International Journal of Advanced Science and Technology
Vol. 29, No. 8, (2020), pp.4076-4085

Applied Industrial Machine Learning Towards Good Banking Credit Score Abiding The CIBIL Guidelines

Shruti Pant[1], U.M.Prakash[2]
[1,2]*dept. Computer Science And Engineering SRM Institute Of Science And Technology Kattankulathur, India*
[1]*shrutipant1710@gmail.com,* [2]*prakashm3@srmist.edu.in*

Abstract

The applications of Machine Learning (ML), and it's great capabilities to reduce strenuous functions and provide more accurate and robust results, has lured banking companies to adopt this approach, in order to keep ahead of the market competitions and gain an upper hand technologically thereby reducing the credit risks. In this rapidly advancing environment, it has become a necessity to understand the Customer's Credit Worthiness for financial organizations, as this instills confidence in the lenders. In the past, many models and tools were developed to coup up with the risk factor concerning the banks, these tools and models were mainly statistical tools with core algorithmic concepts. Now in this ever-expanding domain of the Financial technology domain, the automation of these tasks can be achieved using classification algorithms that aid in labeling the customers based on the data fed. The paper aims to establish reliable models and evaluate their performance reports, of the multiple machine learning models, for selection purpose, the K Fold Cross- validation technique for model selection is used, and furthermore the paper tries tries to enhance the model's parameters for the utmost results,Anaconda Jupyternotebook is the platform used for coding and evaluating the models.To train and test the model's accuracy,a data-set of credit card related features are used. The target variable to be classified is termed as, 'probability of default', bearing in mind the utilization of the CIBIL guide framework for feature evaluation and understanding.

Keywords: *Predictive Modelling, Credit Worthiness, Data Classification, Ensemble Techniques,Machine Learning models, K-fold Cross Validation ,Parameter tuning*

1 Introduction:

Credit Worthiness (CW) is the metric used for evaluating the performance of the customer, requesting for a certain amount of credit, based on his/her past passbook records. It is a three digit representation which is associated as a credit score value by banking Organizations provided to the customer as per set guidelines and parameters. The creditworthiness of an client is evaluated by the credit rating system, set by pre-exisiting methods of evaluations and frameworks. A high credit score grants high CW and vice versa. Some of the crucial and important parameters for evaluation of CW are, credit history, health status and previous credit scores related to the individual all depicts how a person meets debt obligations during the prior loan cycles, this establishes the creditworthiness a person.

Financial Organizations also consider the amount of available liabilities, assets and other valuable goods and possession are used to determine if the customer may be a defaulter or not. In addition, many other features that play an instrumental role are age, income, customer health conditions, employment status, financial obligations, debt owed and re-paied, accounts held by the customer, duration of payment history and the capability of the client to repay debt under prescribed period, are just some of the major factors that contribute toward the derivation of a person's CW. The need for speedy analysis brings in crucial points of interest such as:

1) Ingestion of sensitive large volume of the dataset.

2) Utilization of a standard framework that helps in normal- izing the task performed across banks.

3) Process and massage the relevant dataset into an acceptable format that provides an excellent opportunity to train and test the models at its best.

The results that are evaluated and compared with ML models, in terms of important performance metrics.

2 State Of The Art (Lıterature Survey):

A. "From Data Mining to Knowledge Discovery in Database"[1]

1) Summary: Datamining process applicable to various formats of datasets.

2) Advantages:

- Understood the concept and application of CRISP-DM process.

B. "An Experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring"[2]

1) Summary: Provides an insight upon the various key factors responsible for uderstanding CW.

2) Advantages:

- Provides solid approaches to classification techniques.
- Basic and advanced ensemble techniques require various formating of data to understand key concepts.

C. "Ensemble Methods in Machine Learning"[3]

1) Summary: Ensemble methods are classification tech- niques, used to classify binary or multi-class target variables.

2) Advantages:

- Key understanding of classification techniques.
- Better understanding and strong foundations for XGB- Classifers.

D. "XGBoost: A Scalable Tree Boosting System."[4]

1) Summary: The concept of bagging and boosting is applied on trees, which greatly impact and improve results of the model.

2) Advantages:

- Generally performs well under classification circum- stances.

- Newly sought after model with large community service.

E. "Machine Learning in Incident categorization Automa- tion."[5]

1) Summary: The paper provides an automated solution of SVM to the classification issue not adhering to guidelines.

2) Advantages:

- High automation opportunity leads to newer and greater findings with the model.

