International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 10 | Oct 2020 www.irjet.net p-ISSN: 2395-0072

Credit Scoring : A Comparison between Random Forest Classifier and K- Nearest Neighbours for Credit Defaulters Prediction

Priyanka Dewani[1], Mishika Sippy[1], Gopal Punjabi[1], Amit Hatekar[2]

[1]*Undergraduate Research Scholar, Dept. of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College, Mumbai – 50, Maharashtra, India*
[2]*Assistant Professor, Dept. of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College, Mumbai – 50, Maharashtra, India*

Abstract - *An activity within the banking industry is to extend credit to customers, hence, credit risk analysis is critical for financial risk management. There are various new methods used to perform credit risk analysis. The development of the credit scoring model has been regarded as a critical topic. In this research paper we will analyse a detailed comparison between Random Forest and K Nearest Neighbours algorithm. In this report, we have explained the algorithms and mathematical framework that goes behind developing the machine learning models. We discuss the speed and accuracy of the two Machine Learning algorithms mentioned when we test them on the UCI Credit Card database. After comparison and finding the gender with maximum debt, both the methods are refined and tuned to obtain better precision. Basically, we can conclude with a discussion and comparison of summarizing the best approach to classify these datasets.*

Key Words: Machine Learning, Credit Scoring, Random Forest Classifier (RFC), K-Nearest Neighbours (KNN), comparative study, probability of default

1. INTRODUCTION

Credit scoring is one of the most productive applications and operations research techniques used in banks and finance departments, also they are one of the earliest financial risk management services. Its main idea is to produce a score that any lending company can use to classify applicants into two groups: one group which is credit-worthy and which is likely to repay the financial obligation and another group which is non-credit-worthy and whose application for credit will be rejected due to a high possibility of defaulting on its financial obligations. Basically credit scoring is a typical classification problem. A credit scoring system allows lenders and other financial institutions to allow the creditworthiness of an individual. Some financial-based organizations establish their credit scoring methods. As the credit industry has been emerging rapidly, a large number of consumer credit data are collected by the credit sector of the bank and credit scoring has become a very important issue. Many of the factors contribute to credit scores assigned through the systems check. Factors include payment of interest, length of time using credit, amount of debt a person has and the types of debt that person has. Credit lenders use these methods to determine how much risk a particular borrower places on them if they decide to lend to that person. All these figures are risk-based. If a person has a low credit score, he or she is likely to pay more to borrow money to buy a home or fund a car purchase than someone with a higher credit score. While the credit scoring systems set up a guideline, individual lenders determine which level is acceptable and how much to charge in interest.

Usually, a large amount of redundant information and features are involved in the credit datasets, which often leads to a lower down of the accuracy and higher the complexity of the credit scoring models, so, effectively feature selection methods are necessary for credit dataset with a huge number of features.

Credit scoring was evaluated subjectively according to personal experiences, and later on, it was based on 5C: the character of the consumer, the capital, the collateral, the capacity, and the economic conditions. However, with the large increase in the number of applicants, it is impossible to conduct the work and a wide range of techniques has been applied to solve the credit scoring problem. Basically, those methods can be divided into two techniques: statistical methods (logistic regression, discriminant analysis) and machine learning techniques (like support vector machine, k-nearest neighbour, decision tree, neural network). According to previous studies, machine learning techniques are superior to that of traditional methods in dealing with credit scoring problems, especially in nonlinear pattern classification. For ordinary statistical classification, an underlying probability model should be assumed. The new data mining techniques have been adopted to build the credit scoring models. In addition to expert systems, a lot of classification techniques have been developed used in credit scoring applications. Researchers have developed a variety of conventional statistics models that involve linear discriminant model, decision tree model, rough set theory model, F-score model, and genetic programming model. Moreover, the researchers haven't come across any conclusive proof that one method is irrefutably superior over another. In this paper we would be talking and comparing the K-nearest neighbour algorithm and random forest algorithm.

