Loan Default Prediction Using Spark Machine Learning Algorithms

Aiman Muhammad Uwais[1 [0000−0002−6306−420X]] and Hamidreza Khaleghzadeh[1 [0000−0003−4070−7468]]

School of Computing, University of Portsmouth, Portsmouth, PO1 3HE, United Kingdom
aiman.uwais@myport.ac.uk, hamidreza.khaleghzadeh@port.ac.uk

Abstract. Loan lending has been an important business activity for both individuals and financial institutions. Profit and loss of financial lenders to an extent depend on loan repayment. Though loan lending is beneficial for both lenders and borrowers, it does carry a great risk of the inability of the loan receiver to repay back the loan. This inability is termed as loan default. Loan default prediction is a crucial process that should be carried out by financial lenders to help them find out if a loan can default or not. Successful loan default prediction can help financial institutions to decrease the number of bad loan issues and eventually increase profit. The aim of this paper is to use data mining techniques to bring out insight from data then build a loan prediction model using machine learning algorithms on the Sparks Big Data platform. Six supervised machine learning classification algorithms are applied to predict loan default: Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree (GBTs), Factorization Machines (FM) and Linear Support Vector Machine (LSVM). Accuracy, precision, recall, ROC curve and F measure are used to evaluate the models and the results compared. We achieve the highest accuracy of 99.62% using the Decision Tree and Random Forest Models.

Keywords: Loan default · Prediction · Machine learning · Big Data · Spark.

1 Introduction

With increasing competition in the financial world and due to severe financial constraints, taking a loan has become certain. Individuals and organizations rely on loans for reasons such as overcoming financial limits to achieve their personal goals or for the basic purpose of managing their affairs in times where there are financial constraints [7]. Though loan lending is quite beneficial for both the lenders and the receivers and is considered an essential part of financial transactions, it does carry some great risks [1]. This risk is termed credit risk or loan default.

Murray defines loan default as when a borrower does not make required payments or does not comply with the terms of a loan. Profit or loss of the financial

lender to a large extent depends on loan repayments, that is whether customers are paying back the loans or not (defaulting). Therefore, when loans default, financial institutions will lose money, and it might even lead to bankruptcy and collapse of the institution. By predicting loan default, financial institutions (lenders) can reduce credit risk, prevent loan default and increase profit by evaluating the ability of the borrower to deliver on their obligation of loan repayment i.e. loan default prediction [9]. The process of forecasting when a loan will default or not was initially done manually or semi-manually. With the advancement of statistical computing packages, several machine learning algorithms are used to calculate and predict loan default by evaluating an individual's historical data. But with an ever-increasing amount of data for loan default prediction, there is the need to use Big Data applications. In this paper, we solve this problem by building a high-performing machine learning classifier model using Apache Spark machine learning libraries to predict loan default.

This paper aims to demonstrate the application of Big Data and machine learning in the finance industry. First, exploratory data analysis using data mining techniques is carried out to bring out insights from the dataset. Secondly, we employ Apache Spark machine learning libraries to make accurate loan detail predictions. Six supervised machine learning classification algorithms are applied to predict loan default, and we achieve the highest accuracy of 99.62% using the Decision Tree and Random Forest Models.

The structure of the paper is as follows. Section 2 presents related work. Section 3 describes the research dataset, some exploratory data analysis and data preparation. Modelling is illustrated in Section 4. Section 5 evaluates and compares the presented models. Finally, Section 6 concludes the paper.

2 Related Work

In this section, we review different types of mechanisms that have been employed for loan default prediction on different platforms/architectures.

Wang et al. present a study that uses 4000 samples and 21 attributes to build and evaluate a classifier predictive model. Four algorithms are used: classic SVM, Backpropagation Neural Network, C4.5 and R_SVM. The result shows that the total predicting accuracy of R_SVM is better than other methods [16].

