Default Prediction Model: The Significant Role of Data Engineering in the Quality of Outcomes

Ahmad Al-Qerem[1], Ghazi Al-Naymat[2,3], Mays Alhasan[3], and Mutaz Al-Debei[4]
[1]Computer Science Department, Zarqa University, Jordan
[2]Department of Information Technology, Ajman University, United Arab Emirates
[3]King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Jordan
[4]Management Information Systems Department, The University of Jordan, Jordan

Abstract: *For financial institutions and the banking industry, it is very crucial to have predictive models for their core financial activities, and especially those activities which play major roles in risk management. Predicting loan default is one of the critical issues that banks and financial institutions focus on, as huge revenue loss could be prevented by predicting customer's ability not only to pay back, but also to be able to do that on time. Customer loan default prediction is a task of proactively identifying customers who are most probably to stop paying back their loans. This is usually done by dynamically analyzing customers' relevant information and behaviors. This is significant so as the bank or the financial institution can estimate the borrowers' risk. Many different machine learning classification models and algorithms have been used to predict customers' ability to pay back loans. In this paper, three different classification methods (Naïve Bayes, Decision Tree, and Random Forest) are used for prediction, comprehensive different pre-processing techniques are being applied on the dataset in order to gain better data through fixing some of the main data issues like missing values and imbalanced data, and three different feature extractions algorithms are used to enhance the accuracy and the performance. Results of the competing models were varied after applying data preprocessing techniques and features selections. The results were compared using F1 accuracy measure. The best model achieved an improvement of about 40%, whilst the least performing model achieved an improvement of 3% only. This implies the significance and importance of data engineering (e.g., data preprocessing techniques and features selections) course of action in machine learning exercises.*

Keywords: *Default Prediction, Classification, Pre-processing, Prediction, Features Selection, Generic Algorithm, PSO Algorithm, Naïve Bayes, Decision Tree, SVM, Random Forest, Banking, Risk Management.*

Received February 29, 2020; accepted June 9, 2020
https://doi.org/10.34028/iajit/17/4A/8

1. Introduction

A banking institution can be defined as a financial organization that accepts deposits and convey them into different lending activities. Banks provide accepting deposits services and a variety of loan types besides basic investment products. Accepting deposits from the public is the main function in order to use this capital as loans for borrowers. Major functions of a bank can be summarized as follows: Primary functions which consist of accepting deposits (savings, fixed, current), and Granting loans and advances. Secondary functions, which consist of agency functions and general utility functions [14].

Lending is an important revenue channel for banks and financial institutions, and it's rapidly growing, however, lending institutions are still facing major problems evaluating lenders, as traditional credit scoring and credit assessment methodologies are not as effective as needed, therefore, a more reliable risk assessment model for loan default is needed.

Lending is mainly done through loans as a bank service which is considered as one of the main value propositions financial institutions provide and charging interest rate is one of the main revenue streams [1]. Financial institutions provide loans to borrowers (customers) with the promise that they will pay it back; therefore, there is no real guarantee that they will pay back the loan, and if they stop making loan payments, the profit reflected from an interest rate of the loan will be lost. It is very critical for a financial institution to accurately estimate the riskiness level of borrowers in order to determine their eligibility for loans and the appropriate interest rate. Although credit measurement criteria have been modified and advanced throughout the years, granting loans is still considered a very risky process in the banking industry id not managed properly, consequently, it's the most studied and researched area in the banking industry.

Indeed, loans are paid back according to agreed terms and conditions in the promissory note, and failing to do so is known as "Loan default". Loan default prediction relies on analysis techniques that utilize current and historical information, the behavior of credit customer, loan and settlement information to be able to predict the customer's ability to pay back

the loan on time as well as accurately measure bank profitability.

The prediction of loan default is considered a binary classification of defaulters and non-defaulters from loaners, where different data mining and machine learning techniques can be used as an approach for binary classifications. Building a loan default prediction model in machine learning is considered an optimization task, where the ultimate objective of the task is to increase loan default prediction accuracy (i.e. the accuracy of predicting defaulters as opposed to the overall accuracy of the model). While most previous studies focused on the overall accuracy of the prediction model, more attention needs to be given to enhancing the accuracy of the minority class or the defaulters and this is the main focus and contribution of the paper.

In this paper, we aim at enhancing classification efficiency and accuracy through the application of intensive data preprocessing techniques, and applying three features selection algorithms: Information Gain (IG), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) prior to model building. Thereafter, three different types of classifiers will be utilized which are: Naïve Bayes, Decision Tree, and Random Forest.

This paper is organized as follows. Section 2 provides a literature review. Section 3 explains our approach and our research methodology, including an elaboration of data, preprocessing techniques, models and validation measures. Results are showed in section 4, and section 5 offers the conclusions of this paper.

