Arab American University

Faculty of Graduate Studies

Multi-Channel Prediction for Probability of Bank's Customer Credit Default Based on Machine Learning Techniques along with an earlier Recommendation System.

By

Fadi Hanna Issa Hazboun

Supervisor

Dr. Osama Mansour

This Thesis Was Submitted in Partial Fulfillment of the Requirements for the master's degree in Data Science and Business Analytics.

January / 2023

© Arab American University – 2023 All rights reserved.

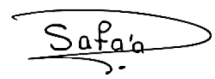
Thesis Approval

Multi-Channel Prediction for Probability of Bank's Customer Credit Default Based on Machine Learning Techniques along with an earlier Recommendation System.

By

Fadi Hanna Issa Hazboun

This thesis was defended successfully on 14/02/2023 and approved by:

Committee members	Signature
1. Dr. Osama Mansour/ Supervisor	
2. Dr. Majdi Owda/ Internal Examiner	
3. Dr. Safa'a AbuJarour/ External Examiner	

Declaration

I acknowledge that the thesis titled “Multi-Channel Prediction for Probability of Bank’s Customer Credit Default Based on Machine Learning Techniques along with an earlier Recommendation System” is my work, has been composed solely by myself, and does not contain work from other researchers, and has not been submitted for any other degree or scientific work except the reference is made.

Name: Fadi Hanna Issa Hazboun

Date: 11 /09/2023

Signature: 

Student ID: 202011757

Dedication

I extend my thanks and gratitude to my family and friends for their support and assistance during my educational journey, specifically my wife for her support, standing by me, and encouraging her to continue my educational path. I also dedicate this thesis to them for their unconditional love and support.

Acknowledgment

Through this opportunity, I would like to express my thanks and deep gratitude to Dr. Osama Mansour for supervising this work, standing by me, and supporting me during the period of preparing the thesis and reviewing my work, as well as providing all advice and observations that influenced this work. I also extend my thanks and gratitude to all the teaching staff of Arab American University for their dedication and their absolute support. specifically, Dr. Majdi Owda and Dr. Amani Owda, for providing advice and guidance that had a positive impact on refining the skills of preparing scientific research and overcoming obstacles.

List of Figures

List of Tables

List of Abbreviations

ACC	Accuracy
AI	Artificial Intelligent
AUC	The area within under the curve
BG	Bagging
BS	Brier Score
FN	False-Negative
FP	False-Positive
GBM	Gradient Boosting Model
GUI	Graphical User Interface
IV	Information Value
JAB	Jordan Ahli Bank
KNN	K-Nearest Neighbors
LDA	Linear Differentiation Analysis
MAE	Mean Absolute Error
ML	Machine Learning
MLT	Machine Learning Techniques
MSE	Mean Squared Error
NN	Neural Network
NPA	Non-Performing Assets
PD	Predict Default
RF	Random Forrest

RMS	Recommendation System
RMSSE	Root Mean Squared Error
ROC	Receiver Operating Curve
SMEs	Small and Medium Enterprises
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
UITR	User Items Trust Records
VIF	Factor Inflation Variance
XGBoost	Extreme Gradient Boosting
XGBT	Extreme Gradient Boosting Tree

Table of Contents

Abstract

Lending institutions and banks are surrounded by a multi-risk business environment, the most important of which are those related to clients' default and non-fulfillment of their obligations towards receivables, which leads to their inability to continue their activities and then failure and bankruptcy. The main challenge is to identify the most important indicators and variables that lead to the customer's default credit and build a hypothetical forecasting model that can effectively and accurately predict the probability of default before it occurs along with the recommendation system (RMSS). This is essential to take the necessary decisions and measures that prevent customers from defaulting in the future and to have access to a credit portfolio that is fairly free from defaulted accounts. This study was conducted on data provided by the Jordan Ahli Bank in Palestine, for 8506 borrowers between regular customers and defaulting customers. The dataset is for this study exclusively and has been approved by the Bank's management. In this study, different machine learning algorithms have been implemented because the probability is an expected value, the accuracy value of the most used methods exceeds 90%, and SVM is superior to other classification models. we propose an RMS based on an item-based collaborative filtering technique (which uses the customer preference data from the algorithm) that works side by side with machine learning algorithms to provide the necessary recommendation to avoid the client getting default. Our findings illustrate important techniques for detecting and predicting credit failures before they occur. Our methodology and results will help regulators, banks, and creditors to meet the challenges of credit granting and financial catastrophe and maintain a clean credit portfolio, by addressing various financial risks and credit rating improvements.

