IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CREDIT CARD APPLICANT WORTHINESS PREDICTION USING MACHINE LEARNING

B.Dhanalakshmi[1] , Dr.K.Merriliance MSc,MPhil,PhD[2]

Department of Computer Applications, Sarah Tucker College, Thirunelveli-7.

Abstract: Machine learning is an emerging technique for building analytic models for machines to "learn" from data and be able to do predictive analysis. The ability of machines to "learn" and do predictive analysis is very important in this era of big data and it has a wide range of application areas. For instance, banks and financial institutions are sometimes faced with the challenge of what risk factors to consider when advancing credit/loans to customers. For several features/attributes of the customers are normally taken into consideration, but most of these features have little predictive effect on the credit worthiness or otherwise of the customer. Furthermore, a robust and effective automated bank credit risk score that can aid in the prediction of customer credit worthiness very accurately is still a major challenge facing many banks. In this paper, we examine a real bank credit data and conduct several machine learning algorithms on the data for comparative analysis and to choose which algorithms are the best fit for learning bank credit data. The algorithms gave over 80% accuracy in prediction. Furthermore, the most important features that determine whether a customer will default or otherwise in paying his/her credit the next month are extracted from a total of 23 features. We then applied these most important features on some selected machine learning algorithms and compare their predictive accuracy with the other algorithms that used all the 23 features. The results show no significant difference, signifying that these features can accurately determine the credit worthiness of the customers. Finally, we formulate a predictive model using the most important features to predict the credit worthiness of a given customer.

Index Terms **- Component,formatting,style,styling,insert.**

I. INTRODUCTION

This research is focusing on application of machine learning (ML) techniques to predict customer eligibility for a credit card.

One of key objective of the bank is to increase the returns. When increasing the returns there is an increase of risk. Banks are faced with various risks such as interest rate risk, market risk, credit risk, off-balance-sheet risk, technology and operational risk, foreign exchange risk, country or sovereign risk, liquidity risk, liquidity risk and insolvency risk. Effective management of these risks is key to a bank's performance. Credit can be defined as the risk of potential loss to the bank if a borrower fails to meet its obligations (interest, principal amounts). Continuously monitoring of customer payments could reduce the probability of accumulating non-performing assets (NPA). Whether to grant or not to grant a loan to a customer is one of key decisions of banks use to reduce probable NPA at the first hand. Credit card as a credit facility instruments banks need to effectively managed credit risk of the facility. The Basel Accord allows banks to take the internal ratings-based approach for credit risk. Banks can internally develop their own credit risk models for calculating expected loss.

There are several manual steps involving when granting a credit card to a customer. Assessing applicant's creditworthiness and checking the eligibility are the key factors and decisions the bank would take about a credit worthiness will not always be accurate. Application of machine learning techniques can eliminate manual paperwork, time-consuming processes and most importantly data driven decision making before granting a credit card to a customer. In this research, different supervised machine learning algorithms were used to develop models and follow the steps in cross-industry standard process for data mining (CRISP-DM) life cycle. Accuracy of models was validated by using different validation techniques.

Motivation

In times of yore, when providing a credit card to a customer, banks had to rely on the applicant's background and the history to understand the creditworthiness of the applicant. The process includes scrutinization of application data with reference documents and this process was not always accurate and customers and the bank had to face difficulties in approving the credit card. But with the digital transformation, there is a growth in Artificial Intelligence & Machine Learning Technology in the past two decades. Therefore, ML techniques being used to evaluate credit risk and automate credit scoring by predicting the customer eligibility correctly using customer demographic data and historical transactional data. Furthermore, ML helps banks to make smarter data – driven decisions for customers; use banking data in a more productive and efficient way; streamline customer interaction by removing manual and lengthy processes.