Reddy and Kavitha [12] use neural networks through attribute relevance analysis in testing class defaulter. Hassan and Abraham [4] use a bank dataset which has 1000 cases; each case with 24 numerical attributes to develop and compare models produced from different training algorithms, scaled conjugate gradient back-propagation, Levenberg-Marquardt algorithm and One-step secant back-propagation (SCG, LM and OSS). The study shows that the slowest algorithm is OSS and the best algorithm is LM because it has the largest R, but that means that is the best for this dataset.

Hamid and Ahmed [3] propose a model for classifying the application of loans to good and bad loans using three algorithms; J48, Bayesian network and Naive Bayes classifier. They use the Weka application for the implementation

and testing. They show that J48 has the best accuracy of 78.378%. Turkson et al. [15] applied 15 different types of machine learning algorithms to predict customers' creditworthiness. The experiment shows that, apart from the Nearest Centroid and Gaussian Naive Bayes, the rest of the algorithms performed well in terms of accuracy and other performance evaluation metrics. Each of these algorithms achieved an accuracy rate between 76% to over 80%.

Odegua proposes the use of the Extreme Gradient Boosting algorithm called XGBoost for loan default prediction. The prediction is based on loan data from a bank with datasets containing 4368 samples and 10 attributes from both the loan application and the demographic of the applicants. Location and age of customers are the two most important features that affect loan default. The XGBoost model had an accuracy of 79%, precision (97%), Recall (79%) and F1 score (87%). Conclusively, the paper provides an effective basis for loan credit approval to identify risky customers from many loan applications using predictive modelling [10]. Lai classifies and predicts loan default using a real-world dataset of 132,029 instances from an international bank using AdaBoost, XGBoost, random forest, multi-layer perceptron and k-nearest neighbours. The experiment shows that boosting algorithms performs better with the AdaBoost method achieving 100% prediction accuracy outperforming the others. ROC and AUC evaluation metrics are used in the model evaluations. Based on the outcome obtained, it is concluded that the application of machine learning techniques is promising in the financial industry [6].

Mohammad et al. present a study on loan prediction by building a logistic regression with a sigmoid function model and analysing the problem of predicting loan defaulters. Logistic Regression models are built, and the different measures of performances are computed. The models are compared based on the sensitivity and specificity performance measures. The best-case accuracy obtained is 81.1%. The researchers made the conclusion that the logistic regression method efficiently detects the right customers to be targeted for granting loans [14].

Patel et al. use various data mining algorithms to predict the likely defaulters from a dataset that contains information about home loan applications, thereby helping the banks to make better decisions in the future. The dataset used has 640,000 instances and 14 attributes. Optimum results are obtained using Logistic Regression, Random Forest, Gradient Boosting and CatBoost Classifier. CatBoost classifier and Gradient Boost provide almost equal accuracy with the Gradient Boosting process giving better results of 84.035%. The researchers concluded that these models can be used to make better decisions on loan applicants in predicting loan default and save financial institutions from undergoing huge losses [11].

Meer uses a dataset consisting of 5,960 records. Two models are built using tuned Logistic Regression algorithms, one model using a tuned Random Forest classifier algorithm and one model using a tuned Gradient Boosting Tree algorithm [8].