Keywords: classification; default prediction; credit risk; machine learning, recommendation system.

Chapter 1

1 Introduction

Financial institutions exist in a business environment surrounded by several risks [1] [2], the most important of which are those risks related to their failure to fulfill their obligations to others, which leads them to their inability to continue their activity, and consequently their failure and even bankruptcy [3] [4] [5] [6]. Therefore, it was very appropriate to search for methods and models that can predict the possibilities of credit default before it occurs, address the areas of default, and take the necessary decisions and actions that prevent customers from defaulting in the future [3] [6].

The current era is witnessing the technical development in electronic banking services and the increasing awareness of customers about the use of electronic payment tools, mobile banking applications, and Internet banking, and entering a new customer experience level, which led banks to increase the demand for high-quality financial engineering, speed of forecasting and financial analysis in value during the past few years [7].

Machine learning (ML) applications in banking enable businesses to automate regular, time-consuming processes that provide a simpler and more personalized customer experience. It also allows working more productively with large data, significantly improves the quality of asset valuation, and forecasts financial performance [8].

The applications of artificial intelligence (AI) have begun to enter the banking sector and their enhanced use recently in and rely on them in credit decisions, as ML models in the credit analysis processes of banks have contributed to the development of several statistical models that can predict the probability of the customer failing or not obtaining bank credit, using a set of variables represented in financial ratios in the first place [8] [1]. Several

models have proven their effectiveness in this area, for some time before the failure occurred [9] [10]. Although the design base of these models is similar, they differ from each other in the degree of the relative weight of these proportions. The Sherrod model is one of the most important models that banks rely on in assessing and minimizing the credit risks of institutions, in addition to predicting the possibilities of their failure, which increases the confidence in the process of making and taking the decision to grant credit or not [11].

The study in [8] illustrates the role of ML in FinTech as well as the use cases of deep learning in banking. Besides, the article focuses on the best effective models for using AI and ML to improve financial services. Along with applying smart recommendation systems (RMSS) to develop solutions that can avoid bank customers from defaulting by relying on the best algorithms of RMSS, machine learning algorithms can predict assumptions without manual interference. supervised learning has a deep background and practical value because default prediction can be automated progressively, especially with variables with multiple characteristics, where it is very difficult to accurately predict hypothetical risk through manual processing [8].

1.1 Background Analysis of Credit Defaults in Banks

The Concept of Bank Credit Default

Credit default means “those credit facilities of all kinds that the borrower obtained from the bank and fails to pay them on their due dates, and therefore the debt turns from current credit facilities to stagnant debit balances (Bad credit). In such an event, serious repercussions can happen, such as getting a poor credit rating” [12] [13], Credit default in

banks is defined as: “the inability of the borrower to pay the debt on the due dates, and this is represented in the principal amount in addition to the interest incurred amount” [12]. In general, the problem of credit default is known as a negative phenomenon that occurs as a result of a defect in the credit relationship between the bank and the customer [8]. The essence of this is the inability of the bank to collect the amount of credit and its interest or the inability of the customer to repay the amount of credit [6] [14].

The Main Concepts of Defaulting the Bank Customer (Corporate or Individual Customers):

A. Financial Insolvency: In its general sense, financial hardship means the inability of the project or client to pay and fulfill its obligations towards others. So that the customer becomes stumbled when it is unable to generate a cash surplus through the activity it practices in a way that is sufficient to meet the burdens and obligations of this activity, and then the customer’s intrinsic ability to pay its current obligations decreases [5] [11].