Statement of the problem

Many researchers have conducted machine learning applications on credit scoring and customer default predictions. Researchers' have concluded that SVM (support vector machine) and ANN (Artificial Neural Network) performed better than other classifiers. However, it is important to study how these two algorithms behave differently with filter based feature selection and balancing imbalanced data which is inherited by nature using Synthetic Minority Oversampling Technique (SMORTE). "To examine two algorithms and identify best classification algorithm to predict customer eligibility for a credit card and to minimize possible credit loss "

II. LITERATURE SURVEY

In[1], The following research reveals the significance of modified classification in estimating new trends. Rigorous evaluation of different classification algorithms viz. logistic regression, decision tree, K-nearest neighbour (KNN) and Naive Bayesian are explored. These findings forecast the finest techniques for discovery of potential defaulters. Our motive is to compare the performance measures between original dataset and original dataset on which principal component is applied. Different algorithms can be compared on the basis of various criterions such as accuracy, precision, F1-score, recall, ROC. We proceeded by applying a general data imbalance handling technique such as smote technique and near miss technique. A comparison is then drawn between the modified dataset with the principle component analysis applied and the imbalance in the original dataset being corrected with the help of under sampling and oversampling. The comparison helps us identifying the best among dimensionality reduction and data imbalance handling techniques on the chosen dataset.

In[2], The naïve Bayes rule (NBR) is a popular and often highly effective technique for constructing classification rules. This study examines the effectiveness of NBR as a method for constructing classification rules (credit scorecards) in the context of screening credit applicants (credit scoring). For this purpose, the study uses two real-world credit scoring data sets to benchmark NBR against linear discriminant analysis, logistic regression analysis, *k*-nearest neighbours, classification trees and neural networks. Of the two aforementioned data sets, the first one is taken from a major Greek bank whereas the second one is the Australian Credit Approval data set taken from the UCI Machine Learning Repository (available at The predictive ability of scorecards is measured by the total percentage of correctly classified cases, the Gini coefficient and the bad rate amongst accepts. In each of the data sets, NBR is found to have a lower predictive ability than some of the other five methods under all measures used. Reasons that may negatively affect the predictive ability of NBR relative to that of alternative methods in the context of credit scoring are examined.

In[3], Credit scoring systems are based on Operational Research and statistical models which seek to identify who of previous borrowers did or did not default on loans. This study looks at the question when will borrowers default not if they will default. It suggests that some of the reliability modelling approaches may be useful in this context and may help identify who will default as well as when they may default.

In[4], This research deals with the challenge of reducing banks' credit risks associated with the insolvency of borrowing individuals. To solve this challenge, we propose a new approach, methodology and models for assessing individual creditworthiness, with additional data about borrowers' digital footprints to implement comprehensive analysis and prediction of a borrower's credit profile. We suggest a model for borrowers' clustering based on the method of hierarchical clustering and the k-means method, which groups actual borrowers having similar creditworthiness and similar credit risks into homogeneous clusters. We also design the model for borrowers' classification based on the stochastic gradient boosting (SGB) method, which reliably determines the cluster number and therefore the risk level for a new borrower. The developed models are the basis for decision making regarding the decision about lending value, interest rates and lending terms for each riskhomogeneous borrower's group. The modified version of the methodology for assessing individual creditworthiness is presented, which is to reduce the credit risks and to increase the stability and profitability of financial organizations.

In[5], Classification using class-imbalanced data is biased in favor of the majority class. The bias is even larger for high-dimensional data, where the number of variables greatly exceeds the number of samples. The problem can be attenuated by undersampling or oversampling, which produce class-balanced data. Generally undersampling is helpful, while random oversampling is not. Synthetic Minority Oversampling TEchnique (SMOTE) is a very popular oversampling method that was proposed to improve random oversampling but its behavior on high-dimensional data has not been thoroughly investigated. In this paper we investigate the properties of SMOTE from a theoretical and empirical point of view, using simulated and real high-dimensional data.

In[6], Recent year researches shows that data mining techniques can be implemented in broad areas of the economy and, in particular, in the banking sector. One of the most burning issues banks face is the problem of non-repayment of loans by the population that related to credit scoring problem. The main goal of this paper is to show the importance of applying feature selection in data mining modeling of credit scoring. The study shows processes of data pre-processing, feature creation and feature selection that can be applicable in real-life business situations for binary classification problems by using nodes from IBM SPSS Modeler. Results have proved that application of hybrid model of feature selection, which allows to obtain the optimal number of features, conduces in credit scoring accuracy increase. Proposed hybrid model comparing to expert judgmental approach performs in harder explanation but shows better accuracy and flexibility of factors selection which is advantage in fast changing market.