2. Related Work

For default predictions and according to Angelini *et al*. [3]; Neural Networks (NNs) are reliable classifiers to design prediction models. In their study, Angelini *et al*. [3] used two NNs approaches, namely Feed-forward NN with ad-hoc connections and Feed-forward Neural Network with classical topology, using a real-world dataset of small business banks in Italy that consists of 11 attributes for 76 small business lending data. Both approaches produced high prediction accuracy models with low error rates, where the Feed-forward Neural Network with classical topology showed 8.6% error rate, and the Feed-forward Neural Network with ad-hoc connections showed only 4.3% error rate.

In another study conducted by Akrani [2], Decision Tree (DT), Support Adaptive Boosting Model (ABM), Vector Machine (SVM), Linear Regression (LR), NN and Random Forest (RF) algorithms were applied to build loan default predictive model on a banking data from UCI machine learning data repository that consists of 1000 instances and 18 attributes, and showed that SVM has the best accuracy. Nonetheless, SVM showed low run time performance and as such, Akrani [2] tuned the SVM algorithm to only incorporate most important features. Furthermore, the study showed through graphical representation that only three variables out of the 18 variables have a negative effect on the credit; duration of credit repays, amount credited and borrowers age.

In [6], it was demonstrated that Support Vector Machine SVM can be used efficiently for credit rating classification in the banking industry, and SVM predicting accuracy may increase when an increased dataset sample is used, and a proper feature selection approach is applied or normal correlation significant test. A real commercial dataset with 70 samples was used, composed of customer's credit info and credit risk evaluation data, and one-by-one back elimination attribute selection method was used to exclude insignificant attributes. On the other hand, using Multiple Discriminate Analysis (MDA) showed over fitting, and although it has the lowest training errors, testing accuracy was unacceptable. And when using Canonical Discriminant Analysis (CANDISC), applying Cumulative Distribution Function to every attribute helped avoiding the over fitting problem, but accuracy results did not exceed 50%.

Jin and Zhu [12] applied different attribute selection methodologies: Random Forest and Correlation Matrix in the model preparation stage to determine the top attributes that determine loan default, and used 5 different data mining classification techniques for prediction model building; DT (CRT and CHAID), NN (RBF, MPL) and SVM. A 3-year dataset was used (2007-2011), that contains various borrower, loan and credit information, and attribute selection using Random Forest was applied to select variables that plays important role in loan default, and it was concluded that SVM showed the best performance among other models but with slight improvement difference.

Hsu and Hung [11] used different algorithms with ensemble methods to build a loan default prediction model; Scaled Conjugate Gradient back propagation (SCG), Levenberg-Marquardt algorithm (LM), and One-Step Secant back propagation (OSS), and applied different filtering methodologies on dataset of real world credit application cases from a German bank datasets of 1000 cases and 24 attributes, then used different parameters for comparison. They concluded that the LM with PLs filter produced the best model with an accuracy of 92%, followed by SCG with the same attribute filter with an accuracy of 89%, and the best accuracy One-Step Secant back propagation (OSS) could reach was 84% using PLs filter as well.

Normalization is a standard technique employed in data mining tasks on averaging users' ratings. It is used in the proposed algorithm to avoid the big difference in the final rating of RS and addresses cold start problem through combining CF and MF techniques.

Reddy and Kavitha [16] showed that using Neural Networks through attribute relevance analysis to build a prediction model increases the speed of Neural Network and feasible accuracy. A simple Neural Network model was used, and Info Gain algorithm was applied on attributes to eliminate less informative variables.

Chen *et al*. [5] proposed a hybrid under sampling approach DSUS (Diversified Sensitivity Under sampling) that combine k-mean clustering, stochastic sensitivity measure and a robust redial basis Neural Network function, to handle the problem of imbalanced class distribution, and compared results with other models. A real loan default 6-month dataset for a P2P institution in China was used. The dataset contains 25,504 records and 339 variables. The proposed methodology showed improvement in the recall and G-mean over the other resampling models.

Shoumo *et al*. [17] focused on applying appropriate dimensionally reduction approach using Recursive Feature Elimination with Cross-Validation (RFECV) and Principal Component Analysis (PCA), noise handling, parameters tuning, using a grid search with cross-validation and on handling the imbalanced data problem. SVM and RFECV based models showed the ability to outperform other regression and tree-based models when predicting credit risk, where Support Vector Machine, Random Forest, Logistic Regression and Gradient Boosting algorithms were used in this paper. Furthermore, it was shown that using the proper penalty parameters and kernels in SVM models has a major impact on model's performance.

Deng in [7] used lending club dataset for the first two quarters of 2019 to determine the main factors that highly affect default risk, and used Log it model to predict borrowers' default in advance. Log it model was used for less computation due to using large dataset, and correlation coefficient analysis was applied to select the top 20 factors according to impact, and an accuracy of 92.8 was achieved. And the proposed approach showed accuracy increase when running the model for a specific attribute value.