B. Financial Failure: The term financial failure indicates that revenues are unable to cover expenses including the cost of money, especially the cost of debt represented by interests and installments or repaying the entire debt on due dates [8].

C. Bankruptcy: This is a legal proceeding involving a person or company that is unable to pay its outstanding debts, or where the assets are unable to cover the indebtedness as the bankruptcy process begins with a petition by the debtor, which is most common. “All assets of the debtor are measured and valued, and the assets can be used to pay off part of the outstanding debt” [11] [4] [15].

1.2 Reasons for Credit Default Related to the Bank

The recording of shortcomings in the credit analysis process and the various precautionary measures surrounding the process of granting bank credit would contribute significantly to creating the problem of credit default [16], and therefore among the most important causes that must be examined and analyzed, are those causes in which the bank bears a significant amount of responsibility for the occurrence of the problem of defaulting on the credit that it grants to its customers. the most important of these reasons can be summarized in the following elements [16]:

- Inability in estimating the value of the collateral.
- Deficiencies in research and credit follow-up and decision-making.
- Mistakes in estimating credit risks and exaggerating profitability.
- Falling into the credit risk concentration.
- Bad management and low administrative efficiency in banks.

1.3 Background of Recommendation Systems.

RMS are very popular in recent years and are used in all walks of life and are often used in a variety of decision-making processes. Recommended systems include movies, music, news, etc. Find out what users are interested in and recommend it to customers with a large amount of information. An RMS is a branch of an information filtering system that is used to learn user behavior or to predict their orientation and preference for items based on the user's experiences, choices, or similar behavior. “There are usually two ways for an RMS to create a recommendation list: collaborative filtering and content-based recommendation” [17]. “The collaborative filtering method builds a model based on the

user's historical behavior" [18] (according to items purchased, aggregated, and rated) along with similar decisions of other users. This model is used to predict the behavior of users according to the items they are interested in, based on the recommendation of the content related items and similar characteristics of the user's behavior are recommended.

In this thesis, we consider customer credit and demographic data as primary sources of information for different classification models, this data was obtained by a specialized team in the Credit Control Department of Ahli Bank. The data will be used to analyze the performance of existing and new customers to predict customer credit ratings.

1.4 Objectives

ML is a field of AI, which is one of the computer science branches. ML applications in banking are very limited, especially in Palestinian Banks, Applying ML techniques in the banking sector can be used as a decision support system for credit granting applications, as well as can detect, recognize, and analyze key features or anomalies from a set of data beyond what the human capabilities can detect.

To the best of my knowledge, the proposed model will improve the use of ML in the banking field by enabling businesses to automate regular, time-consuming processes that provide a simpler and more personalized customer experience. It also allows working more productively with large databases, significantly improving the quality of credit valuation, and forecasting financial performance, especially in the field of credit risk.

Looking and reviewing the open literature we can state that there are very limited studies in ML for banking systems that build their prediction based on one type of credit account without providing any tools related to decision-making supported by submitting the

recommended recommendation to reduce the entry of the customer in the default state. In this thesis, I going to build a novel ML model to predict the probability of a customer default by analyzing all credit accounts using a Multi-channel prediction of customer default probability as well as an intelligent RMS based on “item-based collaborative filtering technique (which uses the customer preference data from the algorithm)” [19], The automated decision from the model can subsequently be relied on by lending institutions and banks (the Credit control department) to allow the appropriate and timely decision to be taken to avoid the client's default before it occurs and to assist in making the appropriate decision whether to grant credit or not in the event of new lending requests.

In this research, novel datasets collected from the Jordan Ahli Bank credit accounts will be manipulated and used to learn a system to be able to predict customer default probability and classify the credit using naive Bayes, decision tree, random forest, KNN, support vector machine (SVM), and XGBoost classification models.