F. "A comparative assessment of ensemble learning for credit scoring"[6]

1) Summary: It is an assessment of ensemble techniques, based upon the three popular techniques of bagging, boosting and stacking.

2) Advantages:

- Bagging performs better than Boosting across all credit dataset.

G. "Decimated input ensembles for improved generaliza- tion"[7]

1) Summary: Usage of ensemble classifiers instead of a single classifier improves generalization performance in many difficult problems.

2) Advantages:

- Input decimation combination improves the generaliza- tion performance.

H. "Bagging predictors"[8]

1) Summary: The paper helps in understanding, of methods for generating multiple versions of a predictor.

2) Advantages:

- The aggregated value for the predictors is taken into consideration when dealing with the selection of multiple versions.

I. "Random Search for Hyper-Parameter Optimization "[9]

1) Summary: The tuning of hyper-parameters of models using grid search technique.

2) Advantages:

- Default parameters, will be replaced with the model's best fit parameters using the grid search technique.

J. "Predicting Credit Worthiness of Bank Customer with Ma- chine Learning Over Cloud"[10]

1) Summary: Utilising domain specific data, to analyse and generate classification models to predict creditworthiness.

2) Advantages:

- The ideas of basic models utilized for classification can be understood in this paper.

K. This article

1) Summary: Provides a global application and improved view on grading the Credit worthiness processes. Adopts fundamental approaches into the field and provides practical application.

2) Advantages:

- Improvement on existing results.
- Newer insights onto the basic functions and results of the analysis performed.

3 Proposed Work:

A. Proposed Data Ingestion Method For Capturing Maximum Essence Of The Data

The feature selection method across the derived multiple data-set is an import method for which the reference of the CIBIL framework is utilized. The features of each structure are obtained by querying the reliable database using SQLAlchemy and stored as a comma-separated (.csv) file locally on the system. The data-sets are stringently examined and only the required attributes are selected and joined across all the data set to derive a master data-set. The join operation performed is purely based on the corresponding shared keyID value available across all the tables. It is highly crucial to select relevant attributes, without including unnecessary ones and also along the way of not losing essential ones.

Now the raw data is processed and majority of its abnor- malities are reduced in order to obtain a cleaned dataset for analysis. The first step of process is termed as null value reduction. In this process, the thumb rule of 5% comes into effect where, if the number of discrepant values is lesser than 5% of the total size of the feature it is eliminated from the dataset. Coming to the other types of null value reduction is when, either the feature has categorical values or it has continuous values. In the case of categorical values, the function of either forward fill or backward fill is applied to cover up the null spaces. In the case of continuous values, the mean of the attribute is taken into consideration and the spaces are filled using that value.

The next process is termed as the outlier analysis. Outlier analysis is performed in order to figure out the distribution of the data and removes all outliers and extreme values that do not satisfy in the distribution. For visualisation purpose a boxplot is the most effective which will provide key insights into the distribution, quartile wise. The general best practice performed is to only consider the data distribution within the 3-sigma points from the mean as this captures +95% essence of the data.

The data cleaned is not yet ready for analysis, as there are attributes that provide no input to the analysis and in turn can reduce the efficiency of the model. A correlation matrix is generated for the attributes. Highly correlated attributes are eliminated as it causes redundancy in the dataset. This is the last step in the process after which we have derived a completely cleaned and massaged dataset ready for analysis.

Figure 1. Prescribed flow of work

B. Equation Representation

The metrics used to measure the performance of an algorithm is considered as a high output value when accuracy, precision, recall, and F1-score is high. This showcases that the algorithm tends to maximize the output value, a sign of good trend.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP + FP}{TP + FN}$$

$$Recall = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

1) One-Error (OE): The process evaluates the number of instances where the top-ranked labels are not a part of the relevant set.

2) Ranking Loss (RL): The metrics takes into account the number of instances where, irrelevant labels are tend to be ranked in a higher order than the relevant ones.

3) Average-Precision (AP): Average precision is the mea- sure that is used to combine the precision and recall values to retrieve results.

$$AveragePrecision = \frac{\Sigma_r P@r}{K}$$

C. Model Phases

The cleaned and massaged dataset from the pre-processing stage is split into its train and test components. This being a supervised approach towards the model, we have labeled the features accordingly into Xi and Y variables. The categorical variable is stored in the Y variable and the rest of the corresponding input variables are stored in the Xi variable where I tend from 1 to n. To find the appropriate model with the best accuracy we utilize the concept of K fold cross- validation technique. In this technique of model selection, the training data is fed into the model in a parallel fashion and corresponding models undergo its phase of training. The models used are Random Forest, KNN, CART, XGBTREE, XGBDART, SVM, ANN, and GBC.

