Project 5

1. Domain and Data

The Data for this project is an artificial dataset with no domain specific relation.

2. Problem Definition

We will try to find best method for feature engineering in order to reduce the dimension of the problem.

3. Solution Definition

We will examine the models' behavior and results when there is no feature reduction as benchmark in first step.

In the second step we will tune our Logistic Regression model to eliminate some features using regularization by implementing l1 penalty and lower C values.

At the third and final step we will combine use of a feature selection model like Select K-Best with regularization to find the best approach for identifying important features and eliminating non-important ones.

4. Metric

Mean accuracy of the model is the metric for deciding if the model performing well and selected features are the important ones. Also the coefficient absolute value threshold for considering a feature important is set to 0.001.

5. Benchmark

6. Results from step 1

Using C value of 1000 has reduced any effect of regularization so potentially all features are contributing in the model. Number of features with absolute value of coefficients more than 0.001 (a metric for deciding the importance of the feature) has been 496, and the accuracy of the model for test set was 0.57.

7. Results from step 2

In this step The Logistic Regression model has been implemented using l1 penalty and lower C values.

The number of important features (the ones with coefficients with absolute values greater than 0.001) will decrease by choosing lower C values. And at some point we start to lose some important features due to strong effect of regularization. Although the accuracy of the model is still improving by lowering the C value but that's not due to eliminating unimportant features.

At our first attempt, C values are increased by power of ten starting at 0.001 and up to 1000, the results show that for C values over 0.01, regularization doesn't eliminate many of the features and for C values less than or equal to 0.01 regularization have eliminated almost all the features.

So for the second attempt C values are assigned from range (0.01, 0.03, 0.0005) which consists of 40 different values. In this case we can see smaller changes in number of features selected. Highest test score (0.612) is observed at C values around 0.0162, but this is equivalent of choosing just two features which is not favorable, the next best test score (0.610) observed at the C value of 0.0225 and it keeps 7 features as important features which is more suitable for our case.

The selected features are, feat_004, feat_048, feat_088, feat_205, feat_317, feat_424, and feat_475.

8. Results from step 3

In this final step features are filtered using K-best selection model and then a grid search is used to run the model for different set of model tuning parameters.

The final results was selecting 8 features with the mean accuracy score of 0.616.

The selected features are, feat_064, feat_105, feat_128, feat_241, feat_336, feat_338, feat_442, and feat_475.

9. Comparing results obtained by LogisticRegression and KNearestNeighbors in Step 3

The same process is done using KNearestNeighbors instead of logistic regression, and this time the abest achieved mean accuracy score was much higher (0.9) and number of selected features are 13.

The selected features are, feat_048, feat_064, feat_105, feat_128, feat_241, feat_336, feat_338, feat_378, feat_442, feat_453, feat_472, feat_475, and feat_493.

10. Comparing features identified by Step 2 and Step 3 as important

There is only on common feature between the selected set of features of these two steps.

Considering the minimal improvement in test scores with two almost different set of selected features, we can say the Logistic Regression model is not the best classifier to use with this set of data.

11. Feature Engineering Recommendations for a potential next phase in project

In the next phase, using real data will help determine which model to start with and what range of features are usually best suited for the specific domain and data.