What's behind the features?

Janani Sundaresan

***Abstract*—This paper analyses the 'Dataset Challenge' from Symphony Ayasdi and finds the important features through dimensionality reduction technique and how a classifier model is built to predict the class. The final conclusion summarizes all the sections and talks about the future scope and the enhancements that can be made to improve the predictive model that was built.**

I. INTRODUCTION

THe dataset consists of 1554 attributes. "The "class" field is the target variable. The predictive model was built to distinguish class '1' from class '0'.

II. EXPLANTORY DATA ANALYSIS

The data was initially checked for missing values and special characters. The observations with value "NA" were replaced with 0 since there were only two variables with those values. The target variable 'class' had more than half of the observations under '0', nearly 72% thereby indicating an imbalance in the number of observations belonging to both target categories.

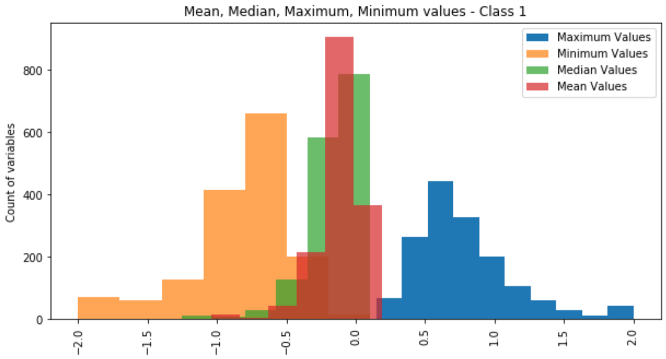

Figure 1: Statistical Distribution for Class 1 data

In Figure 1, the mean, median, maximum and minimum values for all variables with class as 1 were calculated and plotted in the histogram. There is large overlap between mean and median indicating that the number of outliers is less. Also, we can see a smaller number of variables having the extreme maximum and minimum values. Most of the variables are having values between -1 to 1.

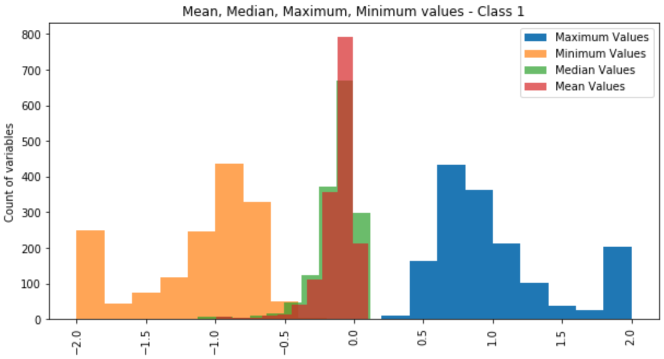

Figure 2: Statistical Distribution for Class 0 data

In Figure 2, the mean, median, maximum and minimum values for all variables with class as 0 were calculated and plotted in the histogram. The mean and median overlap almost 90%. Also, we can see a large number of variables having the extreme maximum and minimum values, thereby indicating a large range of values for the variables. Most of the variables are having values between -2 to 2.

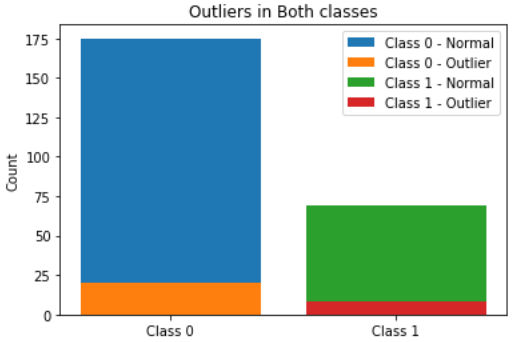

Figure 3: Outliers Comparison

Isolation Forest was used to identify the outliers. Class 0 had 87% normal observations and Class 1 had 95% normal observations.

III. DIMENSIONALITY REDUCTION

Usually, dimensionality reduction technique is used to reduce the number of features when the data has higher number of dimensions than the number of observations. The dimensionality technique used here is Principal Component Analysis. Pca returns the weighted linear coefficients of the features. The feature with highest weight is selected from each component. Since the maximum number of components selected for analysis cannot be greater than the number of observations, 272 was selected as the number of components. Here, three different types of PCA kernels and three different types of scalers were used to find the best combination which distinguishes the classes. In Figure 4, we can see, linear kernel is the best and robust scaler and standard scaler do a good job in classifying when the first principal two components are plotted.

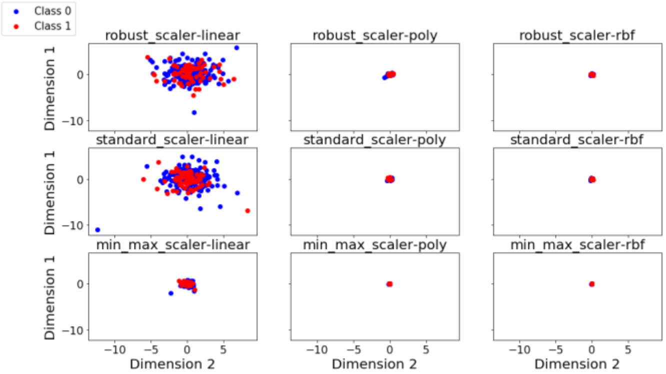

Figure 4: PCA Plot – PC1 vs PC2

IV. FEATURE IMPORTANCE

The feature selection technique used here to identify the best features to predict the target variable is PPSCORE. “The predictive power score is an asymmetric, data-type-agnostic score that can detect linear or non-linear relationships between two columns. The score ranges from 0 (no predictive power) to 1 (perfect predictive power).” [2] Higher the score better would be the prediction. The PPS matrix was plotted and the row in the ppscore matrix for ‘class’ variables gives the best univariate predictor for the target. Usually, the variables with PPS higher than 0.5 would selected as features for prediction. In the below plot, we can see the highest ppscore achieved by any variable is only 0.25.

