A Hybrid Feature Selection and Hybrid Prediction Model for Credit Risk Prediction

MSc Research Project
MSc Data Analytics

Purvi Shetty
Student ID: x20122322

School of Computing
National College of Ireland

Supervisor: Prof Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Purvi Shetty
Student ID:	x20122322
Programme:	MSc Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof Athanasios Staikopoulos
Submission Due Date:	16/12/2021
Project Title:	A Hybrid Feature Selection and Hybrid Prediction Model for Credit Risk Prediction
Word Count:	8580
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Purvi Shetty
Date:	15th December 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Hybrid Feature Selection and Hybrid Prediction Model for Credit Risk Prediction

Purvi Shetty
x20122322

Abstract

Borrowings in the consumer financial market have increased dramatically over the last few years. As a result, the risk of loss due to borrower payment failure has increased. Credit risk mitigation is a major challenge for lending institutions such as banks. Machine learning techniques for credit scoring and default prediction are assisting financial institutions in reducing credit risk. In the consumer financial market, an accurate prediction is critical. Even minor improvements to the prediction model can help banks evade colossal losses. Data mining methodologies such as feature selection and single classifier have been applied and studied in the credit risk domain. But the effects of hybrid models are not much explored. In this study, we propose a hybrid classification model containing XGBoost, CatBoost, and LightGBM combined using a stacked generalization technique. And a hybrid feature selection model is created using Feed Forward, Weight of Evidence(WOE), Anova, Extra trees, Random forest, and L1 feature selection. The results are combined using the voting ensemble approach. Oversampling technique SMOTE is employed to balance the datasets. Lastly, the approach is generalized using three datasets from the credit risk domain. The results show that the hybrid feature selection technique outperforms traditional methods for all three datasets and can be generalized for the Credit risk domain. The stacked model outperformed state-of-the-art for large and medium datasets with an AUC value of 96% and 87%, respectively. But for small datasets, we found single classifiers were beneficial. We were able to identify major indicators in the credit risk domain. This approach will help banks and other lending institutes to improve the performance of the credit risk models and help backup business decisions.

Keywords- Credit Default Prediction, CatBoost, Deep learning, Ensemble approach, Feature Selection, LightGBM, Machine Learning, stacked generalization, voting mechanism, XGBoost

1 Introduction

Over the past decade, there has been a colossal growth in the consumer financial market. The number the loans in the market has increased. Loans range from personal loans to professional loans, housing loans, small business loans, car loans, education loans, etc. Along with the loan, there is an increase in the number of people using credit cards. With exciting offers, cashback, and discounts on using credit cards, there is a rise in demand for credit cards. With such evolution in the financial market, risk management is a critical issue faced by financial institutes in the field of credit. Organizations such as banks that seek profits by providing services such as loans and credit cards are aware of the credit risks are willing participants. For example, A personal loan of up to 7.5 lakh Rupees does not require collateral. Thus, risk-taking and risk management are the major driving forces for banks and institutes for profitability Khemakhem and Boujelbene (2015).

Over the years, there has been a rise in the number of credit defaults, especially after the pandemic. A loan is classified as a Non-Performing Loan(NPL) when the repayment is not done within 90 days of the due date. European Central bank reported NPL worth €550 billion. This amount is likely to rise to €1.4 trillion by the end of 2022, due to the economic crises caused by the global pandemic (Bank; 2020). Allied Irish Banks (AIB) has reported an increase in NPL rate from 5.4% in 2019 to 7.3% in 2020. Meanwhile, the Bank of Ireland noted a rise to 5.7% from 4.4% (Times; 2021). Defaults can be caused to many reasons social, economic, unemployment, inflation, low GDP, unexcepted crisis, etc. Predicting credit default is vital because defaults cause the banks to lose money which renders the ability of the banks to give out new credits and ultimately reduces the bank's profitability.

Using data mining, we can compute the credit score for loans, predict credit defaults. Data mining provides us with getting tools for credit risk management. Credit risk is a critical challenge faced by financial institutes as it determines profitability. Hence, accurate prediction is crucial in the credit risk domain. Even a slight increase in the performance of the prediction model aids in evading massive monetary loss. A reliable credit risk prediction system can reduce the cost incurred by manual credit scoring or prediction system and can help back-up decision-making for the banks. This is encouraging banks to invest more in credit risk prediction Dhaiya and Singh (2016).

Over the past four decades, credit risk prediction has evolved from manual computation and mathematical statistics to data mining technologies. In the 1990s, Logistics Regression (LR) was the most widely used classification method for credit scoring (Li; 2019). Apart from LR, support vector machines, the nearest neighbor has been widely used. The development of the decision tree led to its prominent use in the field of credit risk. With the birth of neural networks, it provided a significant contribution in the credit default domain as well (Li; 2019). There have been very few experiments in the credit risk domain using a hybrid model for feature selection and prediction models.

In this research, we will develop a hybrid feature selection method combining six different feature selection techniques. It will enable us to combine the advantages of the filter, wrapper, and tree-based feature selection methods using the voting mechanism to obtain an optimal feature subset. A hybrid prediction model combining three prediction models using a stacked generalization. This hybrid model will help us tackle the problem of bias associated with a single method which occurs due to the relational nature of the dataset Singhi and Liu (2006). This leads us to the research question.

To what extent can a hybrid approach for feature selection and prediction model using ensemble technique and SMOTE data balancing technique be effective in credit risk prediction?

The following objectives are identified to address the research question and carry out the implementation.

- Study the previous and recent works in the credit risk domain to gather a better understanding of the problem and investigate best practices and state-of-the-art methods.

- Create a feature selection model combining the wrapper and filter methods to create a hybrid model.

- Balance the datasets using SMOTE.
- Build a hybrid classifier model using stacked ensemble method.
- Evaluating the model using appropriate performance measures.
- Generalizing the approach by using three datasets with different sizes.

The core contribution of this projection is to develop a novel hybrid approach to tackle bias associated with single methods and improve the performance of the credit risk prediction models. The datasets used in this approach are of varying sizes. The small dataset has only 1000 records, and hence the hybrid stacked model may not show high results.

The rest of the research paper contains related work 2 where a detailed analysis of the previous and current work in the credit risk domain is provided. Next is the methodology 3, design specification 4, and Implementation 5 sections which will cover all the aspects related to methodology and how it is implemented to allow reproducibility. The later part of the report consists of evaluation and experiments conducted and discussions 6 followed by conclusion 7 giving the brief overview of the results and the possible area of expansion.

2 Related Work

The related work consists of three segments. The first portion concentrates on the use of the machine and deep learning techniques used in credit risk prediction. The second component sheds light on the feature engineering methods utilized in credit risk prediction and similar domains. The last section contains the works on ensemble methods used in credit risk and similar domains.