D. Key Equation

Accuracy: The accuracy of a model is evaluated by the correct classification of the instances.

$$Accuracy = \frac{TP+TN}{No.Of\ Instances} \times 100 \quad (1)$$

True Positive: True Positive identifies the amount of correctly classified instances, to the total number of instances.

Prediction Rate: It's defined as the correctness among the test data, the formula is defined as:

$$PredictionRate = \frac{TP}{TP+TN} \times 100 \quad (2)$$

Recall: Correctly classified positive case is defined as recall/ sensitivity.

$$PredictionRate = \frac{TP}{TP+FN} \quad (3)$$

E. Comparison Of Results

The cleaned and massaged data is pipe-lined using the K- fold cross validation technique and fed into multiple ensemble techniques each one coming out with its own results.

1) Random Forest: Random Forests is also termed as nearest neighbor approach, can be. Start by visualizing what forest is, you'd reach the conclusion that as a tree (weak) combines with others and a forest (strong) is formed. Random forest is an ensemble approach in which the weak learners combine to form strong learners. This helps improve efficiency. If single tree analogs to a single classifier, then the forest analogs to a group of classifiers. Assuming you are familiar with decision trees since the random forest algorithm follows the basic decision tree approach where an input flows down the tree just as sunlight would on a sunny day.

For the analysis performed, the accuracy score derived is 80.76%.

2) K Nearest Neighbour : In this classification algorithm, the test data is given the majority class. The heuristic selection of value K signifies the closest neighbours of a datapoint in the plane. Using general distance theorem such as manhattan distance, the points are grouped to its closest corresponding class.

3) Support Vector Machine: This is a classification algorithm of the supervised category.In this algorithm the data are plotted in a three-dimensional canvas and a planar slicer is utilised to classify these points into their corresponding classes.

4) XGBoost classifier: It is an ensemble (i.e. meta) machine learning algorithm that builds a strong model based on many weaker ones sequentially. To do so, it uses gradient descent. This technique when being used with decision trees, is called gradient boosting trees. The library is written in C++ and offers many useful wrappers in higher-level languages (Python and R for instance) hence it is fast.

The results are tabulated below and their corresponding graph comparison showcases the difference in the results derived :

Algorithm	Accuracy
RF	95.713
KNN	95.629
CART	90.824
XGBgbtree	95.702
XGBDart	95.7025
SVM	95.707
ANN	93.462
GBC	95.650

Table 1. Accuracy Table Of Various Algorithm

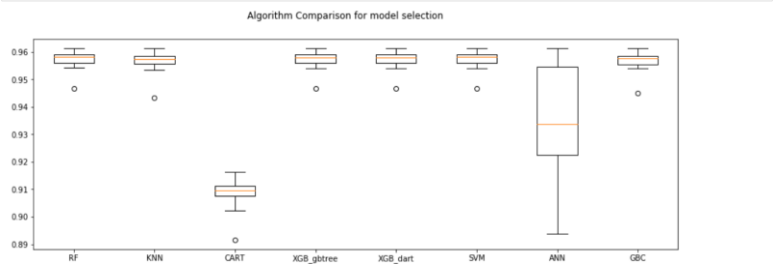

Figure 2. Accuracy Box Plot

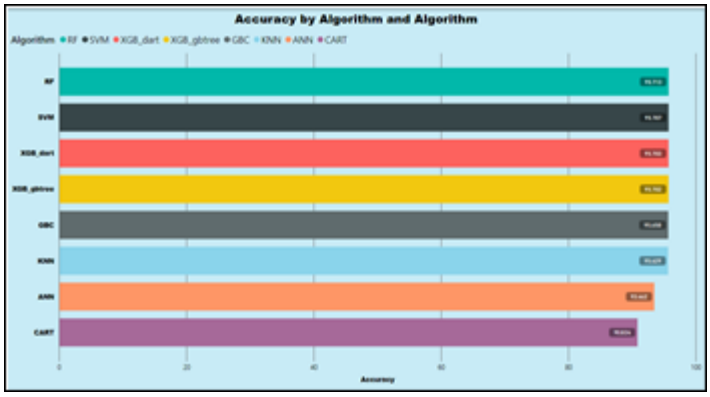

Figure 3. Accuracy Comparison

5) Optimization: The selected model of Random Forest and XGBoost Classifiers were utilising default set of parameters for which they were recieving the certain results.To improve the performance of the model, the utilisation of two techniques for hyper parameter tuning is applied.