2. RELATED WORK

In [1] the accuracy of different data mining techniques for predicting the credit card defaulters is compared. The dataset used in this research is from the UCI machine learning repository based on Taiwan's credit card clients default cases [2]. It has 30,000 instances, and 6626 (22.1%) of which are default cases. There are 23 features including credit limit, gender, marital status, last 6 months bills, last 6 months payments, etc. These are labelled with 0 (refer to non-default) or 1 (refers to default). The experiment ranks the following algorithms - artificial neural network, classification trees, naïve Bayesian classifiers, K-nearest neighbour classifiers, logistic regression, and discriminant analysis. The best performance was shown by K nearest neighbour with an accuracy of 82% on the training data and 84% on the test data. To get an actual probability of "default" they proposed a novel approach called the Sorting Smoothing Method (SSM). In [3], it validates a heuristic technique to mine ability default debts earlier in which a threat opportunity is precomputed from all preceding facts and the threat opportunity for the latest transactions are computed as quickly they take place. Except for this heuristic technique, it additionally makes use of a currently proposed device getting to know technique which has now no longer been carried out formerly at the centred dataset. In [4] previous classification studies, three non-parametric classifiers, Random Forest (RF), k-Nearest Neighbour (kNN), and Support Vector Machine (SVM), were reported as the foremost classifiers at producing high accuracies. The performances have been compared of these classifiers with different training sample sizes. In this study, we examined and compared the performances of the RF, kNN, and SVM classifiers. All classification results showed a high overall accuracy (OA) ranging from 90% to 95%. Among the three classifiers and 14 sub-datasets, SVM produced the highest OA with the least sensitivity to the training sample sizes, followed consecutively by RF and kNN.

In the work of [5], the authors present a comparison between Naive Bayes and Support Vector Machine (SVM) regarding sentiment classification using various datasets and the TF-IDF vectorizer. They further discuss models, data processing, ways to improve metrics like speed and precision using a grid search to determine the best-fitting model. An attempt to explain which model is better is done by displaying the results graphically and its future scope is also specified.

3. METHODOLOGY

This section explains the process of the experiment performed using the Random Forest Classifier and K Nearest Neighbours. For step one we collected the data from the UCI Machine Learning - Default of Credit Card Clients Dataset. This dataset had some missing data and anomalies which were taken care of by filling in the null values and eliminating the anomalies from the data set. In this paper, the data has been cleaned using scikits and pandas inbuilt libraries. The accuracy and precision of the credit scoring default prediction is done using two machine learning models – random forest classifier (RFC) and K-nearest neighbours (KNN). After cleaning the data set, it is split into the training data and testing data and it is passed on to both the models. Then after comparing the various metric values the best model of the two is selected.

Further grid search is used to find the hyperparameters that will give the best results. For random forest classifiers the hyperparameters like criterion, max depth, max features, n_jobs, etc. are used and for K-nearest neighbours the hyperparameters like leaf size, weights, algorithm, metric, etc . After this, the dataset is evaluated on the tuned models and the various metric parameters are calculated. The above helps us to determine which model performs better before and after tuning.

3.1 K – NEAREST NEIGHBOURS (KNN)

The K-nearest neighbours (KNN) algorithm is a type of supervised machine learning algorithm that can be used for classification and regression predictive problems. KNN is a lazy learning algorithm since it does not have a specialized training phase and uses all the data for training while classification. It is also a non-parametric learning algorithm because it does not assume anything about the underlying data.

The KNN algorithm uses 'feature similarity' to predict the values of the latest data points which can be assigned a worth supported by how closely it matches the points within the training set. It is a simple algorithm to understand and interpret which is very useful for nonlinear data because there is no assumption about data in this algorithm. Also, it is a versatile algorithm as we can use it for classification as well as regression. It has relatively high accuracy but there are far better-supervised learning models than KNN.

