WARSAW UNIVERSITY OF TECHNOLOGY
Faculty of Mathematics
and Information Science

Advanced Machine Learning

Project of

Feature Selection Methods

Done by:

Amir Ali
Stanislaw Matuszewski

M.Sc. Data Science

12 June 2022

1. Abstract

In this project, we implement five different feature selections techniques which are ANOVA F-value between label/feature – (SelectKBest), FPR test – (SelectFpr), Family-wise error rate – (SelectFwe), Select from Random Forest Classifier model, and Kernel PCA method. And for the prediction methods we experiment different methods which are Logistic Regression, Quadratic Discriminant Analysis, Linear Discriminant Analysis, Random Forest Classifier, Support Vector Machine, XGBoost Classifier.

2. Data Information

2.1 Artificial Dataset:

There are 2000 items. There are 500 predictor variables. The variable to predict is encoded as 0 or 1. There are 600 observations to predict in the validation dataset.

2.2 Digits Dataset:

There are 5000 items, each having 6000 predictor variables. The variable to predict is equal to `4` or `6`. There are 1000 observations to predict in the validation dataset.

3. Data Preprocessing

Both datasets are clean and there is no missing value as well. The only we think we did is feature scaling. Feature Scaling is the most important part of data preprocessing. If we see our dataset, then some attribute contains information in Numeric value some value very high and some are very low. This will cause some issues in our machinery model to solve that problem we set all values on the same scale there are two methods to solve that problem first one is Normalize and Second is Standard Scaler. We apply here Standard Scaler which basically converting data feature to have normal distribution of mean of 0 and variance of 1. Below you can see the formula:

$$z = \frac{x_i - \mu}{\sigma}$$

Figure 1: Standard Scaler [1]

4. Feature Selection Techniques

In this project we experiment with different feature selection methods which are

4.1 SelectKBest

The SelectKBest method selects the features according to the k highest score. By changing the 'score_func' parameter we can apply the method for classification [2] .

4.2 SelectFPR

It controls the total amount of false detections. Parameters score_funccallable, default=f_classif. Function taking two arrays X and y, and returning a pair of arrays (scores, pvalues) [3].

4.3 SelectFwe

Family wise error rate (Fwe) is the probability of incurring at least one false positive among all discoveries [4].

4.4 Select Feature from Random Forest Model

Random Forests are also used for feature selection reason because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees [5].

4.5 Kernel PCA

PCA as a tool for feature selection is to select variables according to the magnitude (from largest to smallest in absolute values) of their coefficients [6].

5. Classification Techniques

5.1 Logistic Regression

Logistic Regression estimates the probability of an event occurring, such as 0 or 1, based on a given dataset of independent variables.

5.2 Quadratic Discriminant Analysis

QDA assumes that each class follow a Gaussian distribution. The class-specific prior is simply the proportion of data points that belong to the class.

5.3 Linear Discriminant Analysis

Linear Discriminant Analysis focuses on finding a feature subspace that maximizes the separability between the groups.

5.4 Random Forest Classifier

Random Forest contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset

5.5 Support Vector Machine

The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

5.6 XGBoost Classifier

XGBoost Classifier provides parallel tree boosting and is the leading machine learning library classification.

6. Experiments

6.1 ANOVA F-value(Select kBest)

The first method depending on ANOVA F values was implemented with the sklearn's implementation of SelectKBest method. In case of artificial dataset the highest accuracy – 85,4% - was achieved by Random Forest algorithm with use of seven selected features. In the second case the accuracies were gradually increasing with the increasing the number of used features. XGBoost classifier turned out to be the best classifier among others.

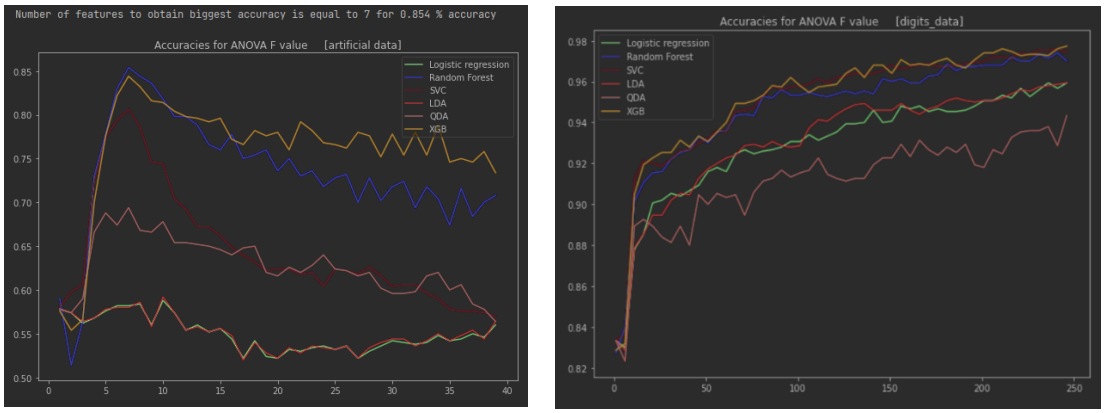

Figure 2: Artificial and Digit Data (ANOVA F-Value)

