Name: John Royce C. Punay Course: PSMDSSC-101 - Data Mining
Student no: 2013048

1.) Visit the T.I.P. library and download a paper from the IEEE Digital Library. The paper should be related to supervised learning. Synthesize the paper using the Trend-Problem-Issue-Objectives-Contribution approach.

➔Trend

- This paper is all about the **supervised machine learning algorithms for credit card fraud detection.** Today, the usage of a credit card has become extremely commonplace. Basically, the user makes payment of the total money without the need to carry a large pile of cash. The form of cashless payment is useful but comes with its own risks. The credit card information can be collected illegally and can be used for fraudulent transactions. In addition to this, fraud is increasing dramatically with the expansion of modern technology, resulting in the loss of money. Although prevention technologies are the best way to reduce fraud, given time will usually find ways such measures.

➔Problem

- The card information is prone to fraudulent transactions where the intention is to gain monetary value. To overcome this problem many algorithms can be used to detect any fraud activity. But only one model can provide accuracy and best suited to detect credit card frauds. Specific algorithms based on AI and neural networks are also implemented to provide accuracy in the prediction of fraud. In addition, data mining techniques are also included to create more efficient fraud detection system. Furthermore, fraud detection is a continuously evolving discipline because whenever the detection is in place, criminals will find a way of adapting the strategies.

➔Issues

- The development of new fraud detection methods is made more difficult by the fact that the exchange of ideas in fraud detection is limited. It does not make sense to describe fraud detection techniques in great detail in the public domain, as this gives criminals the information that they require to evade detection. Data sets are not made available and results are often censored, making them difficult to assess. Many fraud detection problems involve huge data sets that are constantly evolving.

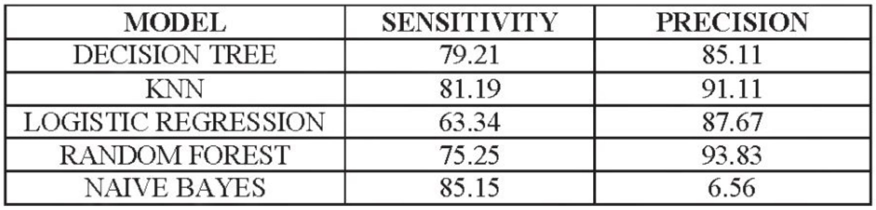
Processing these data sets in a search for fraudulent transactions or calls requires more than the mere novelty of statistical model, and also needs fast and efficient algorithms: data mining techniques are relevant.

➔Objectives

- The objective is to evaluate all supervised machine learning algorithm that determines which is the best fit and suited for fraud detection. The criteria for comparison is the evaluation of performance by measuring sensitivity, precision, and time.
    - Sensitivity
        - This is the measurement of a **false negative** and **true positive**. Below is the formula
            - Sensitivity = (TP)/(TP + FN)
    - Precision
        - This is the measurement of the positive predictions that are **truly positive**, and **false positive**. Below is the formula.
            - Precision = (TP)/(TP + FP)
    - Time
        - This is used for the performance evaluation by calculating the time for the training model and predict the data. The actual time is the approximate time predicting the model.
- Model used
    - Decision Tree
        - The general motive of using Decision Tree is to create a training model that can use to predict the class or value of target variables by **learning decision rules** inferred from prior data(training data).
    - K nearest neighbor
        - In this model, the class label of the test datasets based on the class label of the neighboring training data elements. The similarity between the two elements is measured using Euclidean Distance
    - Logistic regression

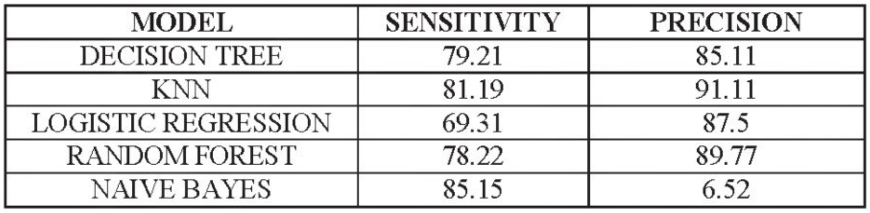
        - It is a statistical model which makes use of a logistic function to model a binary dependent variable.
    - Random forest
        - This model is basically an ensemble classifier, i.e. a combining classifier that uses and combines many decision tree classifiers.
    - Naive Bayes
        - It is a form of probabilistic classifier model; this implies that it has the ability to make predictions for multiple classes at once.

- Observations
    - PERFORMANCE EVALUATION OF VARIOUS MODELS AT THRESHOLD VALUE OF 0.5



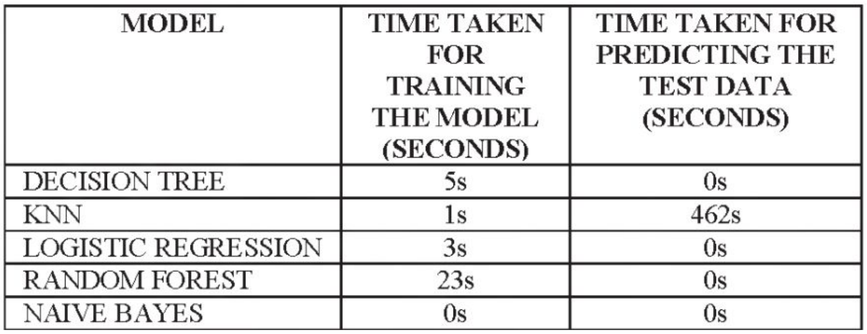
MODEL	SENSITIVITY	PRECISION
DECISION TREE	79.21	85.11
KNN	81.19	91.11
LOGISTIC REGRESSION	63.34	87.67
RANDOM FOREST	75.25	93.83
NAIVE BAYES	85.15	6.56

- Observations
    - PERFORMANCE EVALUATION OF VARIOUS MODELS AT THRESHOLD VALUE OF 0.4



MODEL	SENSITIVITY	PRECISION
DECISION TREE	79.21	85.11
KNN	81.19	91.11
LOGISTIC REGRESSION	69.31	87.5
RANDOM FOREST	78.22	89.77
NAIVE BAYES	85.15	6.52

    - TIME TAKEN FOR TRAINING AND PREDICTING DATA USING VARIOUS MODELS



MODEL	TIME TAKEN FOR TRAINING THE MODEL (SECONDS)	TIME TAKEN FOR PREDICTING THE TEST DATA (SECONDS)
DECISION TREE	5s	0s
KNN	1s	462s
LOGISTIC REGRESSION	3s	0s
RANDOM FOREST	23s	0s
NAIVE BAYES	0s	0s

➔Contributions

- Base on the observations above, the decision tree is the most suitable supervised machine learning algorithm for fraud detection. With the help of modern technology, malicious activity is lessened.

Reference

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9057851