1.5 Thesis Outline

The rest of this work is coordinated as follows. Chapter 2 is a literature review of the related work in ML for banking applications and some techniques used to predict and classify credit default and techniques used in building RMS in multiple environments. In chapter 3, the proposed method will be illustrated for both ML tasks and the intelligent RMS. In chapter 4, the results for classification and regression using different evaluation metrics will be presented. Chapter 5 will represent the conclusion and future work.

Chapter 2

2 Literature Review

This chapter provides a summary and analysis of relevant publications on these thesis topics on default risk in banks, ML in financial institutions and banks, ML in the financial field and its impact on all categories of financial institutions, and smart RMS complementing ML.

2.1 Introduction

The use of machine learning (ML) and data analytics in banking institutions is more than a trend and has become essential to keep pace with market competition and reduce credit risk. In recent years, the creditworthiness of the customer has become more and more important for financial institutions. Many of the earlier credit risk models, which are "statistical tools, are used to infer the future probabilities of customers of becoming default.

2.2 Default Risk in Banks

The exponential increase in the volume, diversity, and speed of data generated by various banking and commercial transactions, poses a major computational and storage challenge for data analysis and intelligence tasks" [10], The risks of granting credit constitute the greatest concern for banking activity, and the risks vary according to the internal and surrounding circumstances, and also differ according to the different types of operations practiced by financial institutions. The risks greatly affect lending operations, and therefore it is necessary to know these risks and devise ways to reduce them [20]. "Many studies have focused on studying the impact of ML, as a sub-field of Artificial

Intelligence (AI)" [20], in helping anticipate the credit default probability before it occurs and reducing the risks of credit grants. The study in [20] presents an analysis of the adequacy of credit classification models using the loan database for a bank in Brazil and explores ML techniques namely: SVM, decision trees, AdaBoost models, and random forests, The study [20] then "comparing their predictive accuracy with a criterion based on a logistic regression (LR) model. The comparisons are analyzed based on the usual classification performance metrics" [20].

Bankruptcy is one of the biggest potential risks in banking institutions, which is represented in the bank's inability to commit to refunding depositors' money when requested, One of the important reasons that lead to the bankruptcy of the banking institution is the inability to collect the borrowed money, in terms of the capital granted and accrued interest, and therefore the bank has a cash liquidity shortage, and the loss of depositors' liabilities such as deposits and savings accounts and inability to pay interest, from On the other hand, the bank is not able to return the money to its owners if requested, and as a result bankruptcy occurs, as happened in the global financial crisis in 2008, Several recent studies have predicted Bankruptcy , and they have achieved a promising direction in this field using ML to predict bank failures and bankruptcies[21] [4]. The study in [21] introduces a framework of SVM from the field of ML to build a prediction model for bank failures, and compares it with Ohlson's bankruptcy prediction model, [22] proposes a new form of stress testing that informs interested parties of the necessary actions that a particular bank must take. This is taken to avoid failure shortly, the model came up with a prediction accuracy of 99.9%.

Some studies have demonstrated the significant effects of hyperparameters on ML

techniques used to assess credit risk. The study [23] “analyzed the limitations of existing studies and research trends” [23]. It was found that cluster and hybrid models with neural networks and SVM are more adopted for credit scoring, Non-Performing Assets (NPA) prediction, and fraud detection. The study [5] concluded that good prediction requires the availability of comprehensive data sets related to the risks of granting credit in the region to which the financial institution is subject. The authors in [14] proposed an LR algorithm to predict the probability of loan default of actual data of a commercial bank, by combining with an artificial neural network to improve the predictability of the model, the experimental results revealed that the proposed combined model outperforms the two basic models in four evaluation criteria: “accuracy (ACC), and the area within under the curve (AUC), the Kolmogorov-Smirnov statistic (KS), and Brier score (BS)” [14]. In addition to hyperparameter tuning, it is also very important to experiment with many ML algorithms, then the study [6] compares the results of these algorithms to find the most appropriate models for the data [12]. A group of nonparametric ML algorithms and two conventional parametric models were compared in the credit default swap study that reflects credit risk. The company has one of the most widely traded credit derivatives, and artificial neural networks have proven to perform better than other parametric and nonparametric models. The results of this study [11] in the proactive assessment of the economic health of financial institutions to predict bank failures indicate that the random forest (RF) method has high predictive performance compared to the rest of the models used with the performance of neural networks also almost equally, Then the study [8] concluded that the measures related to profit and capital factors of higher importance are a marginal contribution to predicting bank failure.