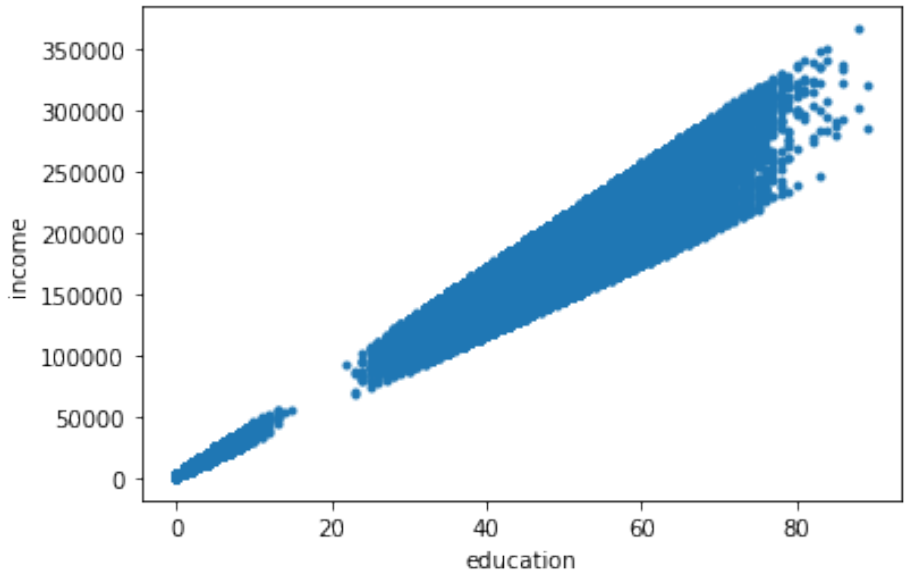

Fig. 1. Scatter plot of income vs education.

3 Research Dataset

3.1 Dataset Characteristics

The dataset has been obtained from the Kaggle website and created for pedagogic purposes for a common loan default prediction task. The data is generated in such a way that default prediction machine learning models are likely to be biased against women and minorities [5]. The dataset contains 640,000 instances and 14 features with the default attribute as the target feature. The other features are the minority, sex, ZIP, loan size, payment timing year, rent, education, income, job stability and occupation. We use 70% of the dataset for training and the rest is used for testing. The dataset consists of two class default labels; true and false.

Figure 1 displays the scatter plot of income against education using pandas matplotlib python function available on Spark. The figure shows that as the educational level increases, the income of the applicants increases. So, it implies that the people with a higher level of education have higher incomes. This depicts a positive correlation.

Figure 2 shows the histogram distribution of the amount of loan size taken. For the first 1000 rows of the dataset around 5000 is the highest amount issued.

Figure 3 displays the default class for the two gender types. Non-defaults (false or blue bar) are the highest meaning for both gender types, more people were able to repay their loan as compared to those that defaulted.

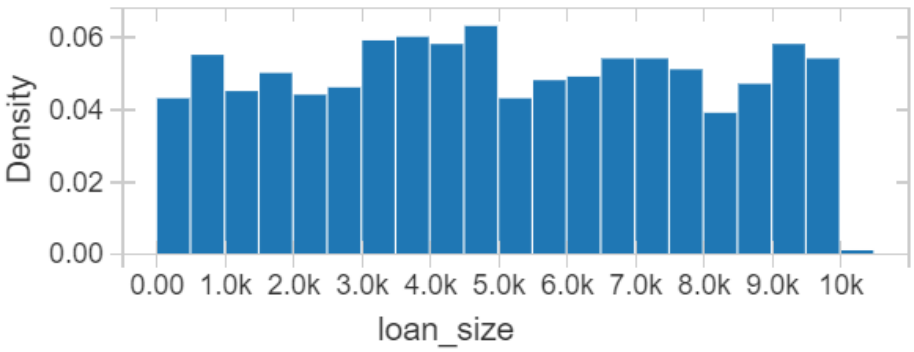

Fig. 2. Distribution of loan size for the 1000 rows in the dataset.