The above research focused on using different classification models and a comparison between them, or added features selection technique to improve prediction, while in this paper we used various preprocessing techniques, multiple classification models and multiple features selection algorithms to cover more than one aspect that might impact prediction accuracy and performance.

3. Approach and Methodology

Data pre-processing is considered a significant and crucial initial step in data analytics and data mining projects. This is because the output of this stage is inputted to the model to obtain final results. Therefore, data preprocessing not only impacts the accuracy of the model, but also it impacts its performance and efficiency [19]. Our approach in this study involved a thorough exploration of data, and the application of multiple preprocessing techniques prior to the classification stage. Once the data was tuned, three different classification algorithm models were used to predict loan default. Thereafter, the results of the competing models were compared with each other in terms of precision, recall and F score. Moreover, the results were also compared with the prediction results of the same model prior to data preprocessing in order to highlight the impact of data preprocessing on the results of the machine learning models. Figure 1 illustrates our approach.

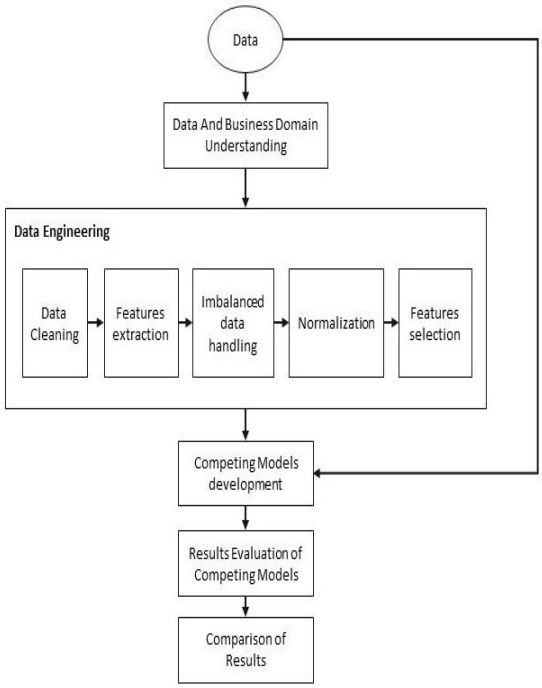

Figure 1. Approach of the study.

A key hidden step within the "Data Engineering" stage is data understanding, and business understanding specially of used data. This is significant as data used to feed the model and it has a high impact of model performance. Using irrelevant data may lead to faulty results. In our experiment; we initially used all available features, and surprisingly the results of all models showed very high performance (above 99%), even before applying any data preprocessing techniques, nor features selection algorithms, which seems to be too good to be true. When investigating by pulling the main features that led to this high performance, the top three attributes were: Recoveries (Post charge off gross recovery), Collection Recovery Fee (Post charge off collection fee) and Total Recovery Percentage (Principal received to date). By applying data and business processes understanding it was determined that those three attributes are biased toward charge off customers (i.e., defaulters), and such information is only

available after the borrower charge off, and would not be available for customers who provide their payments regularly, as they belong to business process that get triggered when the borrower stops loan payments, therefore, they should be excluded from the model. This highlights the importance of data understanding and domain knowledge expertise in machine leaning exercises.

3.1. Data Collection

The data set that is being used for this paper is a loan data set from the lending club, with 145 features and around 43k records. Data includes loan details, e.g. amount, purpose, interest rate, and installments. Customer details, e.g., demographics, employment, Debt-To-Income ratio (DTI), Fair Isaac Corporation (FICO), and credit lines. customer's behavioral details, e.g., revolving balance, revolving utilization, delinquency, and payments. Table 1 represents the data dictionary of final attributes utilized in our model.

Table 1. Data Dictionary of model's features.

Attribute	Meaning
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
grade	LC assigned loan grade
sub_grade	LC assigned loan subgrade
emp_length	Employment length in years
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
issue_d	The month which the loan was funded
Credit_requirements	Indicates if borrower comply with credit requirements
purpose	A category provided by the borrower for the loan request.
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
addr_state	The state provided by the borrower in the loan application
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 year
earliest_cr_line	The month the borrower's earliest reported credit line was opened
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_record	The number of months since the last public record
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
mths_since_last_delinq	The number of months since the borrower's last delinquency.
mths_since_last_record	The number of months since the last public record
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_prncp	Principal received to date
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
recoveries	post charge off gross recovery
collection_recovery_fee	post charge off collection fee
last_pymnt_d	Last month payment was received
last_pymnt_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
next_pymnt_d	Next scheduled payment date
last_credit_pull_d	The most recent month LC pulled credit for this loan
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
pub_rec_bankruptcies	Number of public record bankruptcies
tax_liens	Number of tax liens
debt_settlement_flag	Flags whether or not the borrower, who has charged-off, is working with a debt-settlement company.
debt_settlement_flag_day	The most recent date that the Debt_Settlement_Flag has been set
settlement_status	The status of the borrower's settlement plan. Possible values are: COMPLETE, ACTIVE, BROKEN, CANCELLED, DENIED, DRAFT
settlement_date	The date that the borrower agrees to the settlement plan
settlement_amount	The loan amount that the borrower has agreed to settle for
settlement_percentage	The settlement amount as a percentage of the payoff balance amount on the loan
settlement_term	The number of months that the borrower will be on the settlement plan
loan_status	indicates if the borrower paid the loan in full or defaulted