2.1 Machine and Deep learning techniques in Credit Risk Prediction

Machine Learning is a notable development in the field of data science and artificial intelligence. Machine and deep learning are used to build recommender systems, detect patterns and trends and back up business decisions. In recent years, machine learning is enabling banks and other financial institutes to generate more accurate reports for credit risk. Thus, helping banks from losing money and take appropriate measures in time. Logistic Regression(LR) is a well-performing and the most widely used modeling approach for credit risk prediction by financial institutes such as banks. To find and test alternatives to LR, a comparative study on using new machine learning modeling techniques with different predictive power was carried out by Fitzpatrick and Mues (2016) on mortgage default. A large real-world mortgage loan dataset was used for this study. The dataset was collected by the central bank of Ireland, consisting of 55 percent of Ireland's mortgage market data. The techniques used in the empirical study were Generalized Additive Models(GAMs), Boosted Regression Trees(BRT), Penalised Logistic Regression, and Random Forests (RF). The reason for selecting these models is their nonlinear effect on response variables, and they scale relatively well with any size data. The reasons for technique selection are well articulated in the research paper. The performance of the models was accessed using the h-measure and Area Under the curve(AUC). The models are tuned using the grid search method. The results show that boosting and tree-based approaches perform relatively well. The boosting techniques can be a valuable addition to credit risk prediction.

In the past decade, banks have seen an increase in credit card circulation. With the ever-increasing credit card demand, the credit default risk has also increased. Banks face a huge risk on the returns. As it is said higher the risk higher is the return. Financial institutes are now using machine learning to assess this risk. Yang and Zhang (2018) carried out a comparative study of various machine learning approaches on credit card data. The machine algorithms selected for the study are Neutral network, SVM, xgboost, logistic regression, and LightGBM. The reason for using these algorithms is not clearly documented by the author. Out of all the selected algorithms, LightGBM and xgboost performed well. Thus, proving to have a good application value in the banking and finance domain. The performance metrics used in the study are AUC and F1-score. The issue with this study is that the model is directly applied without any tuning to improve the model performance. Also, the methodology section is not described to allow reproducibility.

The has been a rapid growth in the use of credit cards in the energy industry. But, there are very few studies aimed to focus on measuring the credit risk in the energy industry. A recent study was carried by Tang et al. (2019) to measure the credit risk in China's energy industry. The data used in the study consists of monthly data of customers using the credit card of Postal Saving Bank over the period from Apr 2014 to JUN 2017. The author has applied the Random Forest(RF) algorithm to scientifically measure and create an effective risk assessment model. This study showed that monthly credit card expense and overdraft ratio has a remarkable impact on credit risk. A sample data of 25474 customers was obtained to create the final dataset. The data was randomly down-sampled to select 883 non-default samples. Non-default samples were later combined with 317 default samples to create a final dataset of 1200 rows. In the study, the initial 23 features were manually selected by the author and later verified using the random forest algorithm. To achieve optimum results, parameters such as ntree and mtry are tuned. Accuracy was used to assess the performance of the model. However, the reliability of any model can not be assessed using accuracy alone. Other performance parameters need to be checked to test the reliability of the model. Also, data imbalance could be addressed using different techniques.

Real-world data for the banking and finance domain especially, the credit risk data is highly unbalanced. Very few studies conducted employ techniques to handle imbalanced data. To address the issues with imbalanced data in the credit risk domain Alam et al. (2020) has explored various resampling techniques. A comparative study was carried out to find the best-suited resampling technique for credit risk. In this study, the undersampling methods explored are cluster centroid, Near Miss, and Random undersampling. The Oversampling techniques explored are Synthetic Minority Oversampling (SMOTE), random oversampling, borderline SMOTE, and k-means SMOTE. Out of all the resampling techniques used, SMOTE outperformed all other methods for credit card data. After resampling performance of the model increased tremendously. The accuracy of the Taiwan credit card data increased from 66.9% to 89%. The classifier used for this experiment is the Gradient boosting algorithm. The author has used other evaluation criteria such as AUC, precision, and recall. Thus, enabling the model to be used for commercial use by banks and other institutes.

In financial data, it is common to have features or variables that evolve with time. These features are called dynamic features. However, in the classic modeling methods, these time dependencies are generally ignored. Deep learning has shown vast capabilities of extracting high-level dynamic features from vast data. In the study Hsu et al. (2019), a Recurrent Neural Network(RNN) is used for feature extraction along with GRU on the credit card default dataset. This provides leverage to time dependencies present in the dynamic features. The input dataset was first preprocessed using the feature RNN feature extractor. It contained a total of 5 static and 18 dynamic features. The extracted features were later trained using the enhanced RNN-RF model. The model provided better performance which is evaluated using Area Under the Curve(AUC) and index lift. The data imbalance was considered, and oversampling technique SMOTE was used to balance the data. The downside to this experiment was that neither the neural network nor the random forest model was tuned. Tuning the model would have enhanced the performance of the RNN-RF model.

2.2 Feature Selection Techniques in Credit Risk Prediction

The raw data in the credit risk domain are large, containing a lot of measurable properties. These proprieties are referred to as features. Choosing appropriate features is crucial to enhancing the productivity of any machine learning model. For selecting important features, detailed analysis by domain experts is required. But such manual analysis is often costly and slow. Hence computational methods are employed for feature selection. Hajek and Michalak (2013) provided a comparative study between filter and wrapper feature selection techniques on corporate credit rating. Two datasets were used in the study. One contains 852 US companies and the second with 244 European companies. Different classifiers were used to check the performance after feature selection. The experimental result shows that the wrapper method performed as compared to the filter method on both datasets. But the performance on the US dataset was not quite good since the attributes have weak correlations. The performance is measured using the TYPE I error, misclassification cost, and accuracy, making it reliable. Although the performance in terms of accuracy improved, the computational time required is very high. Another problem with the paper is that it has no methodology section. Hence reproducibility is not possible.

Classification models often run a risk of over-fitting due to the large dimensionality of the data. To address this issue of high dimensional Yu et al. (2021) proposed a trait-driven approach for credit classification. This trait-driven approach consists of categorization of data, feature selection, and classifier selection. If the feature dimension is greater than the sample size then, Principle Component Analysis(PCA) is used for dimensionality reduction. Whereas the feature is less than the sample size, no feature extraction s required. PCA is used since it prevents information loss and overfitting. The performance of the model is evaluated using accuracy, confusion matrix, and AUC. Various combinations for feature selection methods can be tested to increase the performance of the model.