The output of k-NN depends on its use of classification or regression:

Class membership is obtained as the result of kNN classification. An object is know by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (where K is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour system. In k-NN regression, the output is the property value for the objects. This value is the average of the values of k nearest neighbours.

3.2 RANDOM FOREST CLASSIFIER (RFC)

The random forest algorithm is a supervised classification algorithm that can be used for regression and classification. As the name suggests, this algorithm creates a forest with a

number of trees. The forest is an ensemble of decision trees, generally trained with the bootstrap aggregating method commonly known as the bagging method. The logic behind the "bagging" method is that a mix of several base models is combined to produce an optimized model that provides an overall better result. In simple words, a random forest builds several decision trees and combines them to get a more accurate and stable prediction. More the number of trees, the higher is the accuracy of the results. The decision tree is rule-based. When a dataset is trained with features and targets, the tree will give an algorithm with a set of rules which can be used for prediction on the testing dataset. In a random forest finding of nodes and splitting the feature nodes are done randomly. It is based on the construction of a myriad of different decision trees composed of forest that are then aggregated the diversity of these trees come from aspects of the construction of the forest. Each tree is built on a random sample of observation of the bagging method, and a random set of features is chosen to split nodes (feature sampling). It has built-in hyperparameters to increase its prediction efficiency and speed like n jobs, random state, minimum sample leaf, n estimators, max features, etc.

There are several advantages of random forest algorithms like it can avoid overfitting problems when there are more trees in the forest. Both classification and regression can use random forest algorithms and missing values are taken care of. The random forest classifier can also be modelled for categorical values and these algorithms are very tough to beat performance-wise. It is also a fast, simple and flexible tool.

3.3 TERMINOLOGIES

- Accuracy - It is the fraction of the number of correct predictions over the total number of input samples.
- ROC - It is a graph depicting performance of a classification model at all classification thresholds. This curve plots two parameters namely, True Positive Rate and False Positive Rate.
- F1 score- The Harmonic Mean between recall and precision is called the F1 score.
- PRECISION - It aims to answer what proportion of the positive identifications are actually correct. The higher the precision, the more positively identified data is correct.
- RECALL - It is the ratio of number of correct positive results and the number of all relevant samples that should have been identified as positive. Recall and precision are based on a measure of relevance.

4. EXPERIMENT

The following section discusses the datasets, algorithm and performance measurements.

4.1 DATASET DESCRIPTION

This dataset was obtained from UCI Learning This dataset was initially used in 2007. The research was aimed at comparing the predictive accuracy of probability of default. The dataset has 30000 instances and 23 attributes listed in 23 columns.

Table – 1: Explanation of Dataset

VARIABLE	PARAMETER	VALUE ASSIGNED
X1	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit	
X2	Gender	1 = male; 2 = female
X3	Education	1 = graduate school; 2 = university; 3 = high school; 4 = others
X4	Marital status	1 = married; 2 = single; 3 = others
X5	Age	
X6	The repayment status in September, 2005	History of past payment. We tracked the past monthly payment records (from April to September, 2005)
X7	the repayment status in August, 2005	

. . .	. . .	
X11	The repayment status in April, 2005	
X12	amount of bill statement in September, 2005	Amount of bill statement (NT dollar)
X13	amount of bill statement in August, 2005	
. . .	. . .	
X17	amount of bill statement in April, 2005.	
X18	amount paid in September, 2005	Amount of previous payment (NT dollar)
X19	amount paid in August, 2005	
. . .	. . .	
X23	amount paid in April, 2005.	