In[7], Imbalanced classification problems are often encountered in many applications. The challenge is that there is a minority class that has typically very little data and is often the focus of attention. One approach for handling imbalance is to generate extra data from the minority class, to overcome its shortage of data. The Synthetic Minority over-samplingTEchnique (SMOTE) is one

of the dominant methods in the literature that achieves this extra sample generation. It is based on generating examples on the lines connecting a point and one its *K*-nearest neighbors. This paper presents a theoretical and experimental analysis of the SMOTE method. We explore the accuracy of how faithful it emulates the underlying density. To our knowledge, this is the first mathematical analysis of the SMOTE method. Moreover, we analyze the effect of the different factors on generation accuracy, such as the dimension, size of the training set and the considered number of neighbors *K*. We also provide a qualitative analysis that examines the factors affecting its accuracy. In addition, we explore the impact of SMOTE on classification boundary, and classification.

In[8], Machine learning is playing a prominent role in current era. In this modernized world almost all the applications are manipulated and controlled by machine learning algorithms. By the use of historical data there are possibilities to predict the future. Even though a number of researchers are working on various machine learning algorithms, the performance and exactness of the algorithms still remains as a challenge. This work focuses on the performance analysis of various classification algorithms in terms of precision, recall, f-measure etc., to predict the bank loan approval status.

In[9], This research is focusing on application of machine learning (ML) techniques to predict customer eligibility for a credit card to mitigate possible future credit risk which may affect the bank's financial stability and credit performance. Credit card is a credit facility given for a customer by banks and finance companies around the globe. The credit facility has a credit risk for the banks and financial companies. The repayments are least assured and it often ends up as a non-performing credit facility (NPL). To mitigate credit risk banks are assessing applicant's creditworthiness and checking the eligibility before granting a credit facility. The decision is mostly based on traditional credit scoring models and credit worthiness will not always be accurate. This project aims to help banking and financial institutions to identify and interact with creditworthy customers by using predictive models. We used Artificial Neural Network (ANN) and Support Vector Mechanism (SVM) to develop models. Under ANN we have tested models using different sizes of batches, low and high learning rates. Linear SVM and Nonlinear SVM both models used to evaluate the best SVM method. Statistical methods under filter-based feature selection methods applied for feature selection. Model accuracy checked using Mean Absolute Error, Confusion Matrix, Area Under Curve (AUC) for training and test data. We have evaluated three classifiers and we observed that Nonlinear SVM is performed better than ANN and linear SVM. Nonlinear SVM model Accuracy is 0.88, Precision is 0.88, Recall is 0.90 and AUC is 0.89. Accuracy, Precision and Recall values are higher in Nonlinear SVM than ANN and Linear SVM. Recall rate is 0.90 means the model predicts positive class 90% correctly. We also realized that customer behavior might be different from country to country and application of several real banking datasets not limited to customer demographic and sociocultural but also other credit facility features including COVID-19 impact to be an area of concern for researchers. Furthermore, whether there is a relationship between Nonlinearity in highly imbalanced class problems with SMORTE application is another area of concern for researchers.

In[10], Artificial neural network is an information processing system which is influenced by the human brain and works on the same principles of the biological nervous system. They possess the ability to extract meaning from complex and intricate data, by detecting trends and extracting patterns from it. This paper illustrates the ability of neural network model and linear regression model constructed to predict the creditworthiness of an application accurately and precisely with minimal false predictions and errors. The results are shown to be similar for both the models, thus, models are efficient to use depending on the type of application and attributes.

III. METHODOLOGY

Data Collection: The dataset collected for foretelling loan failure clients is foretold into Training set and testing set. Generally 8020 proportion is applied to dissociate the training set and testing set. The data model which was created using Decision tree is applied on the training set and hung on the test take fineness, Test set forecasting is done. There are the 12 attributes.

