Project Classifymeister End Evaluation

Team Members :- Ajitesh Shree, Sharvil Athaley

Objective :-Report on a study for predicting loan default based on Random Forest Algorithm.

Introduction:

1. This paper is based on using a random forest approach to predicting loan defaults for a P2P platform: Lending Club (Note- Peer2Peer platforms are online platforms that connect lenders and borrowers.)
2. The researchers also used other methods such as decision trees, support vector machines, logistic regression and compared them with the RF approach.
3. Dataset used was obtained from the Lending Club for the first quarter of 2019. It contained more than 115,000 loan data of users with over 102 features for each.

Selection and Engineering of Features:

- The 102 features available were reduced using Recursive Feature Elimination to 30 features with high correlation with the target variable.
- The non-numeric features were on OneHotEncoded and then were feature scaled.
- They then plotted a correlation graph for the 30 features and used it to reduce it to 15, keeping the independent features and removing the highly correlated ones (formula).
- They finally used the Random Forest algorithm to rank the features in importance.
- The researchers found that the data contained a categorical imbalance. Over 98% of the samples were normal while only 2% were defaulters and this could have led to inaccurate predictions, so they used SMOTE(Synthetic Minority Oversampling Technique) to rectify this problem. This method takes k-nearest

neighbors of each minority point and uses them to construct new samples (formula). This is done until the sample size of the minority is suitably large.

Algorithm:

- Random forest is a supervised learning algorithm that is used for classification and regression tasks.
- It is an ensemble learning method that constructs multiple decision trees during training and makes predictions based on the mode of the classes or the mean prediction of the individual trees.
- The decision tree is a tree-like structure where non-leaf nodes represent feature tests, branches represent attribute values, and leaf nodes store categories.
- The decision process starts at the root node, tests feature attributes, selects branches based on attribute values, and continues until a leaf node is reached, which determines the final decision or prediction.
- Gini Index was used to split the attributes in the multiple decision trees.

Evaluation:

- Accuracy, F1 score and AUC value were used to evaluate the RF method and compare it against other methods used.
- Using all of these characteristics the RF method slightly outperformed Decision Tree method and significantly outperformed Support Vector Machines and Logistic Regression methods with an Accuracy of 98%, F1 score of 0.98 and AUC value of 0.983.

Conclusion:

Random Forest algorithm outperforms other methods in predicting loan defaulters for P2P platforms and has a strong ability of generalization.