4.2 ALGORITHM

Input: default credit data
Output: Accuracy, precision, f1 score, recall, roc, graphs

1. Importing data
2. Extracting relevant data
3. Gathering insights
4. Splitting data into testing and training set
5. Using grid search to get best parameters
6. Running the code
7. Storing various metrics
8. Plotting graphs

4.3 PERFORMANCE MEASUREMENTS

a) Accuracy = (TP+TN)/ (P+N); where: P = TP+FN and N = TN+FP
b) Recall = TP/ (TP+FN);
c) Precision = TP/ (TP+FP);
d) F-1 Score = 2TP / (2TP+FP+FN);
e) ROC CURVE: It is a graph which shows the performance of a model at all thresholds by plotting two parameters - True Positive Rate(TPR) and False Positive Rate(FPR) where,

TPR=TP/(TP+FN)
FPR= FP/(FP+TN)

* P = The number of real positive cases in data
* N = The number of real negative cases in data
* TP = True Positive; TN = True Negative
* FP = False Positive; FN = False Negative

5. EXPERIMENTAL RESULTS

This section contains the graphs obtained from the insights on the dataset and scoring prediction using Random Forest Classifier and K-Nearest Neighbour.

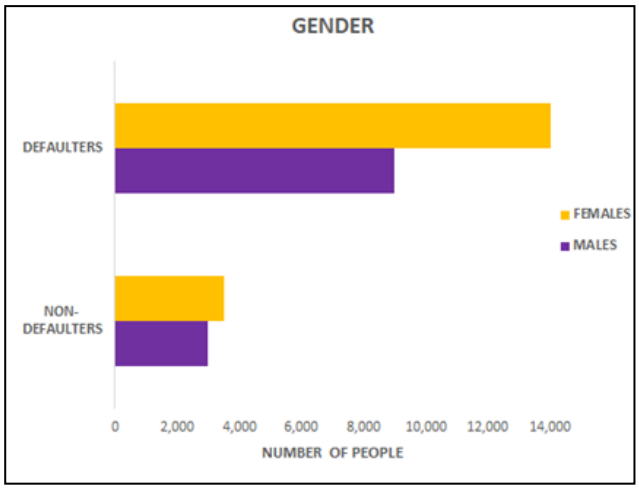

Chart – 1: Credit Score Defaulters - Gender

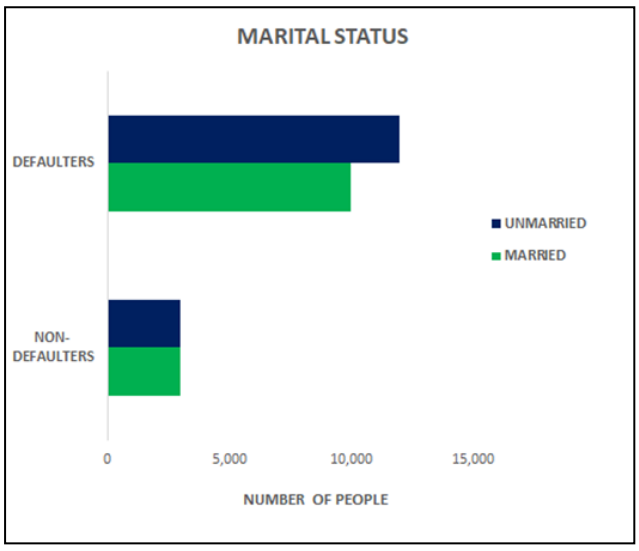

Chart – 2: Credit Score Defaulters – Marital Status

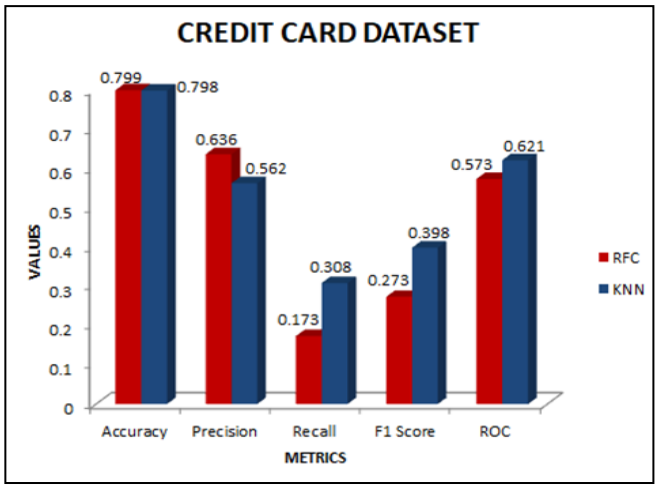

Chart – 3: Metric values of RFC and KNN before tuning

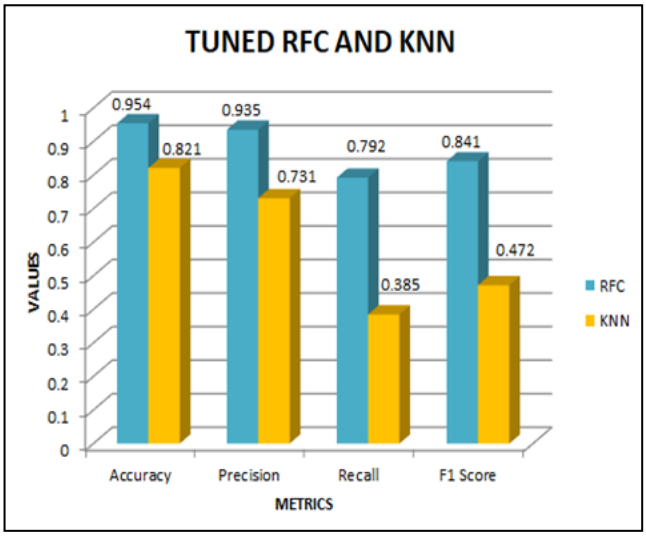

Chart – 4 : Metric values of RFC and KNN after tuning

6. CONCLUSIONS

- We have observed that before tuning, metrics of KNN were better than RFC. However, post tuning the parameters of RFC improved and turned out to be better than that of KNN.
- As expected, after tuning the precision of RFC has increased but the recall increased significantly as well which is not desired. This will hamper the overall performance as there should be a good balance between the two parameters.
- The precision and recall of KNN have increased. Since the increase in recall is very less, the performance will not be affected negatively.