6.2 FPR test

FPR test results depending on different values of alpha=i/10000. For artificial dataset Random Forest classifier achieved the highest accuracy with use of seven selected variables. Eventually, we decided to use this feature selection in final prediction for artificial dataset. In the second case, 5 out of 6 algorithms returned similarly high accuracy, on the other hand they made a use of more than 1000 features.

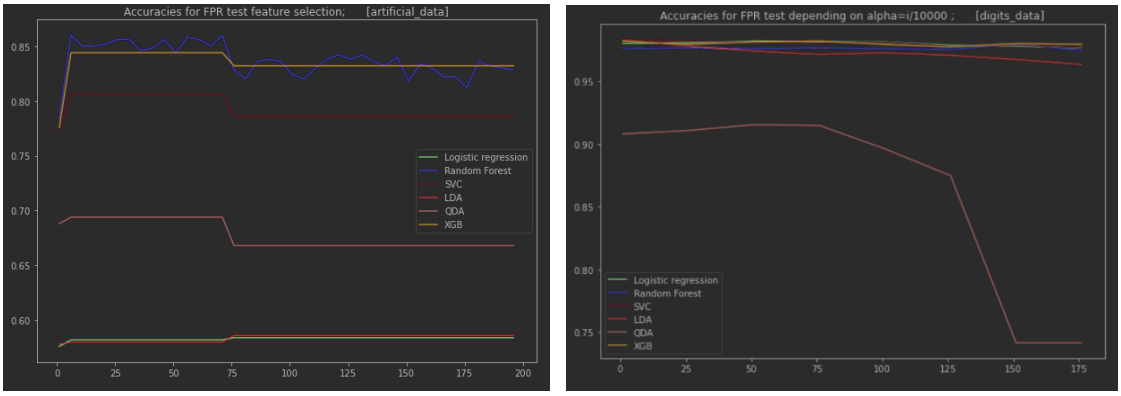

Figure 3: Artificial and Digit Data (FPR test)

6.3 Family-wise error rate

FEW rate results depending on different values of alpha=i/1000. The accuracy metrics obtained for digits dataset seemed satisfactory even in comparison to other feature selection methods, however, it required more than 1500 features. In case of artificial dataset, changing the alpha values across given interval did not change significantly set of selected features.

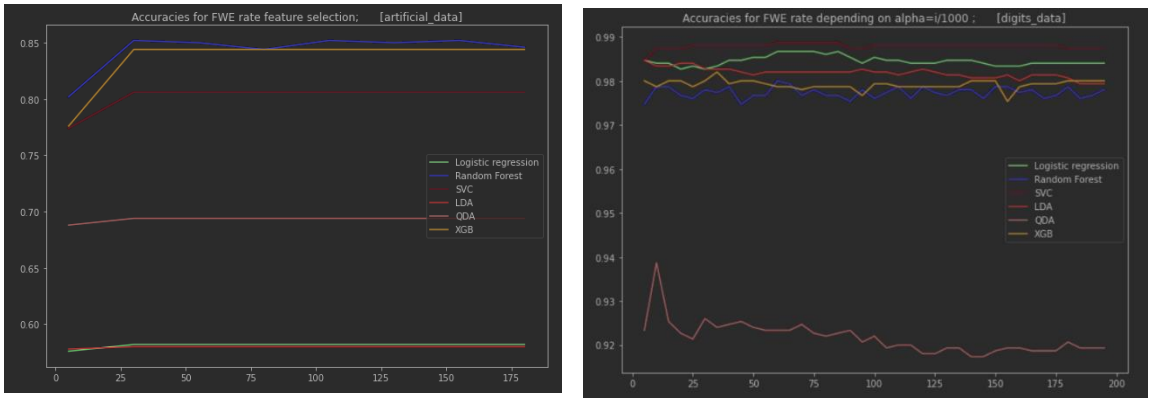

Figure 4: Artificial and Digit Data (FEW test)

6.4 Selection from Random Forest Classifier

With the help of SelectFromModel method provided by sklearn library. We checked this method for various number of used trees from 1 to 800. The achieved results were in both cases lower in comparison the previous ones.