3.2. Data Preprocessing

Unsurprisingly, real world data is usually incomplete, inconsistent, unclean and might contain many different issues. Hence, data preprocessing becomes very significant as it entails the utilization of various

techniques that transform raw data into more consistent, complete and clean data by resolving data errors issues and discrepancies. This section presents the steps considered for the preprocessing stage.

3.2.1. Data Cleansing

The following data cleansing techniques were applied to the dataset:

- Attributes with almost null values were deleted.
- Attributes with unique values or only one value were deleted.
- Unstructured attributes with long free text were removed at this stage.
- Empty rows were removed.
- "loan status" attribute was used to extract the "loan status" class attribute, and "credit requirement" attribute that indicates whether the borrower meets credit policy requirements or not.
- Date attributes were used for feature extraction, where new features were created according to numbers of days between the date and current date.
- "xx" at the end of the "zip code" attribute was replaced to "00".
- "emp length" attribute was transformed from categorical values (1 year) to numerical values (1), 1 was assigned to "< 1 year" value as records have similar behavior of "1 year" in relevance to the class attribute and 0 was assigned to "n/a".
- Null values in our dataset were handled using imputation techniques where numeric missing data were replaced with 0, categorical with "none" and date missing values were replaced with the current date.

After applying the different techniques mentioned above, the original dataset was reduced to 46 features

3.2.2. Imbalanced Data Handling

Imbalanced data is a common machine learning issue where aggregation of instances in one class is more significant than the total number of instances of the other class, which makes the classifier more biased towards the larger class. Oversampling based approach is one of the commonly used methodologies to overcome this issue [4]. Most previous research studies concentrated on enhancing the overall performance of the model without looking deeply at the low performance of the minority class that is caused by imbalanced data. For binary classification that is carried out for default prediction, such a problem becomes sensitive for decision making [18]. In this paper Synthetic Minority Over-Sampling (SMOTE) approach was implemented to increase minority class 100% using k=2 of K-Nearest Neighbors. SMOTE was chosen as it does not create copies of records from the minority class, instead it creates synthetic samples to increase minority class and makes it equal or close to the majority class.

3.2.3. Data Normalization

Data Normalization is one of techniques that is commonly applied over data as part of data preprocessing in machine learning. The main purpose of normalization is to have a unified scale of numeric feature values in the dataset, without perverting differences in ranges and also keeping the same ratios of values differences. It is required only when numeric features have various value ranges and not for every machine learning modeling.

Data normalization is used to eliminate the unit of measurement of data, for easier data comparison, and it usually means to scale features to have values between 0 and 1. Data normalization is a process that reduces data redundancy [13].

In this study, data normalization was applied to all numeric features using Min-Max approach for easier data comparison and to reduce data redundancy. Thus, all numeric features were having values between 0 and 1.

3.2.4. Features Selection

Classification efficiency, speed, and precision can be improved by decreasing features space, and noise features can be eliminated as well. Information gain is one of the commonly used features selection methodologies that selects key features from the dataset and deletes dispensable once [9]. IG measures the level of "information" a feature provides about the class using "Entropy" measure, where entropy is calculated as follows:

$$\text{Entropy} = -\,pk \log 2\, pk \qquad (1)$$

pk: The proportion of instances belonging to class k ($K = 1, \ldots, k$), and $0\log 2\ 0 = 0$.

When applying IG; features with low entropy had a better ranking, and the top ten features were selected to be used in the classification stage within the model out of the 46 features available within the processed data set.

GA-based feature selection is another feature selection methodology that selects optimal features by randomly choosing the population initially, and check their fitness to the environment using an objective function. It also uses a repeated evolutionary process to improve the population until the optimum is reached [8]. When applying GA on the processed data that contains 46 features; fourteen features were selected as relevant features and the rest were ignored in the classification phase within the model.

Particle Swarm Optimization (PSO) algorithm is the third algorithm that was used in this paper for feature selection. This method was originally inspired by the movement behavior of birds' flock. It is easy to

implement but it is computationally intensive, and a very powerful methodology that has been used in many fields including features reduction; by removing irrelevant and noisy features from the original dataset, and keeping only important and relevant features [1]. When applying PSO; data features space was reduced to ten features out of the 46 features available within the processed data set.

These three features selection approaches were applied independently on the preprocessed dataset prior to classification, and classification efficiency improvement was calculated and compared at the end.