- The accuracy of both KNN and RFC have ameliorated after tuning.
- The F1 score has increased for KNN and RFC but when compared, the percentage increase is more in the case of RFC and less in the case of KNN.
- The total training time for RFC was more than KNN.
- When it comes to comparing the marital status of the people whose loan paying capacity is judged, it turns out that unmarried people do not pay their loans on time as compared to married people. There are approximately equal numbers on both sides who clear their dues timely. However, unmarried people are generally defaulters.
- When it comes to comparing the genders of the people whose loan paying capacity is judged, it turns out that females do not pay their loans on time as compared to the males. Approximately equal numbers of males and females clear their dues in time but as a whole the majority of defaulters are females.

ACKNOWLEDGEMENT

We highly appreciate the guidance and support provided by our parents and our professor Mr. Amit Hatekar.

REFERENCES

[1] Yeh, I-Cheng, and Che-hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert Systems with Applications 36.2 (2009): 2473-2480.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] "UCI Machine Learning Repository: Default of credit card clients Data Set". Available: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#. Accessed: December. 23, 2017.

[3] Sheikh Rabiul Islam, William Eberle, Sheikh Khalid Ghafoor. Credit default mining using combined machine learning and heuristic approach. July 2018

[4] Phan, Thanh-Noi & Kappas, Martin. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors. 18. 18. 10.3390/s18010018.

[5] Prayag Patel, Ananya Arora, Saud Sheikh & Prof. Amit Hatekar. (2020) Support Vector Machine versus Naive Bayes Classifier: A Juxtaposition of Two Machine Learning Algorithms for Sentiment Analysis. International Research Journal of Engineering and Technology (IRJET).