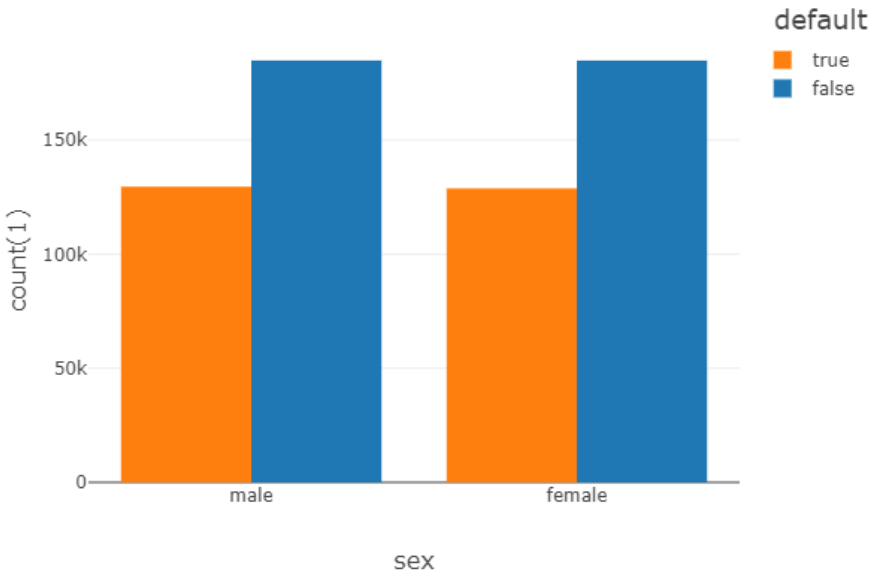

Fig. 3. Default class for sex attribute categories.

Figure 4 displays the number of minorities that paid up their loan (false) and those that defaulted (true). At a glance, the highest number of people that defaulted is the minority ethnic groups while those that did not default are the non-minorities.

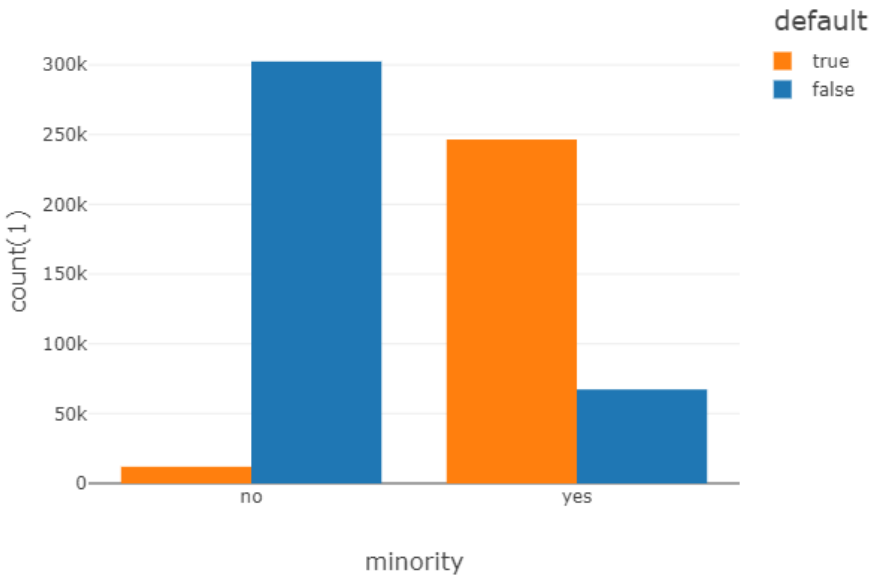

Fig. 4. Minority default count.

3.2 Exploratory Data Analysis

In this section, exploratory Data Analysis using data mining techniques is carried out on the data to bring out some insights. Figure 5 shows the correlation Heatmap of the dataset. A correlation heatmap displays a 2D correlation matrix between two discrete dimensions, using coloured cells to represent data on what is usually a coherent scale. The rows show the values of the first dimension, while the column shows the second dimension. The colour of the cell is proportional to the number of measures that correspond to the value of the dimension. Based on the heatmap diagram in figure 5, one can conclude that the target variable (default) is most positively affected by some features such as rent and negatively correlated with job stability.