3.3. Models Construction and Validation

In this section, we will briefly define the different data mining classification methodologies used in this paper, and highlight cross-validation evaluation criteria used to compare results in the following section.

3.3.1. Naïve Bayes

A supervised data mining classification technique that assumes independence among predictors. It has less computational complexity and memory requirements than other competing models, but it enjoys good performance and prediction accuracy [10]. Naïve Bayes classifier contemplates the contribution of each feature in independent of correlations between features. It is one of the commonly used classification algorithms as it is simple, efficient and can work on small dataset.

3.3.2. C4.5 Algorithm

A supervised decision tree-based classifier with good performance and accuracy. C4.5 algorithm depends on information gain and gain ratio when constructing the decision tree. Pruning is proposed for this technique to avoid over fitting [20]. In this paper, we built two models using a C4.5 decision tree; an unpruned tree and pruned tree and we compared the results of both techniques.

3.3.3. Random Forest

Random Forest algorithm is also known as a random decision tree. It is a prediction technique for classification and regression problems that achieves the best possible solution by constructing multiple uncorrelated decision trees and uses majority voting for the final result [15]. It is also robust to outliers and noise. Moreover, Random Forest algorithm enjoys good prediction accuracy and handles over fitting properly.

Model validation is an important step to evaluate the quality of the model along with its outcomes. Model validation assists in choosing the model that will perform best over unseen datasets. The optimal model should perform well over both the training and the test datasets and also it should perform well over time. Different strategies and methods are currently being used to evaluate machine learning models according to number of splits being applied over the dataset. Cross validation is one of the commonly used technique to evaluate the stability of machine learning models, and how accurately the model performs in prediction different classes in practice. Cross validation helps in avoiding over fitting and under fitting of the models.

10-folds cross-validation was used in this study to evaluate classification models. Using this methodology, the data is randomly divided into ten sets of data; nine are used as training set, and the tenth is reserved for testing purposes. This process is iterated ten times with different training sets. Figure 2 is used to demonstrate how cross validation works.

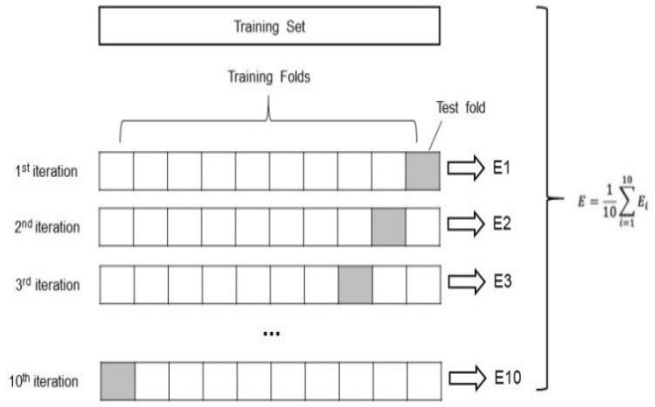

Figure 2. 10-Folds Cross validation across algorithms.

Then precision, recall, and F1 measures were calculated for each model and were used as evaluation criteria. These measures are calculated as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Precision, and it is also known as positive predictive value, is a model evaluation measure in machine learning that calculates the fraction of relevant occurrences among retrieved occurrences. It evaluates how confident and precise the model in prediction by calculating the percentage of relevant results. Recall, on the other hand, which is also known as sensitivity, is the fraction of true positives that were found from the total number of true positives, it measures how good the model is in retrieving all positive occurrences. Consequently, there is a trade off between precision and recall; i.e., obtaining high recall will result in lowering precision. Using one of these measures over the other is dependable on business requirement on what matters more; recalling all

relevant occurrences or be more confident in occurrences that were recalled. A balanced measure of model performance is the F1 measure. It is a combined measure of both precision and recall.

4. Results

Four classification models were built; Naïve Bayes, unpruned C4.5 Decision Tree, pruned C4.5 Decision Tree, and Random Forest. Each one of these models was run four times and evaluation measures were recorded. In the first iteration; models used unprocessed data, and the other three iterations used processed data with three different features selection algorithms as explained earlier. Evaluation measures of used models using unprocessed data set are shown in Table 2 below.

Table 2. Results of unprocessed data measures.

Model	Class	Measures		
		Precision	Recall	F1
Naïve Bayes	Charged off (Defaulted)	87.20%	64.40%	74.10%
	Fully Paid	90.90%	97.40%	94.00%
	Weighted Avg.	90.10%	90.30%	89.70%
C4.5	Charged off (Defaulted)	93.50%	89.50%	91.50%
	Fully Paid	97.10%	98.30%	97.7%
	Weighted Avg.	96.40%	96.40%	96.40%
C4.5 (Pruned)	Charged off (Defaulted)	99.30%	88.00%	93.30%
	Fully Paid	96.80%	99.80%	98.30%
	Weighted Avg.	97.30%	97.30%	97.20%
Random Forest	Charged off (Defaulted)	52.30%	46.80%	49.40%
	Fully Paid	90.70%	92.40%	91.50%
	Weighted Avg.	84.90%	85.50%	85.20%

Random Forest algorithm shows a very low predicting power of defaulted borrowers with an F1 measure of 49.40% when using unprocessed data, followed by Naïve Bayes with F1 of 74.10%. Nonetheless, Tree based algorithms showed a relatively good prediction with F1 measure of 91.50% and 93.30% for C4.5 and pruned C4.5, respectively.