2.3 ML in Financial Institutions and Banks

The importance to achieve a good predictive probability is to evaluate the variables that should be relied upon. Some studies [12] [24] [25] [8] have looked at the important aspect of the extent to which the variables are used to influence the performance of the model. The researchers investigated three potential ML tasks to estimate the risk of default: "predetermined overdue days cutoff by binary classification, regression of the overdue days, and learning-to-rank which learns to optimize the risk-related ranking for the full range of probability" [12] [24], The study was based on the analysis of changing consumer characteristics, credit scores, and features analysis of customer income, balance using machine language algorithms. The study [25] found that ML predictions can capture the dynamics of changing credit cycles as well as default rates [25], They provided an analysis of the application of ML algorithms to a set "of sixteen financial variables from the financial statements of listed Taiwanese companies as inputs for the six approaches" [25]. This data was entered for "six classification algorithms, traditional SVM, HACT, GA-fuzzy hybrid clusters, XG-Boost, DBN classifier, and DBN-SVM hybrid, to predict financial distress" [25]. "The XG-Boost model produced the most accurate prediction performance" [8], Fundamental factors for credit decision-making and fair transparency of grants are credit risk prediction, effective loan processing, model monitoring, and reliability. In [26] binary classifiers based on deep and ML models were built on real data in predicting the probability of loan default, The results of the study found that tree-based models are more stable than models based on multi-layer artificial neural networks. The authors [2] introduced "four ML methods (random forest (RF), extreme gradient boosting tree (XGBT), gradient boosting model (GBM), and neural network (NN) to predict

important factors affecting repayment" [2], among the factors studied were borrowers who passed video or mobile verification, employment, residence, or education level, as these factors were most likely to default, the study showed that RF shows outperformance over other rating models, The study in [27] assess individual information to determine if a person can meet the criteria for credit financing, the study focused on comparing the performance of five common "ML classifiers used for credit scoring, Naive Bayesian Model, LR Analysis, Random Forest, Decision Tree, and K-Nearest Neighbor Classifier. the results of the experiment indicated that Random Forest performed better than others in terms of accuracy, recall, and AUC" [27]. The authors in [28] presented various hypothetical "prediction models based on credit scoring methods and computational techniques using ML algorithms" [28] for a loan finance database in Brazil, the study concluded that the accuracy of the models improves with the increase in the number of days of past due used to determine the default variable. "The best prediction results were obtained using traditional ensemble techniques bagging (BG) and RF" [28].

For surveys on default probability prediction. In early proposals for statistical default prediction, the focus was primarily on linear classification models such as LR or linear regression as linear differentiation analysis (LDA) where covariates pass into the model as a linear combination, The authors in [4] presented a study on the prediction of Default Probability Prediction through statistical methods, where the author relied in this study on the Shumway's model (c = 1) or the pooled proportional hazards model (c = 0) when an accurate hypothetical probability is needed. the types of lending and the products offered to customers vary from one bank to another.