3.3 Data Preparation

Data preparation is the process of preparing the raw dataset to be suitable for the machine learning algorithms. The initial pre-processing task entails null value removal and attribute data types adjustment. The data preparation steps are listed below:

1. Feature selection: The attributes that influence the prediction are selected based on the correlation Heatmap and other social factors that influence loan default. The attributes that are selected and used for model building

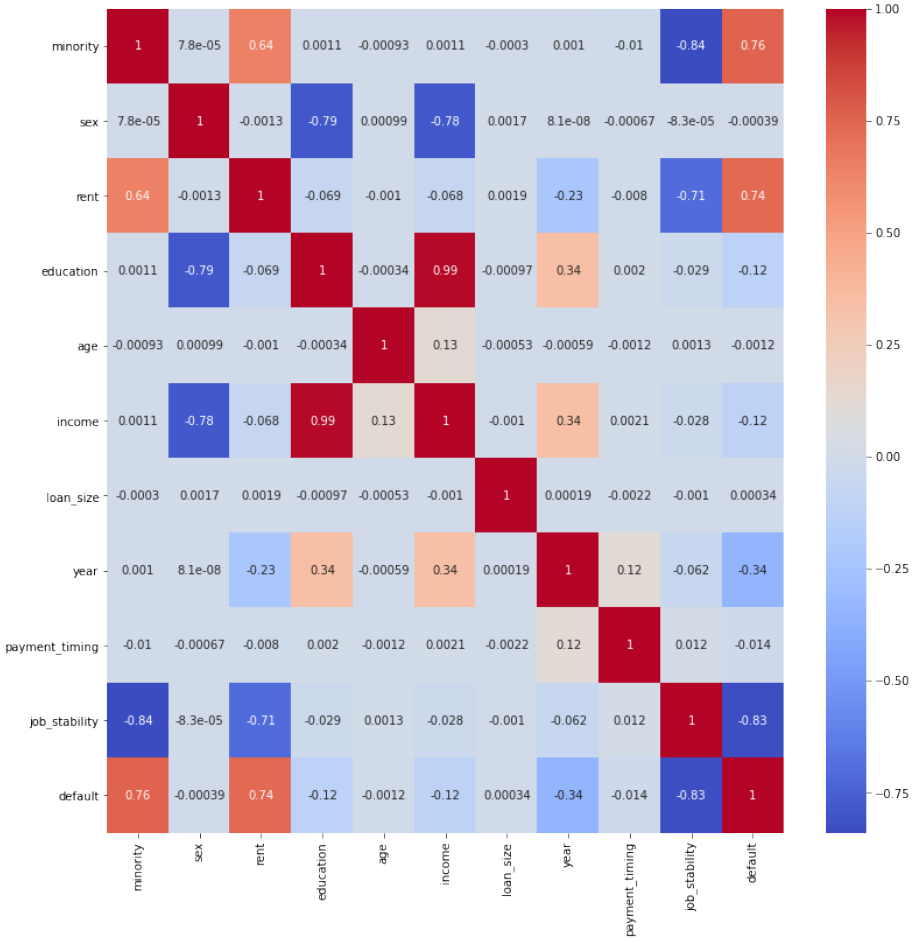

Fig. 5. Correlation Heatmap of the dataset.

are: minority, sex, rent, education, age, income, loan size, payment timing, job stability, year, default.
2. Addressing class imbalance problem.
3. Converting the categorical attributes into a numerical format.
4. Splitting the dataset into training (70%) and test (30%) sets.

4 Modelling

4.1 Implementation Platform

With the ever-increasing amount of data that is in gigabytes and terabytes generated by the financial institution to evaluate loan default, there is the need to

use Big Data applications that will efficiently and accurately predict loan default no matter the quantity of the data. Apache Spark contains machine learning libraries and is the most suitable Big Data application to carry out this task. Apache Spark is a unified computing engine and a set of libraries (framework) for parallel Big Data processing. It supports widely used programming languages (Python, R, etc.), libraries (SQL, streaming, machine learning, etc.) and can run from laptops to server clusters. Sparks provides a unified platform for developing Big Data applications [2]. It also has a machine learning library, known as MLlib, to perform a variety of machine learning tasks.

4.2 Modelling Algorithms

As explained earlier, loan default is the inability of a borrower to pay back his loan. So, when a loan default is true, it means the borrower has defaulted and cannot pay back the loan or meet up with the terms of the loan. However, if the loan default is false, it implies the borrower can meet up with obligation and pay back the loan.