Evaluation measures of Naïve Bayes, C4.5 Decision Tree (unpruned), C4.5 Decision Tree (pruned) and Random Forest after applying data preprocessing are shown in Table 3, Table 4, Table 5, and Table 6, respectively. Where IG, GA and PSO algorithm were used as feature selection methodologies for each classifier.

Naïve Bayes Model showed an increase in prediction power for defaulted borrowers from 71.40% to 78.50% (increase of 7.1%) after applying data preprocessing, and when using either Info Gain or Generic algorithm for features selection. On the other hand, an impact of 6.5% increase in F1 measure for the Charged-off class prediction occurred when using PSO for features selection.

Table 3. Results of Naïve Bayes classifier.

Model	Algorithm	Class	Measures		
			Precision	Recall	F1
Niave Bayes	Info Gain	Charged off	99.50%	64.80%	78.50%
		Fully Paid	88.80%	99.90%	94.00%
		Weighted Avg.	91.60%	90.70%	89.90%
	PSO	Charged off	99.20%	64.10%	77.90%
		Fully Paid	88.60%	99.80%	93.90%
		Weighted Avg.	91.40%	90.40%	89.70%
	Generic	Charged off	99.50%	64.80%	78.50%
		Fully Paid	88.80%	99.90%	94.00%
		Weighted Avg.	91.60%	90.70%	89.90%

Table 4. Results of C4.5 decision tree.

Model	Algorithm	Class	Measures		
			Precision	Recall	F1
C4.5	Info Gain	Charged off	96.60%	95.80%	96.20%
		Fully Paid	98.50%	98.80%	98.60%
		Weighted Avg.	98.00%	98.80%	98.00%
	PSO	Charged off	97.40%	95.80%	96.60%
		Fully Paid	98.50%	99.10%	98.80%
		Weighted Avg.	98.20%	98.20%	98.20%
	Generic	Charged off	96.60%	95.80%	96.20%
		Fully Paid	98.50%	98.80%	98.60%
		Weighted Avg.	98.00%	98.00%	98.00%

For Tree based models (C4.5 and C4.5 pruned) improvements ranged between 4.1%-5.5% were showed in prediction power for defaulted borrowers. Also, using Info Gain and Generic algorithms for attributes selection showed the same impact on improvement results; i.e. from 91.50% to 96.20%, and from 93.30% to 97.60% for C4.5 and C4.5 Pruned. On the other hand, PSO algorithm increased F1 of Charged off class from 91.50% to 96.60%, and from 93.30% to 97.40% for C4.5 and C4.5 Pruned.

Table 5. Results of C4.5 decision tree (pruned).

Model	Algorithm	Class	Measures		
			Precision	Recall	F1
C4.5 Pruned	Info Gain	Charged off	99.10%	96.00%	97.60%
		Fully Paid	98.60%	99.70%	99.20%
		Weighted Avg.	98.70%	98.70%	98.70%
	PSO	Charged off	99.40%	95.60%	97.40%
		Fully Paid	98.40%	99.80%	99.10%
		Weighted Avg.	98.70%	98.70%	98.70%
	Generic	Charged off	99.10%	96.00%	97.60%
		Fully Paid	98.60%	99.70%	99.20%
		Weighted Avg.	98.70%	98.70%	98.70%

Table 6. Results of random forest.

Model	Algorithm	Class	Measures		
			Precision	Recall	F1
Random Forest	Info Gain	Charged off	98.50%	85.90%	91.80%
		Fully Paid	95.20%	99.50%	97.30%
		Weighted Avg.	96.10%	95.90%	95.90%
	PSO	Charged off	98.60%	86.70%	92.30%
		Fully Paid	95.50%	99.60%	97.50%
		Weighted Avg.	96.30%	96.20%	96.10%
	Generic	Charged off	98.30%	87.10%	92.40%
		Fully Paid	95.60%	99.50%	97.50%
		Weighted Avg.	96.30%	96.20%	96.20%

Applying Data preprocessing and features selection techniques prior to using Random Forest for classification had the highest improvement impact on the prediction power of Charged off class. Indeed, F1 measure for Charge off class increased from 49.40% to 91.80%, 92.30% and 92.40% when using Info Gain, PSO and Generic algorithms respectively for attributes selection, with an average increase of 42.77%.

The following matrix contains a calculation of weighted average F1 improvement percentage of each model and feature selection methodology when applying preprocessing techniques as shown in Table 7.

