Big Data Analytics

1.1 Contents

1. Introduction to Data Sets

The purpose of developing the prediction model is to anticipate donors and non-donors using the information. The data collection used to train the prediction model is known as the donor data set. The data set that the predictive model must predict is known as the prospective data set. To accurately forecast both donors and non-donors, the model must be more precise. The TARGET B column in the donor data set indicates whether a person is a donor or not. If it is a donor, the TARGET D column displays the amount that will be contributed. To uniquely identify the person, use the Control Number. For forecasting a donor or non-donor, some columns are crucial, while others are not as crucial. Additionally, several columns have missing data. Due to the significance of the data and the possibility that the forecasts may be affected, certain missing numbers cannot be ignored. The prospective data collection only includes donor information; it lacks goal values. Predictive modeling will be used to make predictions about the data set.

1. Data Preprocessing

Data preprocessing is a phase in the data mining and data analysis process that converts raw data into a format that computers and machine learning algorithms can understand and evaluate. Machines like to process information that is neat and orderly; they interpret data as 1s and 0s. Therefore, it is simple to calculate structured data like whole numbers and percentages. However, unstructured data must first be cleaned and prepared in the form of text and graphics before analysis. A data mining approach called data preprocessing turns row data into a usable and effective form. There are 3 types of data preprocessing,

- Data Cleaning
- Data Transformation
- Data Reduction

The donor data set has missing values for the variables like AGE, INCOME_GROUP, MEDIAN_HOUSEHOLD_INCOME and PER_CAPITA_INCOME. The accuracy of a prediction model depends more on that knowledge. Those vacant vales must be filled in some fashion. There are two options: either fill them in or ignore them. Because values are so important, we cannot dismiss them. Therefore, we must apply the proper technique to fill in the missing values. The

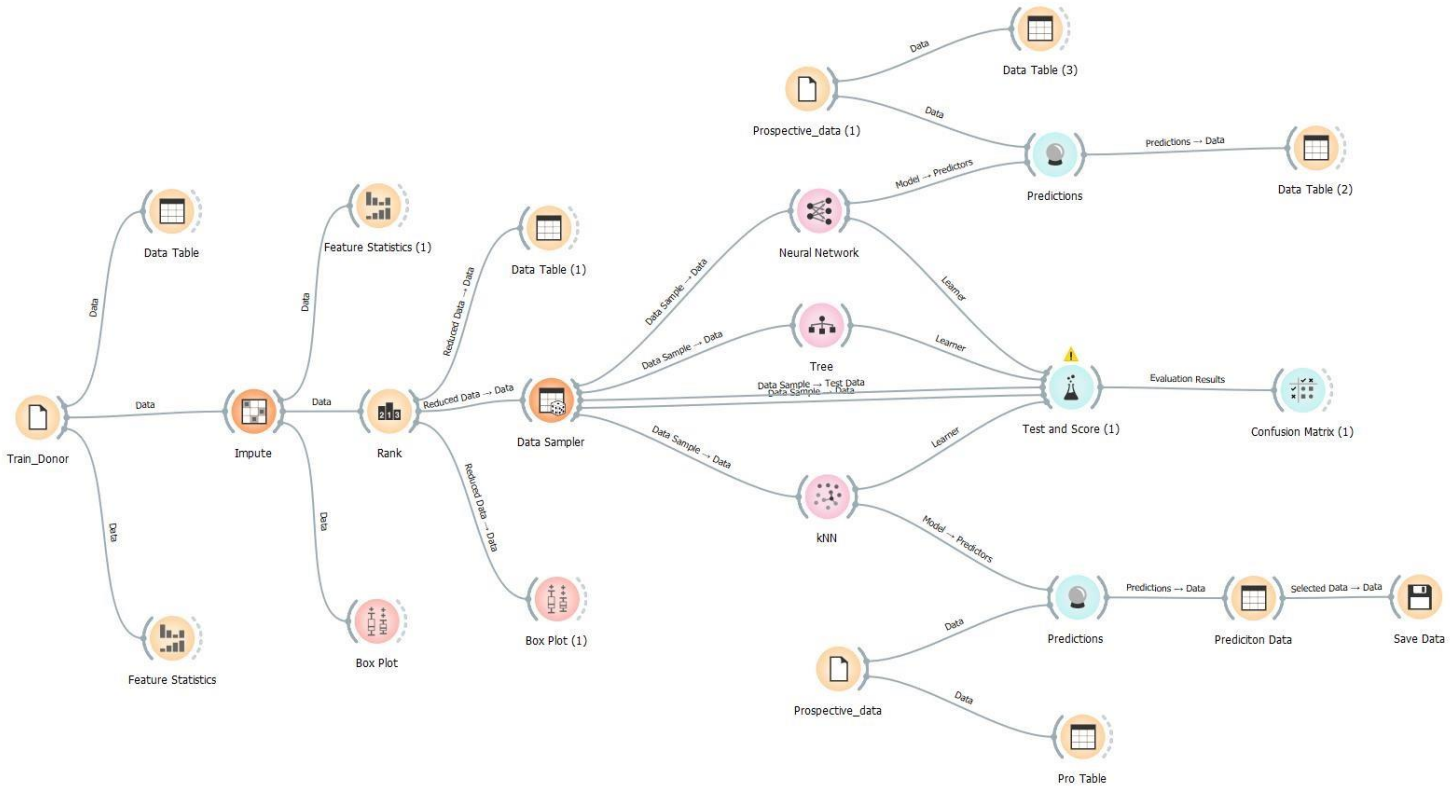
median value of the pertinent column was used to fill in the blanks in the numerical columns for the donor data set.