Since the problem we are trying to solve aims to successfully classify values between two categories, true and false, the problem falls within the binary classification problem. Six supervised machine learning classification algorithms that are available on spark MLlib namely, Logistic Regression (LR), Decision Trees (DT), Random Forests (RF), Gradient Boosting Trees (GBM), Factorization machine (FM) and Linear Support Vector Machines (LSVM) are applied with the training data used to train the models and the testing data used to evaluate the models. Table 1 presents a brief description of the algorithms. We have used the Spark MLlib API on the Databricks environment to implement the models.

5 Model Evaluation and Result

This section shows the model performance evaluation where 30% of the whole dataset is used for model testing and evaluation. Model evaluation is an important component of model development in which it evaluates the performance of the developed predictive models. Therefore, we carry out a comparison of the performance of the proposed models.

For model evaluation, we consider ROC curves, accuracy, recall, precision and F score derived from confusion matrices. A ROC curve is a graphical method that shows the sensitivity and specificity of a classifier model. A confusion matrix is a fundamental two-dimensional matrix that contains information about the actual and predicted categories of the classifier. Accuracy, recall, precision and F-score are then obtained from the confusion matrix parameters: true positive (tp), true negative (tn), false positive (fp) and false negative (fn).

5.1 Result Comparison

Table 2 shows the overall evaluation metrics for the six developed machine learning classification models for loan default prediction. From the table, the accuracy

Table 1. Table of machine learning algorithms used.

Algorithm	Description
Logistic Regression (LR)	LR is a supervised classification algorithm used to predict a categorical response by predicting the likelihood of outcomes. It is a generalized linear model that predicts the likelihood of outcomes. Logistic regression is one of the features available in spark.ml that can be used to predict a binary or multiclass outcome by using binomial or multinomial logistic regression respectively.
Decision Tree (DT)	DT is one of the most common supervised learning techniques used to solve classification problems. It has a structure of a tree with node and leaf representing features and class labels respectively. It is easy to understand, less data cleaning is required, and non-linearity does not affect the model's performance, but it may have overfitting problems.
Random Forest (RF)	RF is flexible and easy to apply supervised ML algorithm that produces, even without hyper-parameter tuning, a top-notch result maximum of the time. It can be used for both classification and regression duties.
Gradient-Boosted Trees (GBTs)	GBTs classifier is a supervised learning classification algorithm. It is a collection of trees that trains a set of decision trees with "weak" constraints and uses boosts to combine predictions.
Factorization Machines (FM)	FM is a supervised machine learning algorithm that is used to solve classification problems. Interactions between features are estimated even in problems with huge sparsity. The spark.ml implementation supports factorization machines for binary classification and regression.
Linear Support Vector Machine (LSVM)	LSVM classifier is a supervised machine learning algorithm use to solve classification problems. SVM builds a hyperplane or set of hyperplanes in high or infinite-dimensional space that can be used for classification, regression, or other problems. Intuitively, a good separation is achieved by the hyperplane that has the greatest distance to the closest training data points of any class.

values for DT, RF, and GBTs are almost similar and are the highest (99%) but when critically examining the top values, the RF classifier model has the highest accuracy of 99.619%. We cannot measure the performance of machine learning models only based on their accuracy. In this research, other evaluation metrics like precision, recall and F measure are also considered.

As shown in Figure 6, the DT and RF classifiers have the highest ROC curve value of 99.56%, precision 99.8%, recall 99.2% and F-score of 99.5%. It implies that these two models perform better in loan default classification than the remaining models. The main reason for this observation is that both DT and RF works well for categorical and numerical values, also missing values does not

Table 2. Performance evaluation metrics for the Spark MLlib based models.