Table 7. Results of weighted average F1 improvement.

Model	Feature Selection Algorithm		
	Info Gain	PSO	Generic
Naïve Bayes	0.20%	0.00%	0.20%
C4.5	1.60%	1.80%	1.60%
C4.5 (Pruned)	1.50%	1.50%	1.50%
Random Forest	10.70%	10.90%	11.00%

In most applications; misclassifying the minority class (false negative) is a lot more expensive than misclassifying the majority class (false positive). In the context of lending, losing money by lending to a risky borrower who is more likely to not fully pay the loan back has higher cost than missing the opportunity of lending to a trust-worthy borrower (less risky). The improvements percentages of evaluation measure F1 of "Charged off (Defaulted)" class between models prior to preprocessing stage (and without using any features selection algorithm), and models after applying preprocessing techniques on the original data set, and using three different features selection algorithms is shown in Table 8 below.

Table 8. Results of "Charged-off" F1 improvement.

Model	Feature Selection Algorithm		
	Info Gain	PSO	Generic
Naïve Bayes	4.40%	3.80%	4.40%
C4.5	4.70%	5.10%	4.70%
C4.5 (Pruned)	4.30%	4.10%	4.30%
Random Forest	42.4%	42.90%	43.00%

5. Conclusions

This paper used Naïve Bayes, Decision tree (unpruned and pruned) and Random Forest classifiers to build loan default prediction models. This paper also applied several data preprocessing techniques, and compared between three features selection algorithms: Information Gain, Genetic Algorithm and Particle Swarm Optimization.

Applying preprocessing techniques proved to be very useful as it significantly enhanced the prediction of the minority class. Improvements obtained were varied across the different classifiers. Using features selection algorithms proved to be positive as well as it enhanced the quality of the models' outcomes. However, the improvements variation amongst the three utilized feature selection methods and across the three used algorithms were not remarkable.

It can be concluded that data preprocessing stage is an important stage when building a classification model, as it has a valuable impact on the quality of the model's outcomes and accuracy. Applying features selection algorithms is very significant as well when having a large dataset. Not only it enhances accuracy, but also it improves the performance of the model. Future work would involve other classifiers and features selection algorithms, as well as using datasets from different banks and in different time frames to enhance the generalizability of our findings. Indeed, future research should examine other important algorithms such as neural networks and show the importance of data processing in enhancing the quality of outcomes.

References

[1] Al-qerem A., Al-Naymat G., and Alhasan M., "Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection," *in Proceedings of International Arab Conference on Information Technology*, Al Ain, pp. 235-240, 2019.

[2] Akrani G., Kaylan City Life (20-Apr-2011), Available: http://kalyan-city.blogspot.com/2011/04/functions-of-banks-important-banking.html, Last Visited, 2019.

[3] Angelini E., Tollo G., and Roli A., "A Neural Network Approach for Credit Risk Evaluation," *The Quarterly Review of Economics and*

Finance, vol. 48, no. 4,pp. 733-755, 2008.

[4] Bentlemsan M., Zemouri E., Bouchaffra D., Yahya-Zoubir B., and Ferroudji K., "Random Forest and Filter Bank Common Spatial Patterns for EEG-Based Motor Imagery Classification," *in Proceedings of International Conference on Intelligent Systems, Modelling and Simulation*, Langkawi, pp. 235- 238, 2014.

[5] Chen Y. Zhang J., and Ng W., "Loan Default Prediction Using Diversified Sensitivity Undersampling," *in Proceedings of International Conference on Machine Learning and Cybernetics*, Chengdu, pp. 1020-1025, 2018.

[6] Chioka (2013, Aug, 30), Available: http://www.chioka.in/class-imbalance-problem/. Last Visited, 2019.

[7] Deng T., "Study of the Prediction of Micro-Loan Default Based on Logit Model," *in Proceedings of International Conference on Economic Management and Model Engineering*, Malacca, pp. 260-264, 2019.

[8] Eulogio R, ORACLE + Data Science (2017, Aug, 12), Available: https://www.datascience.com/resources/notebooks/random-forest-intro, Last Visited, 2019.

[9] Gahlaut A., Tushar K., and Singh P., "Prediction Analysis of Risky Credit Using Data Mining Classification Models," *in Proceedings of 28th International Conference on Computing, Communication and Networking Technologies*, Delhi, pp. 1-7, 2017.

[10] Hassan A. and Abraham A., "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks," *in Proceedings of International Coference on Computing, Electrical and Electronic Engineering*, Khartoum, pp. 719-724, 2013.

[11] Hsu C. and Hung F., "Classification Methods of Credit Rating - A Comparative Analysis on SVM, MDA and RST," *in Proceedings of International Conference on Computational Intelligence and Software Engineering*, Wuhan, pp. 1-4, 2009.