Singh and Sivasankar (2019) performed a study aimed to compare the performance of various predictive models between the traditional single classifiers and the new ensemble techniques to check the potentiality of the clients to whom credits are availed. Filter feature selection method, information gain was used to obtain features with high entropy. The prediction classifier used on the Taiwan credit card dataset is Naive

Bayesian(NB), support vector machines(SVM), KNN, decision tree, random forest, boosting, and bagging algorithms. Among all the models boosting techniques combined with the information-gain selection provided the best results in the study. But the reliability of the model can not be assured because only accuracy was used to measure performance. Also, the reproducibility of the method is not possible due to the lack of a proper methodology section. Filter methods are fast, but they do not necessarily provide an optimal feature subset. So only testing a single filter method for feature selection may not help us obtain the optimal performance for our classification problem. For an accurate prediction, the selection of appropriate attributes is crucial. Trivedi (2020) performed a comparative between Gain-Ratio, Chi-square, and information gain filter feature selection methods. Various ML algorithms such as NB, random forest, decision tree, SVM were applied. This experiment aimed to find the best combination of filter and classifier techniques. Random forest plus chi-square gave the best combination for credit scoring. But the experiment was conducted on the only German dataset. Hence limiting to one dataset and failing to generalize the approach for the domain. Another downfall of the study was the computational time required was quite large for a dataset with only 1000 records.

Classical machine learning algorithms do not have the ability to asset allocation. In peer-to-peer lending looks into the needs of individual investors and hence is different from the traditional financial market. For such a new and emerging market, Ha et al. (2019) proposed a novel approach to analyze the data and the risks associated with individual loans. In the method, feature selection was conducted using the Restricted Boltzmann Machines (RBMs). The top 10 attributes are selected using the RMSE value. Classification is carried out as a comparative study between rule-based, non-linear, and linear classifiers. Dimensionality is the major issue with the credit scoring domain, many feature selection algorithm requires large computational time. To reduce the time required, Jadhav et al. (2018) proposed a novel approach combining the information gain and genetic algorithm wrapper. The performance is tested against SVM and KNN. This method worked well with SVM, but the performance of KNN decreased because information gain affects the structure of the data, and KNN is structure sensitive. selection of the classifier needs to be done after careful assessment of the datasets. Kozodoi et al. (2019) proposed a profit-centric attribute selection method that is based on the NSGA-II genetic algorithm. This approach increases comprehensibility as it focuses on the profit-centric attributes and not on the statistical measure, also addressing the structure change issue.

2.3 Hybrid Techniques and Ensemble Learning Techniques

In recent years, there has been an exponential increase in high-dimensional data. This increase has led to the increasing use of feature engineering. The traditional approach for feature selection shows poor performance with such high-dimensional datasets. Although filter methods are fast, they do not provide an optimal subset. Whereas the wrapper method does provide better subsets, but their computational time is high. Many researchers are now exploring hybrid approaches of combining filter and wrapper methods to gain the benefits of both methods. Amirreza and Hossein (2017) proposed a novel approach of combining fast correlation-based based approach with two wrapper approach improved binary gravitational search algorithm(IBGSA) and ant-colony optimization(ABACOH). The feature selection subsets are combined using the aggregation ensemble method. The author has well-articulated the approach using diagrams. For classification, the KNN technique is used. The approach is generalized using five high-dimensional datasets from the medical domain. The only downfall is that the performance of the approach is validated using accuracy.

A similar approach to the above paper was implemented by Venkatesh and Anuradha (2019). In this experiment, a new hybrid technique is implemented using information gain and recursive feature elimination. First, information gain is used to reduce the dimension of the data, and the optimal feature is selected using the recursive feature elimination technique. Using the filter method first helps reduce the size of the data, which helps reduce the computational time of the wrapper method. The classifier used is random forest. The approach is generalized by using three benchmark datasets, namely Libras movement, clean, and Ionosphere dataset from UCI Repository. This methodology is well-written and self-explanatory. The above two methods have shown promising results but have not been used on datasets from the credit risk domain.

Studies have found that using multiple classifiers can help reduce the risk of overfitting better than single classifiers. They also give better performance and help in generalizing. Li et al. (2021) proposed the use of the super learner ensemble technique for default prediction. In the proposed approach, the author is using six base classifiers on three different datasets. Then the prediction probability is used to compute the weights of every classifier. The value for the classifier's total loss is calculated, and a hybrid model is created using the minimum weighted loss. The robustness of the model is ensured using by using the AUC, accuracy, TYPE I, and II error to evaluate the performance of the approach.

Some approaches for feature selection that work on supervised classification may not work for unsupervised classification problems or vice versa. Das et al. (2017) came up with a novel feature selection approach using a feature association map that works well with both unsupervised and supervised learning techniques. FAM uses a combination of Pearson's correlation and mutual information along with vertex cover and independent set to generate an optimal subset of features. The FAM method implemented is a hybrid method wherein the first filter approach is used to create multiple subsets of features. Later from these subsets, an optimal subset is selected using a wrapper method. Since the wrapper method is applied on reduced feature subsets, the computational time required is greatly reduced. The method is applied across 18 datasets, and the results are compared with other feature selection approaches. Although the theory for FAM is well explained by the author, the application of the FAM in the hybrid algorithm is a little unclear. The author has done an exceptional job in generalizing the approach for both supervised and unsupervised techniques.

In recent years, many studies are concentrated on hybrid models for classification and regression problems to develop a non-biased and reliable model. Many studies suggest that the stacked generalization technique for creating a hybrid model gives superior results when compared to traditional methods. To solve the regression problem between the load demand and electricity generation, which causes overloads, Massaoudi et al. (2021) proposed the approach of stacked generalization. This approach combines the output of meta learners LGBM and XGB. The outputs of these learners are combined using MLP in the final stage of the approach. This approach explored a novel way for Short-Term Load Forecasting. The MLP layers were further optimized using five different techniques like random search, particle swarm optimization, evolution method, simulated annealing, Bayesian optimization. Random search provided the best results for the load forecasting problem. The only downfall of the approach was that after 48 hours, the performance approach decreased. Future work comprises using different deep learning techniques in the meta-learning stage to maintain the performance even after 48-hours.

2.4 Summary and GAP in Research

From the existing work, we can summarize that feature selection play a vital role in enhancing the performance of prediction and classification model. In short, we can say feature selection has a direct influence on predictive performance. Understanding and identifying important features are crucial in the machine learning domain, especially in the banking and credit risk domain, since the datasets in this domain are quite large. Many recent studies have employed a singular feature selection method in the credit risk domain. But a hybrid feature selection method using a voting mechanism does not appear to be used. A single method for classification and prediction can be biased. To solve the issue of this issue of bias, in this research, we will use the stack generalization technique to create a hybrid prediction model. From the existing work, it does not appear that stack generalization is used for hybrid model creation in the credit risk domain. The researches conducted on hybrid models in other domains show improved performance. In the credit risk domain, even a slight improvement is of great value.

3 Methodology

For this study, we used the CRISP-DM methodology. This section outlines the steps and processes taken for the research project. This section is divided into the following subsections data collection, preprocessing and transformation, feature selection, ensemble modeling, hyper-parameter tuning, and Evaluation techniques used in the project. Figure 2 illustrates the flow chart for the proposed research.