	LR	DT	RF	GBTs	FM	LSVM
AreaUnderROC	0.960	0.996	0.9967	0.990	0.882	0.954
tp	72336	76521	76528	75537	73039	71514
fp	1990	122	123	49	20371	2128
fn	4784	599	592	1583	4081	5606
tn	108879	110747	110746	110820	90498	108741
Accuracy	0.9646	0.996	0.996	0.991	0.870	0.959
Precision	0.973	0.998	0.998	0.999	0.782	0.971
Recall	0.938	0.992	0.992	0.979	0.947	0.927
F-score	0.955	0.995	0.995	0.989	0.857	0.949

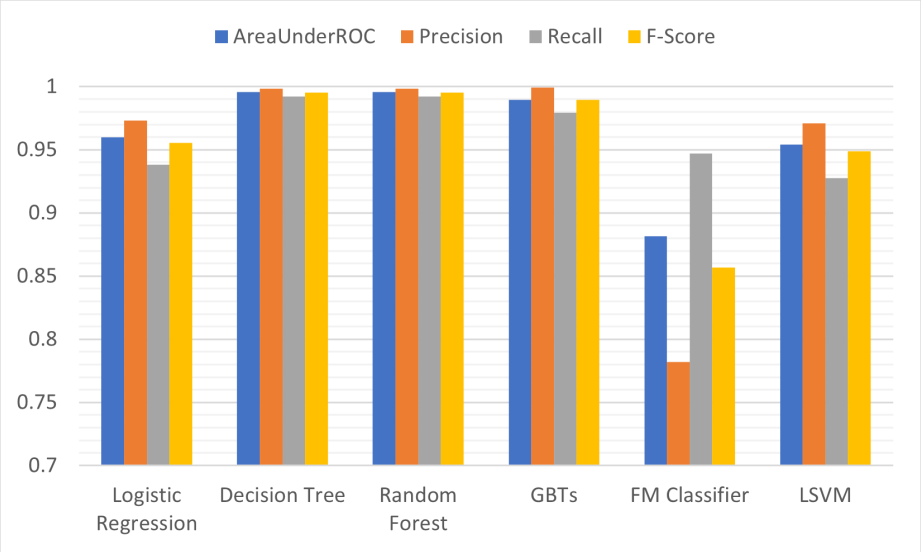

Fig. 6. Performance evaluation.

affect their performance. Following these two models, we have GBT, LR, LSVM and the FM classifiers with F-scores of 98%, 95%, 94% and 85%, respectively. In this research, the factorization machine model is outperformed by the other models. It might be due to the poor performance of FM on dense data. However, FMs perform best in data with high sparsity [13].

As summarised in Table 2, the Decision tree and Random forest models show better performance (highest ROC scores) with a value of 99.5% while the Factorization machine model gives the least performance (lowest ROC scores) with 88.16%. This result is confirmed by the ROC plots displayed in Figure 7.

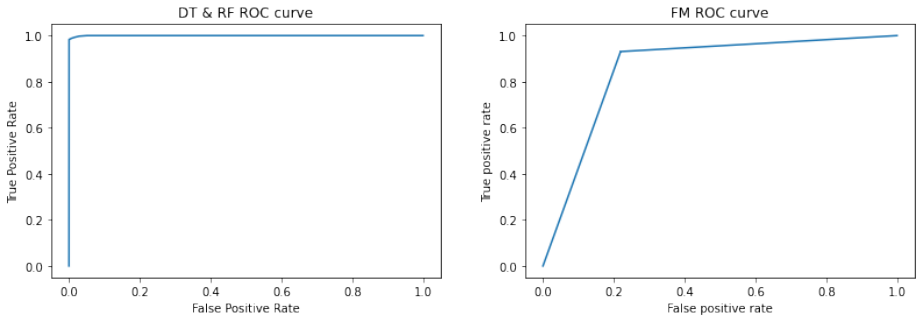

Fig. 7. ROC plots for the DT, RF and FM algorithms.