2. Data Modeling

The target columns of the data collection must be decided before starting the modeling process. Except the TARGET_D and CONTROL_NUMBER all other columns will be features of the data set. The target column used to train the model is TARGET B. The "Select Columns" Widget in the Orange data mining tool can be used to specify the feature values and the target column. There are so many predictive modeling methods that can be used to make the model. KNN, Neural Networks, Decision trees and Random forests are some examples of the models. Each method may have same accuracy level or may not have. By performing all possible predictive model, based on the accuracy of the model, most appropriate model was chosen as the model that fits for the predication. For the donor dataset, The KNN, Trees and Neural network model was performed.

The three models have given three accurate levels for the prediction.

Diagram



3. Evaluation

For the Neural Network accuracy is 73% and for Trees accuracy is 64%. For the KNN, Accuracy is 75%. So, the most accurate models are the KNN and Trees. For Making this predictive model, The KNN was selected as it has highest Accuracy. For the testing of the model, 80% of the data from donor data set was used as training data and 20% used as test data.

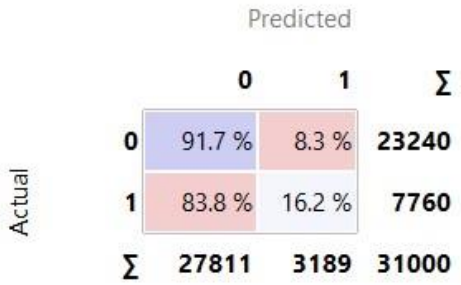


Figure 1 Confusion Matrix of Neural Networks

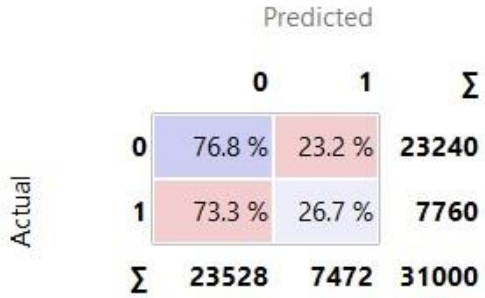


Figure 2 Confusion Matrix of Trees

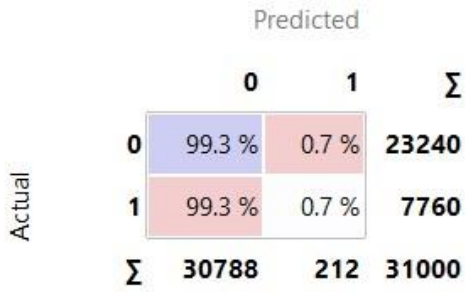


Figure 3 Confusion Matrix of KNN

Considering the confusion matrices of the tree models, The KNN is given the exact prediction of the data set. It has 99% for the True positive and True Negative values. That implies that is has predicted the donors and non-donors nearly 100% correctly than other two methos. That implies the most appropriate model would be KNN method.

4. Conclusion

Using the KNN modeling, the prospective data is to be predicted. Since the accurate of the KNN methos is 99%, the predicted details of donors and non-donors will be 75% accurate.

5. References

https://orangedatamining.com/blog/visualization/ https://www.studytonight.com/post/machine-learning-and-data-visualization-using-orange https://www.jmp.com/en_in/learning-library/topics/data-mining-and-predictive-modeling.html