CSCI-P556
Spring 2019
Assignment 4

Virendra Wali

April 2019

1 Introduction

In this assignment, the task is to do binary classification of data. For this classification Naive-Bayes, KNN, SVM and Ada Boost models are using.This data have 6000 training rows and 1000 validation rows with 5000 features. Below are the steps performed for this assignment:

1. Exploratory Data Analysis
2. Baseline Model
3. Feature Engineering
4. Model Building
5. Discussion and Conclusion

2 Exploratory Data Analysis

- IQR :

  [1] Data Models can contain outliers in varying proportions.[2] Outliers can be of different forms viz. Corrupt Data or Input Errors[3] There are various approaches to detect the outliers in data model.[4] One of the approach is using various strategies in statistical method. One such strategy is called as IQR (Inter Quartile Range)[5] Inter Quartile Range or IQR can be used to detect the outliers by taking into consideration a factor K which has a constant value of 1.5[6] The data values or data samples below K factor of 25 percentile and above K factor of 75 percentile of IQR are tagged as outliers.

Model	Cross Validation Accuracy	Validation Set Accuracy
Naïve Bayes	0.9286666666666666	0.922
K Nearest Neighbor	0.9786666666666667	0.966
Support Vector Machines	0.9693333333333334	0.979
Adaboost	0.9613333333333334	0.964

- Checking Missing and Weird Values :

  [1] Data Models or Samples can contain certain deviations like NaN values, infinity values OR missing values.[2] In order to tackle such situations we can use interpolation, imputation or if the data set is massive in size then we can simply delete the rows with NaN values.[3] In this, assignment in order to tackle the above mentioned situation, the isNull() function was used to scoop out NaN / infinity / missing values.[4] However, no such presence of NaN / infinity / Missing values was detected by the isNull() function. Hence the need to use either interpolation or imputation for detecting the above-mentioned foul values did not arise at all.

- Balance Classes :

  [1] Balance Classes are the one in which labels are used to detect the imbalance in the classes. [2] As per binary classification, data values can be grouped under classes with only two values. [3] However, it necessary that both the groupings have same amount of data or the row values sometimes which the case might not hold true. [4] Added to this, training the data set on imbalanced data might produce result with a lower accuracy. [5] The solution to this problem is considering the weight-age of each of the binary grouping and training the data set based on the weights of the each group. [6] However, upon examining the training data, both binary groups were found to be comprising equal number of data values or rows. [7] As a result the need for training the data based on the weight-age of each binary group did not arise at all.

3 Baseline Models

- Post conducting the exploratory data analysis, Baseline model accuracy's for Naive-Bayes, KNN, Support Vector Machine and Ada Boost model were examined without resorting to feature engineering. Although no feature engineering was conducted on the features, base model still produced a much better accuracy.

4 Feature Engineering

- Dropping Uniquely Identifier Columns:

In this step, a correlation matrix was generated followed by the extraction of columns with unique identifiers. To segregate the columns with unique identifiers, a threshold value of 0.95 was set and all the columns with high correlation were dropped. Couple of approaches were tried to engineer the features to get better accuracy.

- Random Forest Feature Selection:

  In order to extract the most important features from sklearn, Random-Forest classifier and select model were put to use. Added to this, a threshold value of 0.000001 was used to select the features and only those qualified for selection which were above threshold. The number features extracted using this method stand at 3987.

- Random Forest Feature Selection:

  Another tree-classifier i.e XGBoost also used used to extract important features. Much similar to Random-Forest, XGBoost classifier and select model method from sklearn were used to extract the most important features. Again, a threshold value of 0.000001 was used for feature selection and only those features lying above threshold qualified for selection.

5 Model Building

- Naive-Bayes: Naive-Bayes is a probabilistic classifier which uses Maximum a posterior model in a Bayesian setting. There are many methods in sklearn to implement Naive-bayes. Initially, GaussianNB was used as a classifier but the method works on assumption that the data is normally distributed and GaussianNB is mainly used when data is continuous. Post using the GaussianNB as naive bayes classifier, the model generated an accuracy of 75.6 percent accuracy. As a result, instead of GaussianNB, a MultinomialNB was used when the data is discrete. MultinomialNB however clocked a 92 percent accuracy of validation of data.

- K Nearest Neighbours: KNN is non-parametric algorithm which does not works on any assumption for underlying data distribution. Also, it does not needs any training data for model generation. However, the entire data Random set is consumed for the testing which makes testing costlier and slower. In this model, a KNN classifier was used along with normalized and non-parameterized data. An accuracy improvement of 0.06 percent was observed for normalized data. Added, for the value k = 5, a best accuracy was recorded.

- Adaboost: Ensemble Learning is a process using which multiple machine learning models (such as classifiers) are strategically constructed to solve a particular problem. In adaboost we construct multiple weak decision

Model	Cross Validation Accuracy	Validation Set Accuracy
Naïve Bayes	0.9286666666666666	0.922
K Nearest Neighbor	0.9786666666666667	0.966
Support Vector Machines	0.9693333333333334	0.979
Adaboost	0.9613333333333334	0.964

Table 1: Dropping Uniquely Identifier Columns

Model	Cross Validation Accuracy	Validation Set Accuracy
Naïve Bayes	0.9293333333333333	0.925
K Nearest Neighbor	0.9786666666666667	0.968
Support Vector Machines	0.9686666666666667	0.979
Adaboost	0.9626666666666667	0.97

Table 2: Random-Forest Feature Selection

stumps and while predicting the label we take voting of all decision stumps to decide the final label. Here I am using sklearns adaboost classifier with n-estimator parameter is 80.

- Support Vector Machines: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.

  Kernel trick is the method of using linear classifier to solve non-linear problems. The kernel function is what is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable. And finally last but very importrant characteristic of SVM classifier. SVM to core tries to achieve a good margin. A margin is a separation of line to the closest class points. A good margin is one where this separation is larger for both the classes. Here, I am using linear kernel for binary classification. .

6 Discussion and Conclusion

Initially, an Exploratory Data Analysis was performed and the data set was analyzed. Next, the data set was tested on Baseline Model without feature engineering and a decent amount of accuracy was clocked. This was followed by

Model	Cross Validation Accuracy	Validation Set Accuracy
Naïve Bayes	0.943	0.943
K Nearest Neighbor	0.984	0.982
Support Vector Machines	0.98	0.977
Adaboost	0.964	0.961

Table 3: XGBoost Feature Selection

Model	Cross Validation Accuracy
Naïve Bayes	0.057
K Nearest Neighbor	0.018
Support Vector Machines	0.023
Adaboost	0.039

Table 4: Error rate

feature engineering which comprised of 3 approaches : Dropping the uniquely identified columns, Random Forest Feature selection and XGBoost. The first two methods did not generate the requisite accuracy however the third method did result in the significant amount of accuracy. However, the highest accuracy was generated by the K Nearest Neighbors i.e. 98.2percent which was relatively high as compared to others.