2.4 ML in the Financial Field and its Impact on all Categories of Financial Institutions

Banking institutions often focus on granting loans to individuals because individuals have a broad base and some of the most important loan products for individuals are mortgage loans. Credit rating forecasting models typically involve the use of data sets with highly balanced interest event (default) distributions. LR is often used to estimate predicting default (PD), and statistical methods have been used to improve the ability to predict trip probability regardless of sample size [9]. In [29] the authors suggested a model for predicting Mortgage defaults through ML, a linear logistic regression (Logit), and a GTB. The variables of loan ratio, amount, and current interest rate were used in the study. The effects of features were measured by interfering with inputs and estimating shapely values [9]. this study in [15] presents a standard economic method for "estimating the term structures of corporate default probabilities over multiple future periods" [15], conditioned by the firms and macroeconomic covariates. The method allows, according to its assumptions, to combine the "traditional duration analysis of the dependence of event severity on time-varying covariates with the traditional time-series analysis of covariates" [15]. The study [15] found that the estimated structures of default risk rates for individual firms in this sector are highly dependent, in terms of level and form, "On the current state of the economy, and especially on the company's current leverage, as recorded by distance to default" [15], a measure of leverage volatility is common in the banking industry.

Small and Medium Enterprises (SMEs) that mostly rely on bank loans are an essential role in the economy of many countries, This makes SMEs riskier than their larger counterparts parts, and this prompts banks to develop hypothetical prediction models that

specifically address the credit relationship management of SMEs, The study in [30] adopted an LR model estimated on a credit rating agency database, where the results showed that among the observed “financial and economic characteristics of companies, capital structure (in terms of internal and external funds” [30] and terms of external funding source) and interest expense are “more relevant than economic variables as determinants of SME's default” [30]. Conventional statistical models cannot predict the probability of default effectively and accurately. To address this problem, the authors [31] propose a hypothetical prediction model for a heterogeneous set of high-dimensional data through a decision tree model for accurate prediction of customer default, “GBDT, Extreme Gradient Boosting (XGBoost), and Gradient Boosting Machine were used. LightGBM” [31] as individual classifiers to generate a learning-based hypothetical prediction model in a heterogeneous population. The learning model-based feature classification “is applied to the P2P lending credit data, and the individual classifiers are subject to hyperparameter optimization” [31].

2.5 Smart Recommendation Systems Complementing ML.

During the past few decades, with the emergence of YouTube, Amazon, Netflix, and many other web services, RMS has occupied an increasing place in our lives and has become very important in most different areas of life [32]. From various factors, including user preferences, “these systems predict the most likely product that users are likely to buy and be interested in” [33].

These systems work using one of the methods specific to each of them and as needed in a specific way that may be singular or using more than one type of method of recommending systems with each other and works automatically without the intervention of the user [34]. “The purpose of a RMS is to suggest relevant elements to users by using one of the different types of recommender systems” [19].

There is a variety of RMS that have been applied primarily in most areas of life, such as the “Collaborative Recommender system, the content-based recommender system, the demographic-based recommender system, a utility-based recommender system, a knowledge-based recommender system, and a hybrid recommender system” [17], each with a role and impact according to the field of application [17] [33] [35].

2.5.1 Recommendation Algorithm Based on Collaborative Filtering

Collaborative filtering is a recommended method widely used in recommender systems. This algorithm is based on the premise that “things are grouped by groups, and people are divided by groups” [19]. “Users who like the same items are more likely to have the same interests” [32]. Filter-based collaborative RMS are generally applied to systems with user ratings, and the results are used to discern users' preferences for items. Collaborative filtering is a collective wisdom usage model, and “does not require special handling of items, but establishes the relationship between items through users” [32]. Collaborative filtering RMS are currently divided into two types: user-based and object-based.

2.5.1.1 User-Based Recommendations System

“The basic principle of user-based collaborative filtering recommendation is that, based on all users' preferences (scores) for items or information found to be similar to the current user's tastes and preferences” [18], the "neighbor" user group in general applications, an algorithm is used to calculate K-nearest neighbors; Based on the information of the historical preferences of these K-neighbors, “recommendations are made to the current user” [36]. This RMS is a feature in that the content of recommended items may be completely irrelevant, “so you can discover the interests of potential users and generate personalized recommendation results for each user” [36]. The system is easily throttled. Therefore, there are few user-based collaborative filtering systems in the industry.