Dataset Collection:

Data is collected from a variety of sources and prepared for data sets. And this data is used for descriptive analysis. Data is available from several online abstract sources such as Kaggle.com and data.gov.in. We will use an annual summary of credit card worthiness for at least 10 years.

Preprocessing:

The collected data may contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed and so it'll better the effectiveness of the algorithm. We should remove the outliers and we need to convert the variables. In order to flooring these issues we use chart function.

This step is very important in machine learning. Preprocessing consists of inserting the missing values, the appropriate data range, and extracting the functionality. The kind of the dataset is critical to the analysis process. In this work we have used isnull() method for checking null values and lable Encoder() for converting the categorical data into numerical data.

Train model on training data set:

Now we should train the model on the training dataset and make soothsayings for the test dataset. We can divide our train dataset into two tract train and testimony. We can train the model on this training part and using that make soothsayings for the testimony part. In this way, we can validate our soothsayings as we've the true soothsayings for the testimony part (which we don't have for the test dataset)

Correlating attributes:

Grounded on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can include Property area, education, loan measure, and originally credit History, which is since by insight it's considered as important. The correlation among attributes can be associated using corplot and boxplot in Python platform.

Feature Selection:

Feature extraction should simplify the amount of data involved to represent a large data set. The features characteristics extracted from the pre-treatment phase constitute the final set of training. These characteristics include the physical and chemical properties of the soil. Here, we have used Random Forest Classifier() method for feature selection. This method selects the features based on the entropy value i.e., the attribute which is having more entropy value is selected as an important feature for yield prediction.

Split the Dataset into Train and Test Set:

This step includes training and testing of input data. The loaded data is divided into two sets, such as training data and test data, with a division ratio of 80% or 20%, such as 0.8 or 0.2. In a learning set, a classifier is used to form the available input data. In this step, create the classifier's support data and preconceptions to approximate and classify the function. During the test phase, the data is tested.The final data is formed during preprocessing and is processed by the machine learning module.

Applying Machine Learning modules:

In our project, we have used three different supervised machine learning algorithms for credit card worthiness predictionprediction.

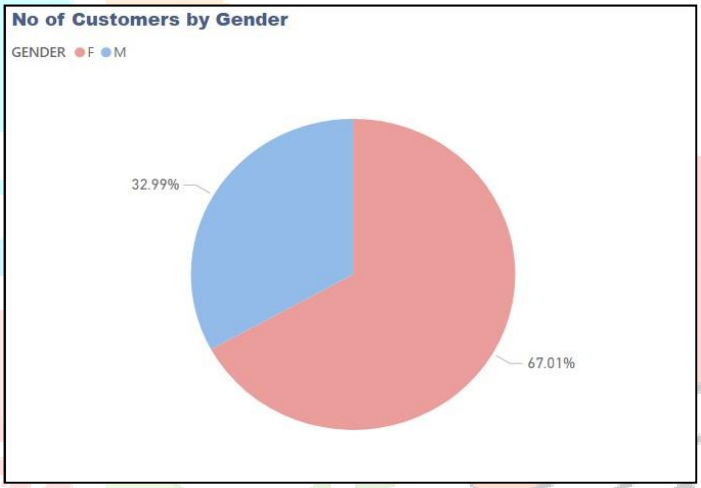
IV. EXPERIMENT AND ANALYSIS

Thischapterdescribesimplementationandresultsevaluationactivitiesindetail.Listofactivitiesdiscussedinhereareexplanatoryanalysisofdata,datapreparationactivities,modelsbuilding,evaluation of models and deployments.

ExplanatoryDataAnalysis

Graphical and numerical representation of data provide better insight about particular data set.Graphicalrepresentationof our dataset is described below.

Figure:1–Noof CustomersbyGender



According to the figure 1 distribution of gender of the customers are 67 % female and 32%aremale.