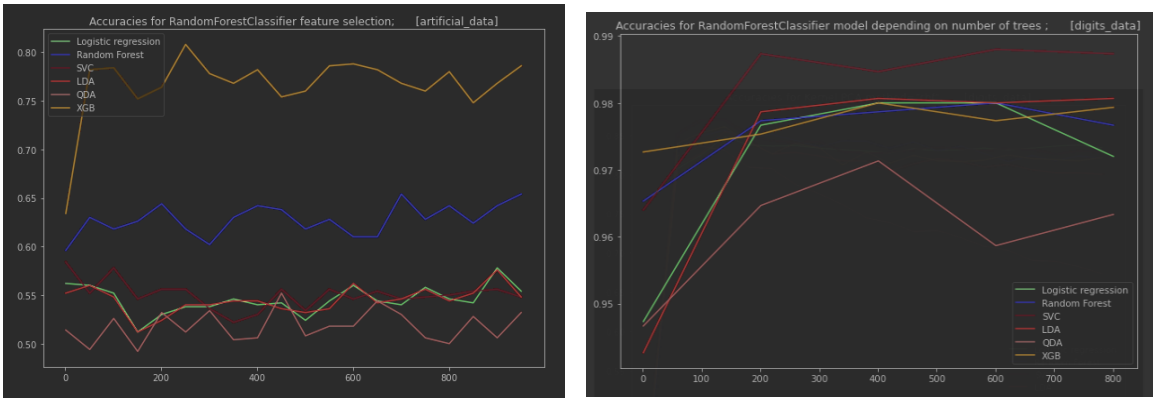

Figure 5: Artificial and Digit Data (Feature Selection with Random Forest Classifier Method)

In case of artificial dataset only XGB classifier crossed the level of 0.8 accuracy. In case of digits dataset the received model used more than 1000 features.

6.4 Kernel PCA for feature Selection

The last approach to feature selection we examined was the use of the Kernel PCA method with Radial Basis Function. Since it creates n principal components instead of selection of custom indicating on relevant variables this approach was strictly additional could not be used for assessing testing dataset.

In case of Digits dataset, for all prediction algorithms we achieved highest accuracy with around 16 principal components. However, their predictions achieved maximum accuracy at 86% and were relevantly worse than the previous methods.

Implementation of kernel PCA on artificial dataset did not any regardless of the different number of principal components used.

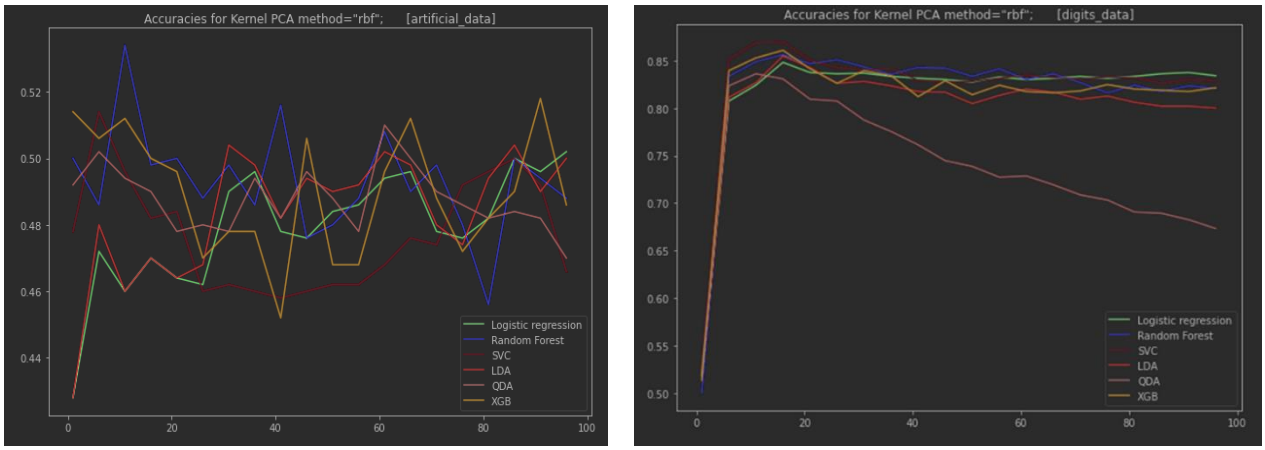

Figure 6: Artificial and Digit Data (Feature Selection with Kernel PCA)

Reference

[1] https://velog.io/@jiselectric/Feature-Scaling-in-Scikit-Learn

[2]https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest

[3]_https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFwe

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFpr

[5]https://scikitlearn.org/stable/modules/generated/sklearn.feature_selection.SelectFrom Model

[6]https://scikitlearn.org/stable/modules/generated/sklearn.decomposition.KernelPCA