3.1 Data Collection

In order to enable the generalization of the proposed approach, we have used three datasets of varying sizes from the credit risk domain. The first dataset used is the Small Business Loan Approval(SBA) dataset [1]. We check if the businesses seeking a loan will default the repayment of the Loan before giving our approval. It is a large dataset with 899164 records and 27 variables and was obtained from Kaggle. This dataset is relatively new and has not been explored yet. The second dataset selected is the credit card default prediction dataset(Taiwan) [2]. It comprises 30000 rows and 25 attributes and was obtained from the UCI repository. The third dataset is a small dataset for German credit scoring [3] with only 1000 records and 20 columns and was taken from the UCI repository. The datasets contain information on every person who obtains a bank loan.

[1]https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied?select=SBAnational.csv

[2]https://archive.ics.uci.edu/ml/datasets/default%2bof%2bcredit%2bcard%2bclients

[3]https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

3.2 Data Preprocessing

3.2.1 Dataset 1: SBA Loan (Size: large)

We drop unique ID and borrower's organization name as these are noisy and useless features. Next, we drop attributes that directly tell us that the loan has been charged off, which are useful after the default occurs, and we also remove time-dependent variables. Hence, we drop ChgOffPrinGr, ChgOffDate, ApprovalFY , ApprovalDate, DisbursementDate. There are four columns BalanceGross, SBA_Appv, DisbursmentGross, GrAppv with a currency symbol, so we perform a transformation to remove the $ symbol from the data. The next step involves identifying duplicate and null values. The data does not contain any duplicate data and eight columns, including the dependent variable MIS_Status. After careful consideration and looking at the data, we have used the mode technique to handle null values. This is because we can observe that the values in the column MIS_Status are positively skewed towards value 0. Mode is the best option we the data is skewed. With the columns having object data type, we have used one-hot encoding for a smaller number of unique values and hashing techniques for an attribute with a large number of unique values. Finally, we create a new column defaulted using the MIS_Status column. If the status is CHGOFF, the defaulted value is 1 and 0 otherwise. We are going to use the defaulted column as the predictor variable. Lastly, we perform scaling so that the values are between 1 and -1.

3.2.2 Dataset 2: Taiwan credit default (Size: medium)

The dataset did not have proper column names. The first step was to assign proper column names using the description of the dataset. This dataset does not contain any text data, only numerical values. Next, we check for any missing or null values. Drop ID column as it is not a useful feature. Education and marriage are nominal attributes. We have used hot encoding using dummies for both these attributes. To do so, we take the unique values, PAY_0 to PAY_6 are ordinal columns, and BILL_AMT1 to PAY_AMT6 are numerical features and hence do not require any preprocessing. Lastly, we perform scaling so that the values are between 1 and -1.

3.2.3 Dataset 3: German Credit Card (Size: small)

The dataset did not have any column names, so the first step of data pre-processing is to add columns according to the description given. Also, all the data in the columns are in the form of codes like A11, A12, and so on. The data is converted into numeric values by looking at the metadata information present in the dataset description. The next step is to read the data and check for null and duplicate values. The German credit card dataset does not contain any duplicate or null values. Lastly, we perform scaling so that the values are between 1 and -1.

3.3 Hybrid Feature Selection

Irrelevant and unwanted features can increase the computational complexity in credit risk prediction due to the large dimensions of the data. It increases the cost and effort required for the modeling process. Feature selection aids in the identification of optimal features, the reduction of complexity, and the improvement of model performance. Feature selection is of three types filter, wrapper, and embedded method. The filter method filters out redundant features. It is computationally fast, but it does not always provide us with the best subset. The wrapper method yields an optimal subset but is computationally slow and costly. The embedded method selects features during the modeling creation process. Sometimes with a single feature selection method, attributes are selected due to the relational nature of the data and may not be an important feature and can be misleading, giving us poor predictive performance. It is called feature selection bias Singhi and Liu (2006). In the proposed approach to solve the bias problem, we will use six feature selection methods Feed Forward, Weight of Evidence(WOE), Anova, Extra trees, Random forest, and L1 feature selection. We combine the results using a voting mechanism(any vote, hard voting, unanimous vote, soft vote, and minority voting) 4.1. The voting mechanism uses the frequency count of features obtained from different feature selection methods providing us with an unbiased feature subsetZhu et al. (2019).

3.4 Oversampling using SMOTE

The datasets in the credit risk domain are generally imbalanced. The datasets selected for the project are highly imbalanced. To deal with the bias resulting due to imbalanced data, we have used the Synthetic Minority Oversampling Technique(SMOTE). In this technique, the examples for the minority class are duplicated. Thus, the is no addition of information or loss of information as in the case of the under-sampling technique.

3.5 Stacked Generalization Classifier

For credit risk prediction, we are going to use the stacked generalization technique to combine multiple classifier models. From the literature review, stacked generalization has the capability of combining multiple models predictions and giving better performance than a single model. Stacked generalization is of two stages. The first stage consists of the base learner's classifiers. In this research, we are going to use CatBoost, LightGBM, XGBoost 4.2. The second stage consists of combining the prediction result of the meta learners and providing a final prediction using the meta learner.

3.6 Hyper-parameter tuning

Optimization is required to get the best model or get the most out of the machine or deep learning model. Hyper-parameter is used to optimize the model based on the dataset. To select the best set of parameters for the credit risk prediction dataset in the research, we have used the Grid Search technique. In grid search, a grid of hyper-parameter is defined, and every position in the grid is evaluated to find an optimal set of parameters. This is called exhaustive search, and it gives the best way to tune the parameters.

3.7 Evaluation

For credit risk prediction sensitivity and specificity are equally important because incorrect prediction can cause severe money losses. For example, a default predicted as non-default will lead to giving the loan to a person who will not be able to pay it back. And non-default predicted as a default means losing a potential client. Hence, we have used F1-score with accuracy for performance evaluation. Area Under the curve(AUC) measures the performance of the model in terms of its ability to differentiate between the classes and is used to check to create a robust model since the datasets are unbalanced.

4 Design Specification

We have used the three-tier architecture. The first layer is the data layer. In this layer, we gather and understand the data. It is vital to understand the data cause it aids in selecting appropriate models for prediction and classification. The data layer in this research project consists of data gathering, data preprocessing and transformation, feature selection, and oversampling. The second layer is the application layer, also known as the business layer. In this layer, we create the prediction models. We are combing multiple models using the stacked generalization technique. The final layer is the presentation layer, where the output for the prediction models can be viewed and analyzed by the end-users.

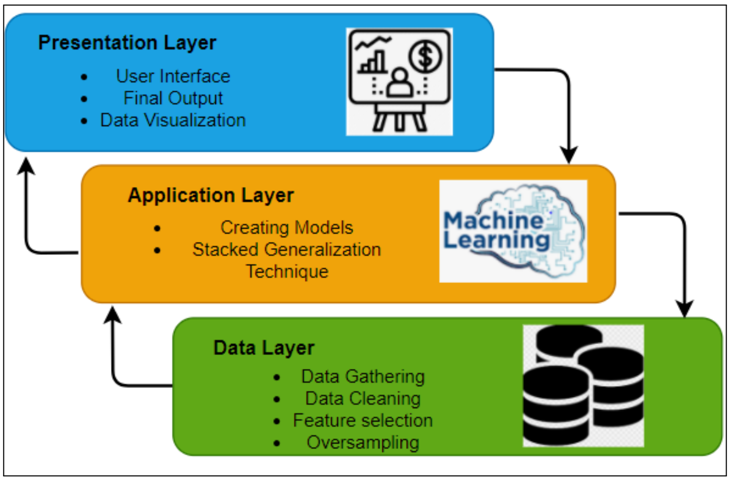

Figure 1: Three-tier Architecture Design

4.1 Hybrid Feature selection using voting mechanism

For the feature selection step, we are using the voting mechanism technique. The methods used for feature selection are

- Forward Selection: Feed Forward is a wrapper method that uses the greedy algorithm for feature selection. In the feed-forward method, features are added into an empty subset to improve the performance. It keeps adding features sequentially until the performance increases. It works well with most models and is therefore used in the implementation.

- Weight of Evidence(WOE): WOE analysis the attribute relevance by measuring the predictive power of a single feature with respect to the target attribute independently. A higher value of WOE tells us how confident we are that the feature will help us predict the probability of the event according. WOE is widely used in credit scoring (Bhalla; 2015) as it provides high interpretability. Thus, we are employing WOE for our implementation.

- Anova: Anova is a statistical method that uses variance as a measure for feature selection. The univariate filter method only shows characteristics of a single variable at a time, whereas ANOVA shows the relationship between two variables.

- Random forest: Random forest is a combination of filter and wrapper methods for feature selection. Each decision tree is made of random extraction of features making them less prone to overfitting.
- Extra trees: The extra tree aggregates de-correlated decision trees to form a forest. It computes a mathematical criteria Gini index to select the best n features from the dataset. It is similar to the random forest but differs on the decision trees are formed.
- L1 feature selection: LASSO or L1 uses the shrinkage method for feature selection. It reduces several coefficients to zero. This leaves features that are truly important discarding other features.

The voting mechanism uses the frequency count of the feature's importance provided by the aforementioned methods and generates subsets based on the number of times a feature is selected. Five different types of voting mechanisms are mentioned in the below table 1.

Table 1: Types of Voting Mechanisms

Unanimous	**Majority**	**Hard Voting**	**Soft Voting**	**Any Vote**
Selected by all methods	Selected by at least 5 methods	Selected by at least 4 methods	Selected by at least 3 methods	Selected by at least 1 method

4.2 Hybrid Prediction Model

To create a hybrid model for credit prediction, we are using the stacked generalization technique. This approach consists of two layers. First is the base learner layer. It consists of all the models that we want to combine. In this project, we are using CatBoost, XGBoost, and LightGBM. The output from these single classifiers is combined and stored to form a new dataset along with the predictor variable. This new dataset is used by a meta-leaner to give the final prediction. The meta learner used is LightGBM.

- eXtreme Gradient Boosting(XGBoost): It is an implementation of gradient boosting. It uses the ensemble method of boosting, where new models are sequentially added into the existing model to correct the error caused by previous models until there is no scope for improvement. The implementation of XGBoost is carried out to achieve efficient computation time and memory resources. Other reasons for using XGBoost are that it can automatically handle missing data, can apply multi-threaded parallelism to accelerate execution time, and enables continuous training. Thus, providing us with high accuracy with minimum errors Son et al. (2019).
- LightGBM: LightGBM is a boosting technique that uses a novel technique of gradient-based one-side sampling to compute the best split and grows tree vertically leaf-wise, whereas XGBoost grows horizontally level-wise. We have implemented LightGBM because it is a relatively new algorithm developed in Jan 2017 known for providing efficient results by handling large datasets while using less memory and supporting GPU learning. It has more than 100 parameters which are difficult and needs to be carefully tuned to get efficient performance Ke et al. (2017).

- CatBoost: CatBoost is a new boosting algorithm that was released in July 2017. CatBoost used an oblivious decision tree to create a balanced decision tree. It divides the dataset by random permutations and applies boosting on them. Thus, making the algorithm easier to fit and reducing the execution time. What makes CatBoost different from other algorithms is that much effort is not required for handling missing data and text attributes since it is capable of handling missing data and text columns. It works exponentially well with a large amount of data, making it suitable for the credit risk domain. Hancock and Khoshgoftaar (2020).

5 Implementation

This section outlines the tools and technologies used for the project implementation and the final stage outcome description. The below table Table 2 provides the details on the tools and technologies used with purpose.

Table 2: Tools and Technologies Used for Project Implementation

Tools and Technologies used	**Reason**
Python	The programming language used for coding and implementation
Jupyter Notebook	Interactive computing Platform
Google Colaboratory	JupterWorking environment
Google Drive	Data are uploaded and stored in Google Drive
Microsoft Excel	Datasets are stored in CSV format.

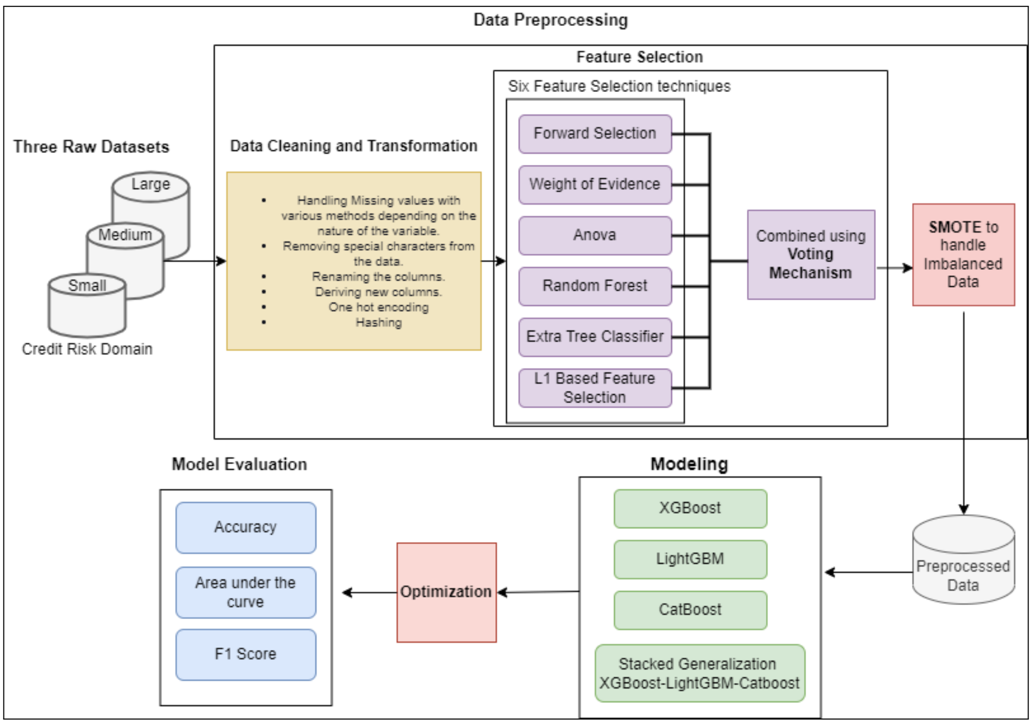

Figure 2: Implementation Flow Chart

Figure 2 illustrates the implementation flow for the research. All three datasets are stored in CSV format and uploaded on google drive. The datasets are mounted in google colab using the drive library. The datasets are then cleaned and transformed. The following steps were taken to clean and transform the datasets: handling missing data, removing special characters from data, renaming columns names, deriving new columns, one-hot encoding, and hashing. After performing these steps for dataset 1, the number of columns is increased to 44 columns, and the row count is the same. For the mid-size dataset 2, the number of variables count is 32, and the row count is 30000. Dataset 3 has no changes in terms of column and row count.

After data cleaning, we implement the proposed hybrid feature selection approach. In this stage, we have applied data six feature selection Feed Forward, Weight of Evidence(WOE), Anova, Extra trees, Random forest, and L1 feature selection on all three datasets. The outputs of these feature selection techniques are converted into binary format and stored in a single data frame. The final score column is created by adding the occurrence of the features in each feature selection technique. The below table Table 3 provides the feature subset i.e number of features selected for each voting mechanism for each of the voting mechanism mentioned in the table Table 2 in design specification section. For all three datasets, Soft-Voting provided the best feature subset.

Table 3: Output for Voting Mechanism

	Unanimous	Majority-Voting	Hard-Voting	Soft-Voting	Any Vote
Dataset 1	9	16	19	23	42
Dataset 2	6	9	11	21	32
Dataset 3	5	10	13	17	18

All three datasets are highly imbalanced. Dataset 1 has 82% defaulted data and 18% non-defaulted value. Similarly, dataset 2 has 78% default payment values and 22% non-defaulted value. Dataset 3 has a 70% good rating and 30% bad credit rating. To create balanced data, we have used SMOTE oversampling technique.

5.1 Prediction Model Implementation

We have implemented XGBoost, CatBoost, and LightGBM individually and later combined them using a stacked generalization approach to get the final prediction. For dataset 1, to optimize the XGBoost algorithm, we have set colsample_bytree to 0.3, gamma to 0.2, learning_rate to 0.1, max_depth to 9. For CatBoost algorithm, we have set depth, iterations, and learning_rate to 10, 100, 0.4, respectively. To maximize LightGBM performance, we have used learning_rate as 0.1, n_estimators as 20, num_leaves as 500, and colsample_bytree as 0.75. Similarly, we have adjusted the hyper-parameters of models Xgboost, Catboost, and LightGBM for dataset 2 and dataset 3. To find the best parameters, we have used a random search for XGBoost, whereas for CatBoost and LightGBM, we have used a grid search algorithm. We store the output of the three models in a data frame and combine all three data frames into one dataset. We make the final prediction on top of this dataset using LGBMClassifier. To combine the outputs and make the final prediction, we use the StackingClassifier function from the sklearn.ensemble package.

6 Evaluation

In this research, we have conducted three experiments for the three datasets used in the project. To assess the performance of the models, we have used Accuracy, F1-score, and AUC. The AUC helps us differentiate between the classes when the datasets are imbalanced. Hence, AUC is important for creating a reliable model. We have applied XGBoost, CatBoost, and LightGBM and ensembled a stacked generalization model by combining the three models. The results with unbalanced and balanced data in specified in this section.

6.1 Experiment 1: Evaluating the Voting Mechanism

Table 4: Results Of all the voting subsets on all the datasets using AUC value

		Unanimous	**Majority**	**Hard**	**Soft**	**Any_Vote**
Dataset 1	**XGBoost**	91	94	94	94	92
	LightGBM	94	94	94	94	94
	CatBoost	92	92	92	95	94
	Stacked	95	95	95	96	95
Dataset 2	**XGBoost**	72	84	82	84	84
	LightGBM	71	84	82	83	84
	CatBoost	70	85	84	84	84
	Stacked	71	87	85	87	86
Dataset 3	**XGBoost**	76	80	78	86	86
	LightGBM	77	79	80	87	84
	CatBoost	73	76	76	82	82
	Stacked	80	80	76	86	84

Table 4 shows the AUC value obtained for using the 5 voting techniques mentioned in table 2. Values are computed for all four models. From the table, we can observe that the maximum value obtained for all the stacked model using the Soft-voting mechanism is the highest. Also, the value for soft-voting for all the base classifiers is more compared to other methods. The Soft-Voting and Any_vote have close values in many cases, but when we look at the number of features selected 3. Soft-voting is giving high results with fewer features when compared to Any_vote. Dataset 3 shows the difference in the values clearly.

6.2 Experiment 2: Dataset 1 - SBA Loan (Size: large) using Soft-voting

For this dataset, the feature subset obtained from Soft-voting gives optimal results compared to the other voting techniques in table 3. The table 5 provides the output for base classifiers and stacked classifier for feature subset for Soft-Voting.

We have applied the base classifiers first without balancing the data, and we can see that the accuracy and F1- score is around 90 to 94%, which is good. But when we look at the AUC values, it is 73% for XGBoost, which is low compared to accuracy. It means that results are inclined to one class due to class imbalance, also known as overfitting. The boosting algorithm gave a higher value of almost 86, and the out stacked generalization

Table 5: Results for Dataset 1 SBA Loan (Size: large) using Soft-Voting

		Unbalanced			SMOTE		
		Accuracy	**F1**	**AUC**	**Accuracy**	**F1**	**AUC**
Dataset 1	**XGboost**	0.903	0.889	0.734	0.944	0.945	0.94
	CatBoost	0.933	0.931	0.858	0.941	0.941	0.94
	LightGBM	0.949	0.948	0.80	0.951	0.950	0.95
	Stacked Model	0.942	0.941	0891	**0.959**	**0.95**	**0.96**

model was able to provide us with an AUC of about 89 percent. After applying SMOTE for balancing, the accuracy increased by 4%, and the AUC value increased by almost 20%. A similar rise can be seen with the result for Catboost and LightGBM. The percentage increase in the AUC performance is 10% and 15%, respectively. Lastly, for the stack model, we obtained the highest performance with an AUC of 0.96%. The variables that are crucial for loan approval are Term, SBA_Appv, Zip, Bank_hash, UrbanRural, RetainedJob, GrAppv, BankState_hash, RevLineCr_Y. From these variables, we gather that the location, amount of the loan, if the business is set up in an urban or rural region or if there are any revolving credits, number of jobs retained by the business are important factors that help us determine if the loan of the small businesses should be approved or not.

6.3 Experiment 3: Dataset 2 - Taiwan credit default (Size: medium) using Soft-voting

Table 6: Results for Dataset 2 (Taiwan credit default (Size: Medium) using Soft-voting

		Unbalanced			SMOTE		
		Accuracy	**F1**	**AUC**	**Accuracy**	**F1**	**AUC**
Dataset 2	**XGBoost**	0.814	0.79	0.64	0.8414	0.8408	0.84
	LightGBM	0.821	0.80	0.66	0.8313	0.8308	0.83
	CatBoost	0.823	0.80	0.664	0.8454	0.8451	0.84
	Stacked Model	0.824	0.80	0.65	**0.8657**	**0.8652**	**0.87**

Out of the five different subsets obtained after hybrid feature selection, for dataset 2, we are using Soft-Voting subsets with 21 features for evaluation as it provided better results than the other four feature subsets in table 3. Table 6 shows us the output for dataset 2 with unbalanced and balanced data. The AUC value for unbalanced data ranges between 64 to 66, whereas the AUC value for balanced data is above 80. We can see that the base model almost gave a similar result, about 83% for AUC. The stacked ensemble approach was able to increase the value to up to 87% and give accuracy and F1 score of more than 86%. These results were obtained by using only 21 features out of the 32 features. From the feature selected by six feature selection techniques, we can say that PAY_0, PAY_2, AGE, LIMIT_BAL, PAY_5, PAY_4, PAY_3, and BILL_AMT1 are very important to determine if the customer is going to default in the credit card repayment. LIMIT_BAL is the amount of credit given, and PAY_6 to Pay_0 is the repayment status of the previous six months.

6.4 Experiment 4: Dataset 3 - German Credit Card (Size: small) using Soft-Voting

For experiment 3, we have used a small dataset with only 1000 records and 20 features. We tried fitting all four models with all five feature subsets provided by the voting technique in table 3. Out of all of the five subsets, the Soft voting subset gave us the optimum performance, and the output of models with soft voting feature subset is recorded in the 7.

Table 7: Results for Dataset 3 - German Credit Card (Size: small) using Soft-voting

		Unbalanced			SMOTE		
		Accuracy	**F1**	**AUC**	**Accuracy**	**F1**	**AUC**
Dataset 3	**XGboost**	0.74	0.736	0.697	0.916	0.917	0.864
	CatBoost	0.736	0.701	0.636	0.884	0.886	0.828
	LightGBM	0.746	0.732	0.684	0.9263	0.9262	0.875
	Stacked Model	0.73	0.737	0.709	**0.875**	**0.882**	**0.859**

From the table 7, we can see that there is a major difference between outputs of the balanced and unbalanced data. For unbalanced data, the higher accuracy was obtained by LightGBM, but if we observe the AUC, it is quite less. A higher AUC was obtained by the stacked model for unbalanced data. After applying SMOTE, the accuracy and AUC value went up by a considerate amount. We were able to improve the accuracy to up to 92 percent and AUC to move that 85 percent. The best performance was provided by the LightGBM model in terms of accuracy, F1-s core, and AUC. The features having a major influence on the credit-ability or credit scoring are the balance amount, the duration of the credit, previous credit status, current employment, savings value, credit amount, and age.

6.5 Discussion

In this research, we have implemented a hybrid feature selection prediction model. For hybrid feature selection, we have used six selection methods and applied a voting technique to derive a feature subset. The experiments are conducted on three datasets of varying sizes in the credit risk domain. For the second experiment, we have used a large dataset. The Soft-voting subset of 23 features provided the best performance as seen in experiment 1 6.1. The performance in terms of accuracy and F1- score shows the good predictive power of the model. Its shows that the preprocessing step for data cleaning and feature selection is done adequately. Also, the variation in the AUC value shows the importance of data balancing. Even though the accuracy of the unbalanced data is quite high, the AUC value suggests that the model is not able to differentiate between approved and non-approved loans. For unbalanced data, a difference of 10% can be seen in the AUC value of the base classifier and stacked classifier. The stacked model outperformed the base classifiers for both balanced and unbalanced data. But for the balanced dataset, the performance for all the classifiers is quite high, up to 95%. So the stacked model can not go beyond that. We can use the LightGBM model for this dataset after SMOTE. It will help avoid computational time used by the stacked model. This dataset is quite new, and not a lot of researches and experiments are conducted on this dataset, and hence no comparison with existing literature is provided.

For the third experiment on the medium-sized dataset, the soft-voting technique outperformed all the other voting techniques. A similar difference in the output of the AUC value as dataset 1 can be observed in the balanced and unbalanced datasets. This also validates our assumption of imbalance data in the credit risk domain. We were able to improve the performance of the credit card data, as well as the base classifier and stacked model, thanks to the hybrid feature selection. In comparison with the Yang and Zhang (2018), the proposed approach was able to increase the AUC value from 79 percent to 87 percent. We compare the result of the Xgboost and LightGBM model in the citepyang2018comparison. The values obtained for the AUC by the same models in our experiment are also greater. The Value for F1-score in the same paper is slightly higher by 3% than our results. Even though the performance of our approach in terms of accuracy and AUC is more, we can still find ways to improve the F-score value of our model. Singh and Sivasankar (2019) experimented with enhancing the accuracy of the credit default dataset. When we compare the results, our approach has about 5% higher accuracy. Moreover, our approach has a high F1-score and AUC, making it more reliable than only using accuracy.

The Third and final dataset is a small dataset with only 1000 records. For this dataset as well soft-voting provided the best results. But the hybrid stacked model performance is less than the base classifier. It may be due to the small size of the dataset. With a small dataset, there is a chance of variance that is the difference in the training and test performance. Our base classifier provided better results with no variance in the performance of training and test data. Hence, there is no variance and no overfitting. A stacked model with simpler models like logistic, SVM, Random-forest might work well with smaller datasets. When we compare our results with base paper Ha et al. (2019) that uses information gain for feature selection and LDA model for prediction provided an accuracy of 76.50%. Whereas the proposed hybrid feature selection model with LightGBM provided an accuracy of about 92%.