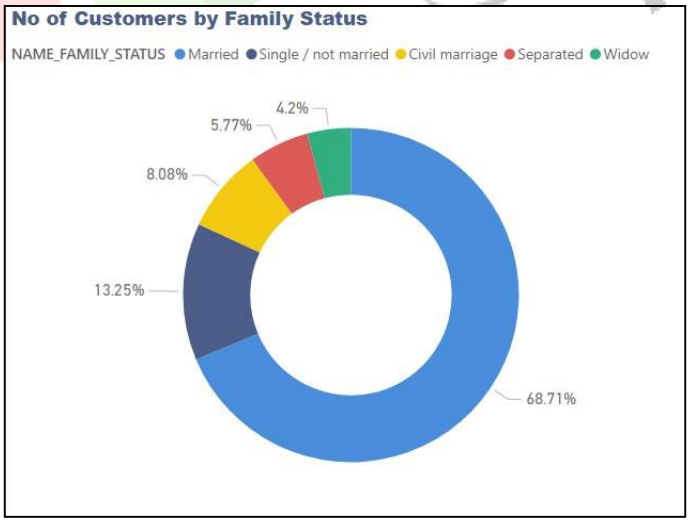


Figure2–Noof CustomersbyFamilyStatus

According to the 4.2 graph No of Customers by family status 69% are married, 13% aresingle,8% are civil marriage, 5%areseparated and 4%arewidows.

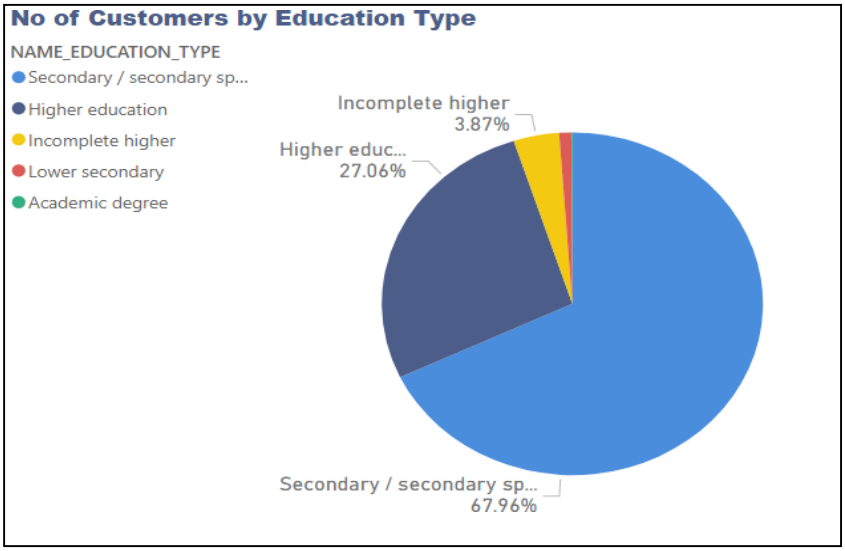


Figure3–NoofCustomersbyEducationType

Asshowninfigure3NoofCustomersbyEducationTypefigure68%havesecondaryeducation,27%arehavehighereducationand 4% arehaveincompletehigher education.

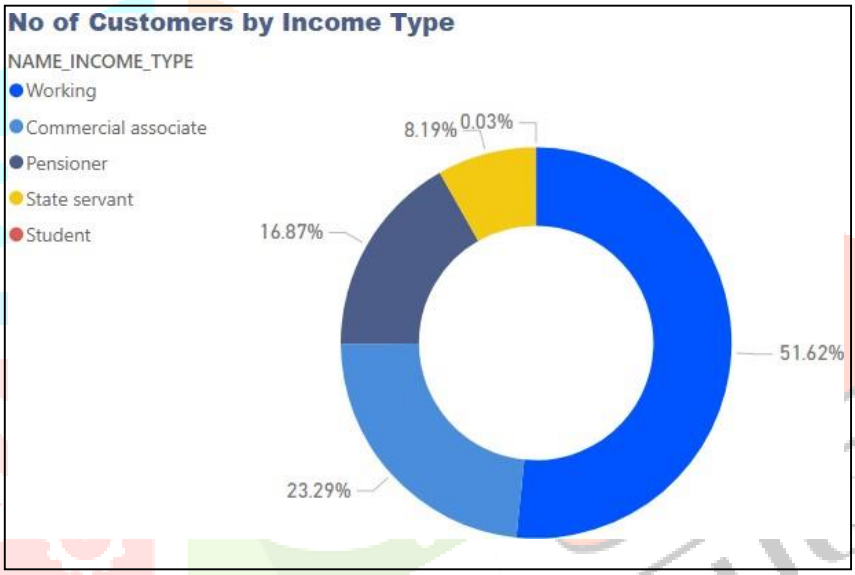


Figure4–NoofCustomersbyIncomeType

Accordingtothegraph4.4NoofCustomersbyIncomeType,62%areworking,23%arecommercialassociates,16%arepensioner, 8%Stateservant.

IV. CONCLUSION

We have obtained the publically available data set and explanatory analysis was carried out tounderstand the data set. Then conducted several activities related to data preparations such asdata preprocessing, feature selections and feature scaling. To achieve a desired outcome, it isvery important to carry out these activities accurately. We have divided the data set into twoparts as a training and test data set and the intended purpose is to validate the accuracy of themodel. Artificial Neural Network, Linear SVM and Nonlinear SVM three predictive modelswere implemented. Performance measures were tested by using Accuracy, Precision, Recall,AUCon eachclassifier.

REFERENCES

1. Agarwal, Abhishek, et al. “Enhancement of Classification Techniques Using Principal Component Analysis and Class Imbalance Handling Methods in Credit Card Defaulter Detection.” *International Journal of Forensic Engineering*, vol. 5, no. 1, 2021, p. 1. *DOI.org (Crossref)*, https://doi.org/10.1504/IJFE.2021.117383.
2. Antonakis, A. C., and M. E. Sfakianakis. “Assessing Naïve Bayes as a Method for Screening Credit Applicants.” *Journal of Applied Statistics*, vol. 36, no. 5, May 2009, pp. 537–45. *DOI.org (Crossref)*, https://doi.org/10.1080/02664760802554263.
3. Banasik, J., et al. “Not If but When Will Borrowers Default.” *Journal of the Operational Research Society*, vol. 50, no. 12, Dec. 1999, pp. 1185–90. *DOI.org (Crossref)*, https://doi.org/10.1057/palgrave.jors.2600851.
4. Orlova, Ekaterina V. “Methodology and Models for Individuals’ Creditworthiness Management Using Digital Footprint Data and Machine Learning Methods.” *Mathematics*, vol. 9, no. 15, Aug. 2021, p. 1820. *DOI.org (Crossref)*, https://doi.org/10.3390/math9151820.

5. Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106. https://doi.org/10.1186/1471-2105-14-106
6. Yadav, Hitesh, et al. "A Novel Hybrid Approach for Feature Selection in Software Product Lines." *Multimedia Tools and Applications*, vol. 80, no. 4, Feb. 2021, pp. 4919–42. *DOI.org (Crossref)*, https://doi.org/10.1007/s11042-020-09956-6.
7. Elreedy, Dina, and Amir F. Atiya. "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance." *Information Sciences*, vol. 505, Dec. 2019, pp. 32–64. *DOI.org (Crossref)*, https://doi.org/10.1016/j.ins.2019.07.070.
8. Karthiban, R., et al. "A Review on Machine Learning Classification Technique for Bank Loan Approval." *2019 International Conference on Computer Communication and Informatics (ICCCI)*,IEEE,2019,pp.16. *DOI.org(Crossref)*,https://doi.org/10.1109/ICCCI.2019.8822014.
9. UCD Michael Smurfit Graduate Business School, Dublin, Ireland, et al. "Credit Risk Prediction Using Artificial Neural Network Algorithm." *International Journal of Modern Education and Computer Science*, vol. 10, no. 5, May 2018, pp. 9–16. *DOI.org (Crossref)*, https://doi.org/10.5815/ijmecs.2018.05.02.
10. Lee, T., and I. Chen. "A Two-Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines." *Expert Systems with Applications*, vol.28,no.4,May2005,pp.74352. *DOI.org(Crossref)*,https://doi.org/10.1016/j.eswa.2004.12.031.