Both these tuning techniques are a part of the ”sklearn.gridsearch” package. In gridsearchcv, from a given range the technique tries every combination for the hyper-parameter values. When searching a 2-dimensional space (i.e. optimizing two parameters), the process of grid search will look like.

The two selected models, namely RandomForest Algorithm and XGBoost Classifiers, were able to provide a satisfactory score when its parameters were set to their default values. The technique of parameter tuning, applies a very simple logic to utilise parameters either in a one to one fashion taking on all the permutation and combinations of the value set, or the parameter values are randomly selected within a specified range and are selected at random.

Each of these techniques have their own advantages and disadvantages. The gridsearchcv technique is a resource consuming technique, it consumes high amount of space and time and computational power. This technique however reaches its saturation and find the best parameter value. RandomizedCV is a hit and trial method where the best case and worst case both are possible for searching parameter values.

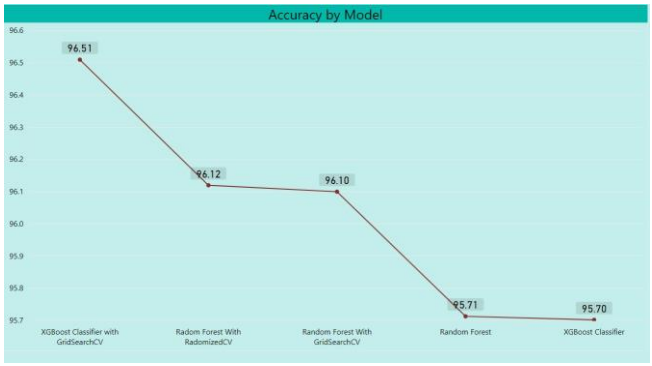

Figure 4. Accuracy Comparison

Algorithm	Accuracy
RF	95.713
XGBgbtree	95.702
RF+GridSearchCV	96.10
RF+RandomizedCV	96.12
XGBoostClassifier+GridSearchCV	96.51

As seen in the above diagram,the models have a slightly better accuracy score compared to their default ones. The increase of the scores from 95% aggregate to 96% , indicates that when the hyper parameters are tuned they provide the models a better result.

4 Conclusıon And Future Work:

Financial organizations are facing a lot of issues in providing loans and other benefits to their customers due to the constant risk of defaulters. The usage of machine learning models to provide this crucial information regarding the customers will allow the banks to bravely make decisions regarding the credits it's giving out. These models when optimised to provide the best results act as a resourceful aid to the organization.

"Jupyter Notebook", was used for the coding of these models and its optimizations.

In the future, I intend to build up an automated risk assessment system over the cloud for financial organizations that will incorporate key features to determine credit worthiness of customers.

References

1. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases",American Associ- ation for Artificial Intelligence. All rights reserved. 0738-4602-1996
2. Loris Nanni , Alessandra Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", Elsevier, Expert Systems with Applications 36 (2009) 3028–3033.
3. Giorgio ValentiniFrancesco Masulli, Ensembles of Learning Machines, Part of the Lecture Notes in Computer Science book series (LNCS, volume 2486).
4. Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust 2016 Pages 785–794https://doi.org/10.1145/2939672.2939785
5. Sara Silva Instituto Universit a´rio de Lisboa (ISCTE-IUL), Lisboa, Portu- gal ; Rube´n Pereira ;

Ricardo Ribeiro, Machine learning in incident cat- egorization automation, Information Systems and Technologies (CISTI), Iberian.
6. Gang Wang, Jinxing Hao, Jian Mab, Hongbing Jiang, "A comparative assessment of ensemble learning for credit scoring", Expert Systems with Applications 38 (2011) 223–230.
7. K. Tumer and N. C. Oza, "Decimated input ensembles for improved generalization," in Proceedings of the International Joint Conference on Neural Networks (IJCNN '99), pp. 3069–3074, Washington, DC, USA, July 1999.
8. L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.
9. James Bergstra JAMES, Yoshua Bengio YOSHUA ,Departement d'Informatique et de recherche op ´ erationnelle ´,Universite de Montreal ´,Montreal, QC, H3C 3J7, Canada, "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research 13 (2012) 281-305 Submitted 3/11; Revised 9/11; Published 2/12.
10. A. Motwani , P. Chaurasiya , G. Bajaj Computer Science Engineering, Sagar Institute of Science Technology Research, RGPV, Bhopal, India Computer Science Engineering, S.V. Polytechnic College, Bhopal, India, Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud .