Predictions of Loan Defaulter - A Data Science Perspective

P. Maheswari
Depatment of CSE
Lakireddy Bali Reddy College of Engineering (A)
Mylavaram, Krishna, Andhra Pradesh
potnuri.maheswari@gmail.com

CH. V. Narayana
Depatment of CSE
Lakireddy Bali Reddy College of Engineering (A)
Mylavaram, Krishna, Andhra Pradesh
cvnreddy.chejarla@gmail.com

Abstract—**With the progress of technology and implementation of Data Science in banking, changes the face of banking industry. Most of the banking, financial sectors and social lending platforms are actively investing on lending. But financial institutions might face huge capital loss if they approved the loan without having any prior assessment of default risk. Financial institutions always need a more accurate predictive system for various purposes. Predicting loan defaulters is a crucial task for the banking industry. Banks have immensely large amount of data like customer's data, transaction behavior, etc. Data Science is a promising area to process the data and extract the hidden patterns using machine learning techniques. This paper uses statistical measures to preprocess the data and build an effective model that will predicts the loan defaulter accurately.**

Keywords— Machine Learning, Data Science, Logistic Regression, Random Forest, KNN, LDA

I. Introduction

Finance sector is one of the earliest applications of Data Science. Financial institutions meet bad debts and losses every year. However, they initially perform paper work while sanctioning a loan which yields to get lot of data. Since from last few years banking improves their analysis of identifying the probability of risk through customer profiling, past expenditures, and customer transaction behavior, etc. Data Science is a combination of various statistical tools, algorithms and machine learning techniques to extract the hidden patterns from the data which helps to turn into in-sights. Now-a-days, financial organizations takes the advantage of data science applications, to study the individual customers banking profile, and providing the appropriate services by segmenting the customers based on their credit history.

Banks will receive number of loan applications every day. Loan is the main asset of banks to improve their profitability. However, assessing the risk is one of the major concerns for banks. This paper classifies that, the customer will be defaulter or not, by performing data science process, i.e., data pre- processing, Exploratory data analysis, building the models using machine learning algorithms and finally evaluate the models using various validation metrics.

II. Literature Review

Fraud detection and credit risk applications are the most popular application of Data Science, particularly well-suited to classification technique. Prediction of loan defaults mostly employs classification algorithms. In classification, data is processed into train and test. Training data is used to build model for prediction and test set is used to evaluate the model. In the first step is gathering information, data from previously approved loan datasets are gathered together. In paper [1] uses Exploratory Data Analysis (EDA), is to provide basic insights of any dataset. The main objective of EDA is to extract the essential patterns and visualize them in graphs, plots. In [2] proposed Decision Tree Induction Algorithm to predict the attributes relevant for loan credibility? In this paper a prototype model is built which can be used by the organization in making the right decision to approve or reject the loan.

In [3] authors were proposed clustering mechanism to improve the accuracy of defaulters in banks based on probability. The experimental results were obtained using KNN algorithm and it is implemented in R.

In [4] states authors stated problem statement with the class imbalance problem. Various approaches are discussed to handle the class imbalance problem. This classification follows binary method which results the output in one of the two variables either default or not. In [5] proposed Naive Bayesian classifier to classify the loan defaulters which is quickly produce the results. It assumes that all the input variables are independent and calculate the prior and posterior probabilities. The naïve Bayes classifier is particularly appropriate when the dimensionality of the inputs is high. Along with its simplicity Naïve Bayes is one of the most sophisticated classification techniques. It well suited to credit-risk manager domain.

To the best of our knowledge, this paper addresses the very popular and novel research works studied in detail is shown in Table 1. The growth of the time series data is increasing dramatically. Furthermore, there are several tools to predict or forecast the time series accurately. Although this is not a clear research objective, but it is interesting to be able to develop more real-time forecasting algorithms and tools.

III. Proposed Methodology

The Data Science process revolves around using machine learning and other analytical methods to produce insights and predictions from data in order to achieve a business objective. The entire process involves several steps like data cleaning, preparation, modeling and model evaluation shown in Fig. 1.

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

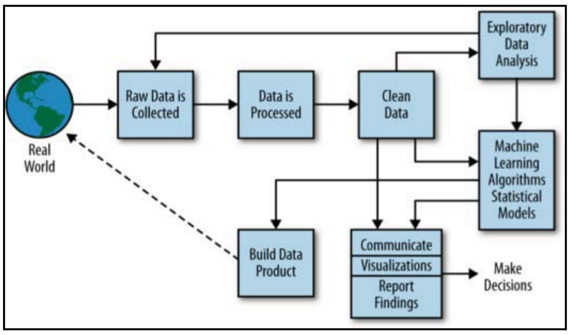

Fig 1. Data Science Process

A. *Business Understanding*

The data collection or the analysis completely depends on the objective need to satisfy. For that initially need to gain the domain knowledge of the particular aspect.

B. *Data Understanding*

Once the business problem is identified and understood, the next step is to gather the data from the various resources. In these days most of the data can be available through multiple resources, the forward step is to understand the data. Somehow in few cases we cannot gather data directly, need to gather data from the ground level. Both the cases can be achieved when we have domain knowledge on particular objective.

C. *Data Preparation*

Data preparation is the most time- consuming process in the process of data science. In data preparation, there are several steps to pre-process the data like, selecting the relevant features, identifying the noisy data, imputing the missing fraction of data using imputation methods, finding outliers and handling them. Creating new set of data from the existing features. This is the major step in data science process because how good will the data was pre-processed results the good models for accurate outcome.

D. *Exploratory Data Analysis (EDA)*

In EDA, the data will be present more effectively. Using statistical measures or metrics we can able to find the meaningful patterns. Further, we can visualize the results in plots, graphs, and histogram, scatter plot, etc. A lot more work can be done in EDA, to tune the data like dimensionality reduction, sampling data and class imbalanced problems to model the data.

E. *Data Modeling*

Data modeling stage is closely related with problem understanding and data understanding, because we need to understand whether the problem statement can be solved through a classification or regression technique. Without understanding the problem we cannot perform the data modeling. After the model built, need to tune the parameters more precisely to get best accurate results.

F. *Model Evaluation*

It is very important to evaluate a model before applying in real-time. The cost and time will depend on this stage. The model is evaluated by using existing data or with new data, how well the model is evaluated, results stable consistency when applied on different platforms.

G. *Model Deployment*

Finally, the model will be deployed into real time application which should produce results according to real-time basis. If any of the above steps goes improperly, all the steps are iteratively repeated frequently.

IV. Implementation

This section describes the implementation details of entire process includes data collection, preprocessing, feature extraction and finally the model building. The model will be validated using performance evaluation metrics.

A. *Dataset Collection*

This paper uses lending club loan dataset available in Kaggle. The dataset was composed of 1.6 million records and 150 features. The entire dataset is in un-processed form consists of categorical data and descriptions. This dataset is re-al-time data which illustrates loan administration experience within the US so-cial circle small business.

B. *Data Pre-processing*

In data preprocessing, initially missing values with 30% of records are re-moved and imputed by using standard imputation methods. Fig. 2 represents the missing fraction of data. Exploratory data analysis (EDA) on each variable will be performed. The EDA helps to extract the hidden patterns. Some of the observations are plotted in Fig. 3 & Fig. 4, where the Sub-Grade plot identifies that as grade worsens the rate of loan defaulters are increased and the Home-Owner ship plot identifies that more home owners and renters are tend to be loan defaulter. Then, identifying the categorical data and converted into shape that can be processed by machine learning techniques.

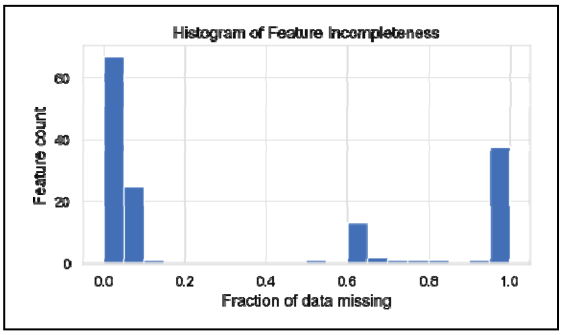

Fig. 2. Missing Fraction of Data

Selection of relevant features using sampling techniques like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). In this study, PCA and LDA both techniques are used for feature selection but, LDA technique, is more likely to select the relevant features which yields to generate accurate models.

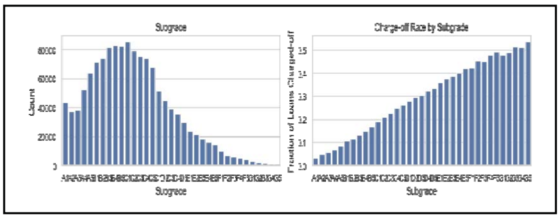

Fig. 3. Increase of Loan defaulters as Sub-Grade worsens

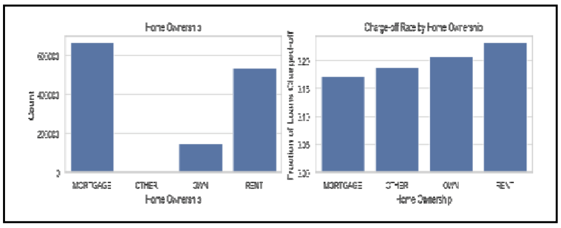

Fig. 4. Home Owners and Renters has high Loan default rate

C. *Feature Engineering and Data Splitting*

Feature Engineering involves feature scaling and feature selection. Through Machine learning pipelines feature engineering had performed using normalization and standardization techniques.