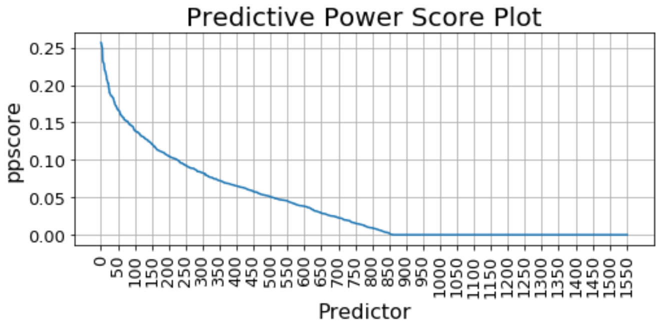

Figure 5: Importance of PC1 variables Plot

The highest ppscore is attained by four variables namely 1487, 1542, 722 and 1359. There are 692 variables that have a ppscore of 0 indicating that they are of no use in predicting the class. From the below plot, we can see the count of variables between different ppscore bins and more variables fall into bins with very low ppscore thereby indicating they are not very helpful in prediction of the class.

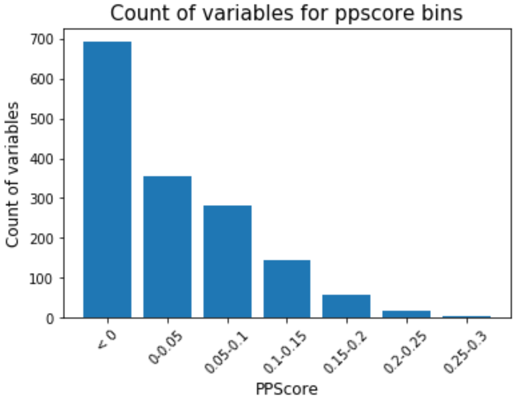

Figure 6: Importance of PC1 variables Plot

The first principal component is strongly correlated with Variables 158, 120, 198, 8, 199. If one increases, then the remaining ones tend to increase as well. These variables are very useful in distinguishing the classes. Furthermore, we see that the first principal component correlates most strongly with the variable 158. The maximum variance explained by all the 272 PC Components together is only 72%. The first component alone explains 14% of variance and the maximum variance of 72% is explained by the first 32 components after which there is no increase in variance, thereby indicating a maximum of 39 features can be considered for the final classification. Some features even have a negative weight indicating there would not be useful in predicting the class.

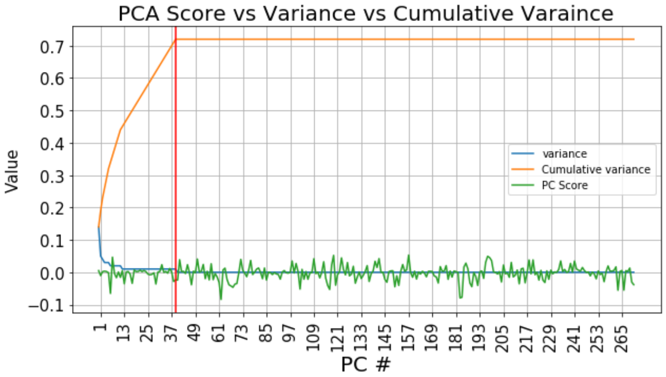

Figure 7: Importance of PC1 variables Plot

V. CLASSIFIER MODEL

In this section, we are going to see how a classifier model was built to predict the target variable. Since this is a classification problem, the ensemble methods and logistic regression models were considered for selecting the best model. Initial data preprocessing was done and the data with all features were used to identify the best model. Accuracy was used as a metric to identify the base model. The analysis showed that the Logistic Regression is the best model. Then dimensionality reduction using PCA was performed to reduce the number of features for different types of scalers. Here three different types of scalers were used – Robust Scaler, Standard Scaler and Min Max scaler. The data was preprocessed using each of these scalers and then their dimensions were reduced. Each scaler selected different number of features.

Hyperparameter optimization was performed based on the features selected by pca using grid search cross validation technique, to find the best set of parameters and the best fit was chosen as the optimal model. The features were selected by linear PCA were used since it had the best accuracy among all other kernels.

It can be inferred from TABLE I below that the Logistic Regression model with linear kernel and robust scaler performs the best among the three.

Scaler	Accuracy
Robust Scaler	.573
Standard Scaler	.682
Min Max Scaler	.719

Table I – Accuracy for different Scalers

VI. CONCLUSION

This report presents different models, different scalers and feature selection techniques that were used to predict heart failure. The same model when tested by keeping all the features, the accuracy increased by 3% thereby indicating the need for further fine tuning which would lead to a better prediction model by carefully considering a better feature set.

References

[1] https://github.com/8080labs/ppscore

Note: Pandas was used only for ppscore calculation since the package does not accept anything other than dataframe as input.