emergency Department indicator - Yes: 1, No: 0 → Binary Encoded
- Age group - one hot → Age Group_1, Age Group_2..., Age Group_5.
- Gender - one hot → Gender_1, Gender_2, Gender_3

- Race - one hot → Race_1, Race_2, Race_3, Race_4
- Ethnicity - one hot → Ethnicity_1, Ethnicity_2, Ethnicity_3, Ethnicity_4
- Type of Admission → one hot encoded similarly into 5 features
- APR Risk of Mortality → one hot encoded similarly into 4 features
- APR Medical Surgical Description → APRMDS_1, APRMDS_3
- Ten one-hot encoded features for each payment typology

- Lastly, each target encoded feature was multiplied into every other target encoded feature to create new features in an attempt to map any corelative effect that they might have

Thus, Finally 106 features so formed were used to train the model