D. *Model Building and Performance Metrics*

Dataset splitting is performed based on issue date of loans, considering earliest date issued loans. This paper worked on large amount of records, through GridSearchCV technique built the model instead of default algorithm process.

a) Logistic Regression with Stochastic Gradient Descent (SGD): The logistic regression algorithm implements using SGD method to process better on larger dataset. This model is a binary classification, measures the relationship between dependent and independent variables to predict the probability of (target variable) loan default.

b) Random Forest: Random Forest is an ensemble classification technique. From the previous studies random forest works better than logistic regression and efficient for processing large datasets. In this work model built on training dataset without standardization which results little low accuracy but classification metrics results similar to logistic regression.

c) K-Nearest Neighbor (KNN): The KNN algorithm is used to solve both classification and regression problems. This algorithm requires feature-scaling. The algorithm results same as logistic regression but the main drawback is, it takes more processing time for lagers datasets.

V. Experimental Results

This section provides the overall performance of the models that are tabulated in Table 1. The models are evaluated using AUC ROC value metric. All the three models results similarly but Logistic Regression model results effectively in terms of accuracy and AUC ROC score is 0.71 that visually shown in the ROC Curve plots in Fig. 5. The final model logistic regression using SGD was finely tuned by adjusting hyper parameters using best parameters and validated with test set, results 0.69 mean-cross validated AU ROC score.

TABLE 1. Performance of Models

Model	Performance Metrics		
	Accuracy	***Precision***	***Recall***
Logistic Regression	0.80	0.81	0.97
Random Forest	0.79	0.81	097
KNN	0.78	0.81	0.97

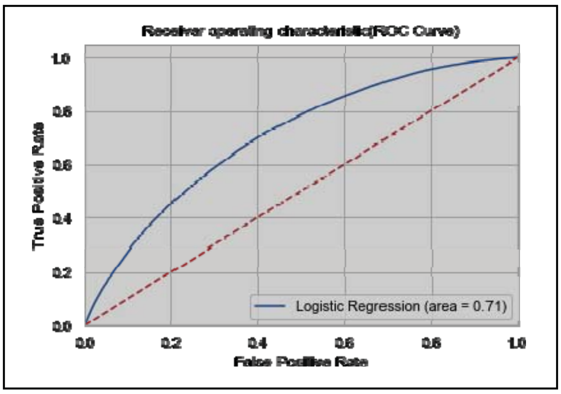

Fig. 5. ROC Curve of Logistic Regression Model

VI. Conclusion

This paper discussed how data science can impact the banking sector to improve their analysis of identifying risk by preprocessing the historical data of customers and building the model using machine learning techniques. Due to huge volume data processing, built the models in cross validation approach using GridSearchCV. The classification techniques logistic regression, random forest and KNN model are built, so far three algorithms results similarly. Among them Logistic regression with SGD training results better predictions than the others.

Few observations made while performing EDA,

- Even income source verified sanctioned loans have higher probability of loan default.
- As the subgrade worsens, there is high possibility of loan default, so we can consider the subgrade feature instead of overall grade feature.
- From the home ownership feature, it is identified that we cannot consider the own home ownership as highest priority to approve a loan.

REFERENCES

[1] M. S. Sivasree, “Loan Credibility Prediction System Based on Decision Tree Algorithm,” Int. J. Eng. Res. Technol., 2015.

[2] Aida Krichene,” Using a naive Bayesian classifier methodology for loan risk assessment,” Journal of Economics, Finance and Administrative Science, 2017

[3] Bagherpour, “Predicting mortgage loan default with machine learning methods,” Univ. California / Riverside, 2017.

[4] Namvar, M. Siami, F. Rabhi, and M. Naderpour, “Credit risk prediction in an imbalanced social lending environment,” arXiv Prepr. arXiv1805.00801, 2018

[5] Goyal and R. Kaur, “Loan Prediction Using Ensemble Technique.,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 5, no. 3, pp. 523–526, 2016.

[6] X.Francis Jency, V.P.Sumathi, Janani Shiva Sri , “An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients”

[7] Dr. K. Chitra1, B. Subashini , “Data Mining Techniques and its Applications in Banking Sector “ , International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal,Volume 3)

[8] Semiu A, Akanmu Abdul Rehman Gilal August 2013), “A Boosted Decision Tree Model for Predicting Loan Default in P2P Lending Communities”

[9] S. Raschka and V. Mirjalili, Python machine learning. Packt Publishing Ltd, 2017.

[10] Pandit, “DATA MINING ON LOAN APPROVED DATSET FOR PREDICTING DEFAULTERS,” Rochester Institute of Technology, 2016.

[11] G. Sudhamathy, “Credit risk analysis and prediction modelling of bank loans using R,” Int. J. Eng. Technol, vol. 8, pp. 1954–1966, 2016.

[12] M. Li, A. Mickel, and S. Taylor, “Should This Loan be Approved or Denied?: A Large Dataset with Class Assignment Guidelines,” J. Stat. Educ., vol. 26, no. 1, pp. 55–66, 2018.

[13] Mahesh Marodkar, “Loan Defaulter’s Application in R programming”.

[14] Shaath, “Credit Risk Analysis and Prediction Modelling of Bank Loans Using R”.