[12] Jin Y. and Zhu Y., "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," *in Proceedings of 5th International Conference on Communication Systems and Network Technologies*, Gwalior, pp. 609-613, 2015.

[13] Kim H. Park C.,Yang H., and Sim K., "*Genetic Algorithm Based Feature Selection Method Development for Pattern Recognition*," *in Proceedings of SICE-ICASE International Joint Conference*, Busan, pp. 1020-1025, 2006.

[14] Mahanipour A. and Nezamabadi-pour H., "Improved PSO-based feature construction algorithm using Feature Selection Methods," *in Proceedings of 2nd Conference on Swarm Intelligence and Evolutionary Computation*, Kerman, pp. 1-5, 2017.

[15] Netti K.and Radhika Y., "A Novel Method for Minimizing Loss of Accuracy In Naive Bayes Classifier," *in Proceedings of IEEE International Conference on Computational Intelligence and Computing Research*, Madurai, pp. 1-4, 2015.

[16] Reddy M. and Kavitha B., "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis," *in Proceedings of International Conference on Signal Acquisition and Processing*, Bangalore, pp. 274-277, 2010.

[17] Shoumo S., Dhruba M., Hossain S., Ghani N., Arif H., and Islam H., "Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking," *in Proceedings of TENCON IEEE Region 10 Conference*, Kochi, pp. 2023-2028, 2019.

[18] Xiaoliang Z., Hongcan Y., Jian W., and Shangzhuo W., "Research and Application ofthe improved Algorithm C4.5 on Decision Tree," *in Proceedings of International Conference on Test and Measurement*, Hong Kong, pp. 184-187, 2009.

[19] Xiang-wei L. and Yian-fang Q., "A Data Preprocessing Algorithm for Classification Model Based On Rough Sets," *in Proceedings of International Conference on Solid State Devices and Materials Science*, pp. 2025-2029, 2012.

[20] Zhang H., Ren Y., and Yang X., "Research on Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree," *in Proceedings of 10th Web Information System and Application Conference*, Yangzhou, pp. 446-449, 2013.

Ahmad Al-Qerem graduated in applied mathematics and MSc in Computer Science at the Jordan University of Science and Technology and Jordan University in 1997 and 2002, respectively. After that, he was appointed as full-time lecturer at the Zarqa University. He was a visiting professor at Princess Sumaya University for Technology (PSUT). He obtained a PhD from Loughborough University, UK. His research interests are in performance and analytical modeling, mobile computing environments, protocol engineering, communication networks, transition to IPv6, machine learning and transaction processing. He has published several papers in various areas of computer science. Currently, he has a full academic post as a full professor at computer science department at Zarqa University-Jordan.

Ghazi Al-Naymat received my Ph.D. degree in May 2009 from the School of Information Technologies at The University of Sydney, Australia. In 2015, I joined the Department of Computer Science, King Hussein School of Computing Sciences at Princess Sumaya University for Technology (PSUT). In addition, I worked as the chair of the computer science department at PSUT from 2017-2019. My research interests include: Data Mining and machine learning, big data, and data science. He has a full academic post as associate professor at Department of Information Technology, Ajman University, Ajman, United Arab Emirates.

Mays Al Hasan obtained her bachelor degree in Computer Engineering from Jordan University of Science and Technology, and she is now studying for a Data Science master's degree at Princess Sumaya University for Technology. She has published couple of papers in different areas of Data Science and Analytics. She is currently working as Technical Product Manager for AI and Analytic at Mawdoo3, and has over 10 years of experience working domestically and internationally in the Analytics and Business Intelligent fields in different industries.

Mutaz Al-Debei is currently working as a Senior Territory Manager for the Public Sector at Oracle. Previously, he was a Senior Territory Manager for Autonomous Data Management & Cloud Technology at Oracle. Also at Oracle, Al-Debei had a previous role as a Principal Cloud Platform Consultant - Big Data & Business Analytics. Before Joining Oracle, Al-Debei was working as the Director of Big Data & Advanced Analytics at INTRASOFT MEA. Moreover and before joining INTRASOFT, Al-Debei was serving as an Associate Professor of Information Systems and Computing at the University of Jordan (UJ), and also as an ICT Chief Consultant at the National Center for Security and Crises Management. He also worked as an IT Manager for Arab Radio & Television (ART) in Jordan Media City and he held other positions in Al-Ahli Bank (Master Card Department) and Royal Scientific Society. Al-Debei earned his PhD from Brunel University London (BUL) in Information Systems and Computing in May, 2010. Furthermore, Al-Debei has received many international and national significant research awards such as Abdul Hameed Shoman Award for Arab Researchers – ICTs, 2015, the prestigious Vice Chancellor's Prize for Doctoral Research from Brunel University London in 2010, the Distinguished Researcher Award from The University of Jordan - three times in 2012, 2014, and 2018. Also, he received best paper awards from UKAIS (2008), and another one from IFIP 8.2 (2010).