From the experiments performed on the three datasets, we can say that the hybrid selection approach using soft-voting has helped us increase the performance of the credit risk prediction when compared to the base papers and can be generalized for the credit prediction domain. It also helps us prevent the bias caused by the relational nature of the data. The hybrid stacked generalized model did not perform well with a small dataset, and for the large dataset, it showed only a small improvement. The stacked generalization model depends on the size and type of data. The base model for the stacked method should be picked after careful inspection of the dataset. Through this research, we were able to learn about the importance of feature selection and different approaches to it. Also, we were able to gain knowledge on ensemble techniques, data balancing techniques, optimization methods, and various techniques for data cleaning and preprocessing.

7 Conclusion and Future Work

The curse of dimensionality is one of the most significant challenges faced by banks. The datasets found in this domain contain a large number of features. Features need to be carefully selected to improve the performance of the prediction model. Another crucial issue with feature selection is the bias that causes a single feature selection model to select features that are not of any value for prediction. Our approach for hybrid feature selection using six different feature selection techniques combined with a voting mechanism was able to solve both issues. For all the three datasets used in the project, the soft-voting technique provided the best results. This Voting mechanism approach implemented in this research has helped improve the overall performance of the implemented models. Hence, we can generalize this approach for the credit prediction domain. Thus achieving the research objectives mentioned in the introduction section 1. For the prediction model as a base classifier, we have used XGBboost, CatBoost, and LightGBM. We were able to increase the performance of the base classifiers as compared to the base papers as referred to in section 6.5 which is in line with our research question. This was possible after careful pre-processing and transformation of the datasets along with feature selection. The stacked model, however, provided improved results with the large and medium-sized dataset. The AUC value obtained for the large dataset is 96%, and for the medium dataset is 87%. Also, the accuracy obtained is 95% and 86%, respectively. For the small dataset of German credit scoring, the result of the base classifier LightGBM is better than the stacked model making it more suitable for smaller datasets. Hence, the hybrid stacked model can not be generalized.

For further studies, we can try exploring the combination of deep learning with machine learning models. Neural networks work exponentially well with noisy and complex data. Neural networks are proven to provide better performance with large datasets. Given the nature of the data in the credit risk domain, a neutral network should yield good results. Also, incorporating the neural network algorithm may help us increase the performance of the stacked model and enable us to generalize the stacked ensemble method for credit risk prediction.

8 Acknowledgement

I would like to thank Prof Athanasios Staikopoulos from the National College of Ireland for his constant support and assistance. With every group and individual session, he guided us and helped us understand the expectation for the research project. I would like to express my heartfelt appreciation to all my professors and classmates for their support throughout the course.

References

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J. and Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets, *IEEE Access* **8**: 201173–201198.

Amirreza, R. and Hossein, N. (2017). A hybrid feature selection approach based on ensemble method for high-dimensional data, *2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, IEEE, pp. 7–9.

Bank, E. C. (2020). Non-performing loans, Website. last checked: 11.11.2021.
URL: *https://www.bankingsupervision.europa.eu/banking/priorities/npl/html/index.en.html*

Bhalla, D. (2015). Weight of evidence(woe) and information value(iv) explained, Website. last checked: 11.11.2021.
URL: *https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html*

Das, A. K., Goswami, S., Chakrabarti, A. and Chakraborty, B. (2017). A new hybrid feature selection approach using feature association map for supervised and unsupervised classification, *Expert Systems with Applications* **88**: 81–94.

Dhaiya, S. and Singh, N. (2016). Impact of bagging on mlp classifier, *International Conferences On Computing For Sustainable Global Development*, pp. 3794–3799.

Fitzpatrick, T. and Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market, *European Journal of Operational Research* **249**(2): 427–439.

Ha, V.-S., Lu, D.-N., Choi, G. S., Nguyen, H.-N. and Yoon, B. (2019). Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning, *2019 21st International Conference on Advanced Communication Technology (ICACT)*, IEEE, pp. 511–515.

Hajek, P. and Michalak, K. (2013). Feature selection in corporate credit rating prediction, *Knowledge-Based Systems* **51**: 72–84.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). Catboost for big data: an interdisciplinary review, *Journal of big data* **7**(1): 1–45.

Hsu, T.-C., Liou, S.-T., Wang, Y.-P., Huang, Y.-S. et al. (2019). Enhanced recurrent neural network for combining static and dynamic features for credit card default prediction, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 1572–1576.

Jadhav, S., He, H. and Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating, *Applied Soft Computing* **69**: 541–553.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems* **30**: 3146–3154.

Khemakhem, S. and Boujelbene, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach, *Accounting and Management Information Systems* **14**(1): 60.

Kozodoi, N., Lessmann, S., Papakonstantinou, K., Gatsoulis, Y. and Baesens, B. (2019). A multi-objective approach for profit-driven feature selection in credit scoring, *Decision support systems* **120**: 106–117.

Li, G., Shen, M., Li, M. and Cheng, J. (2021). Personal credit default discrimination model based on super learner ensemble, *Mathematical Problems in Engineering* **2021**.

Li, Y. (2019). Credit risk prediction based on machine learning methods, *2019 14th International Conference on Computer Science & Education (ICCSE)*, IEEE, pp. 1011–1013.

Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S. and Abu-Rub, H. (2021). A novel stacked generalization ensemble-based hybrid lgbm-xgb-mlp model for short-term load forecasting, *Energy* **214**: 118874.

Singh, B. E. R. and Sivasankar, E. (2019). Enhancing prediction accuracy of default of credit using ensemble techniques, *First International Conference on Artificial Intelligence and Cognitive Computing*, Springer, pp. 427–436.

Singhi, S. K. and Liu, H. (2006). Feature subset selection bias for classification learning, *Proceedings of the 23rd international conference on Machine learning*, pp. 849–856.

Son, H., Hyun, C., Phan, D. and Hwang, H. J. (2019). Data analytic approach for bankruptcy prediction, *Expert Systems with Applications* **138**: 112816.

Tang, L., Cai, F. and Ouyang, Y. (2019). Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in china, *Technological Forecasting and Social Change* **144**: 563–572.

Times, T. I. (2021). Irish banks' non-performing loans expected to increase due to pandemic, Website. last checked: 11.11.2021.
URL: *https://www.irishtimes.com/business/financial-services/irish-banks-non-performing-loans-expected-to-increase-due-to-pandemic-1.4513789*

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches, *Technology in Society* **63**: 101413.

Venkatesh, B. and Anuradha, J. (2019). A hybrid feature selection approach for handling a high-dimensional data, *Innovations in computer science and engineering*, Springer, pp. 365–373.

Yang, S. and Zhang, H. (2018). Comparison of several data mining methods in credit card default prediction, *Intelligent Information Management* **10**(5): 115–122.

Yu, L., Yu, L. and Yu, K. (2021). A high-dimensionality-trait-driven learning paradigm for high dimensional credit classification, *Financial Innovation* **7**(1): 1–20.

Zhu, Y., Zhou, L., Xie, C., Wang, G.-J. and Nguyen, T. V. (2019). Forecasting smes' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach, *International Journal of Production Economics* **211**: 22–33.