Figure 7 shows ROC plots for the DT, RF and FM algorithms proposed in this paper. The Area Under a ROC curve (ROC) is the expectation that a model will give an elevated grade to a randomly selected positive class data point than a randomly selected negative class data point. A ROC curve that traces a diagonal line signifies a poor classification algorithm, and it will randomly guess if a loan will default. In addition, a good performing model will have ROC closer to 1 while the ROC will be closer to 0.5 for poor-performing algorithms [8]. Therefore, as shown in Figure 7, the plot for DT and RF ROC curve is towards 1 showing a better model performance compared with the ROC curve of FM.

6 Conclusion

In this paper, the application of data mining and Big Data techniques in building loan default predictors is studied. Six models were developed using the Spark MLlib API allowing us to obtain the best performing model for loan default prediction. Based on the model evaluation, The random forest model presented the highest accuracy (99.619%). Also, the Random forest and Decision tree models have the best performance in terms of ROC curve (99.56%), precision (99.8%), recall (99.2%) and f-score (99.5%). Therefore, it can be concluded that decision tree and random forest classifier models are the most efficient and accurate in predicting the binary categories of loan default.

Based on the results obtained, Spark machine learning library-based models have shown a promising result in the prediction of loan default in this research. It allows financial institutions (lenders) to be informed of default in issued loans beforehand which will help them reduce financial loss and the cost associated with loan recovery. This will increase profits.

References

1. Adewusi, A.O., Oyedokun, T.B., Bello, M.O.: Application of artificial neural network to loan recovery prediction. International Journal of Housing Markets and

Analysis (2016)
2. Chambers, B., Zaharia, M.: Spark: The definitive guide: Big data processing made simple. " O'Reilly Media, Inc." (2018)
3. Hamid, A.J., Ahmed, T.M.: Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAIJ) Vol **3**(1) (2016)
4. Hassan, A.K.I., Abraham, A.: Modeling consumer loan default prediction using neural netware. In: 2013 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRICAL AND ELECTRONIC ENGINEERING (ICCEEE). pp. 239–243. IEEE (2013)
5. Klaas, J.: Loan default model trap. https://www.kaggle.com/jannesklaas/model-trap, (Accessed on 13/10/2021)
6. Lai, L.: Loan default prediction with machine learning techniques. In: 2020 International Conference on Computer Communication and Network Security (CCNS). pp. 5–9. IEEE (2020)
7. Marqués Marzal, A.I., García Jiménez, V., Sánchez Garreta, J.S.: Exploring the behaviour of base classifiers in credit scoring ensembles (2012)
8. Meer, K.: Machine learning models for mortgage default prediction in pakistan. In: 2021 International Conference on Artificial Intelligence (ICAI). pp. 164–169. IEEE (2021)
9. Murray, J.: Default on a loan, united states business law and taxes guide national credit act (2005). act no. 34 of 2005, republic of south africa (2011)
10. Odegua, R.: Predicting bank loan default with extreme gradient boosting. arXiv preprint arXiv:2002.02011 (2020)
11. Patel, B., Patil, H., Hembram, J., Jaswal, S.: Loan default forecasting using data mining. In: 2020 International Conference for Emerging Technology (INCET). pp. 1–4. IEEE (2020)
12. Reddy, M.J., Kavitha, B.: Neural networks for prediction of loan default using attribute relevance analysis. In: 2010 International Conference on Signal Acquisition and Processing. pp. 274–277. IEEE (2010)
13. Rendle, S.: Factorization machines. In: 2010 IEEE International conference on data mining. pp. 995–1000. IEEE (2010)
14. Sheikh, M.A., Goel, A.K., Kumar, T.: An approach for prediction of loan approval using machine learning algorithm. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). pp. 490–494. IEEE (2020)
15. Turkson, R.E., Baagyere, E.Y., Wenya, G.E.: A machine learning approach for predicting bank credit worthiness. In: 2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR). pp. 1–7. IEEE (2016)
16. Wang, B., Liu, Y., Hao, Y., Liu, S.: Defaults assessment of mortgage loan with rough set and svm. In: 2007 International Conference on Computational Intelligence and Security (CIS 2007). pp. 981–985. IEEE (2007)