2.5.1.2 Item-Based Recommendation System

“Object-based collaborative filtering is similar to user-based collaborative filtering. It uses all users' preferences (results) for items or information” [37]. Find out the similarities between items and items, and then recommend similar items to the user based on the user's past preference information. Item-based collaborative filtering can be seen as a kind of deterioration recommended by association rules However, since collaborative filtering considers the actual ratings of users and only calculates similarity rather than searching for recursive groups, item-based collaborative filtering can be considered to have higher accuracy and coverage higher. Compared to user-based recommendations, object-based recommendations are more widely used, with better algorithm scalability and performance. Since the growth rate of the project is generally constant, the performance does not change much. The disadvantage is unable to provide personalized recommendation results.

2.5.2 Collaborative Filter Algorithm Summary

"The recommendation mechanism based on collaborative filtering is the most widely used" [32].

Advantages of collaborative filtering over content-based filtering

Collaborative filtering offers many advantages over content-based filtering. The most important of them are as follows:

- "Not required to understand item content: Item content does not necessarily tell the whole story, like movie genre/genre etc." [19].
- There is no problem with cold starting an item: even if no information is available about the item.
- It embodies "the change in user interests over time: focusing only on the content does not provide any flexibility in the user's perspective and preferences" [19].
- "Captures the inherent subtleties: This is very true for latent factor models. If most users buy two unrelated" [19].

Among the most popular applications of recommendation, systems are "e-commerce, e-learning, e-library, e-government and e-business services, films, music, TV shows, books, documents, websites, conferences, tourism sites, and learning materials" [19] [33] [38]. The RMS can predict the item or service based on the user profile, and it is not limited to the user only, but it is useful for both service providers and users because of its many benefits for each of them, for example, It reduces private transaction costs by accessing In addition to the element of time and its use, RMS also help users "get personalized recommendations, help users make the right decisions in their online transactions, increase

sales, redefine the users' web browsing experience, retain customers, and enhance their shopping experience" [39] [40].

Many studies provided solutions for financial systems through the application of RMS models, The study [33] presented a system for generating recommendations for the banking sector to increase the sales of the bank's products according to customers' orientation, thus increasing the market value and enhancing banking services. To improve prediction accuracy they use a recommendation model based on collaborative filtering and demographic filtering, the study relied on the customer's evaluation of the product, which the author faced two challenges in applying the bank's cooperative approach: 1- "The lack of explicit evaluations of products by customers for recommendation" 2 "The starting problem cold that makes it difficult to recommend products to a potential or new customer who has no preference at all." The results of the study revealed that the hybrid RMS achieved the highest accuracy in prediction, as this system combines both the recommendation model based on cooperative filtering and demographic filtering.

Similar to the previous study, the researcher [35] presented a proposal developed to serve financial technology specifically in the processes of recommending banking services to the bank's customers, relying on predictive systems to improve the bank's automated and customer-oriented services and achieve the greatest degree of transparency and accuracy of decision-making. Models were developed based on collaborative filtering along with content-based filtering, these models were built to analyze data on loans, mortgages, and retirement plans. The results of the study help the banking sector in improving its services and satisfying customers by recommending products that meet their desires and aspirations

in an automated and intelligent manner and saving material costs to provide a human cadre to follow up on the tasks of systems recommendation.

The study [41] used the RMS in the field of filtering university students and classifying them against the student's academic performance. The RMS provides more appropriate educational plans and programs for the student, specifically for students with low performance. The study relied on a cognitive recommendation model. The model analyzes the student's information before university until the beginning of university education. The model categorizes students in terms of academic performance risks from high risk to low risk, and thus this classification helps to extract the appropriate recommendation to the university administration about directing the student to the correct academic path and recommending an academic program more suitable for the student. The study was based on building classification models using ML models. three types of classifiers, Naïve Bayesian Classifier, Decision Tree Classifier, and Fuzzy Classifier. In the same context and the field of individual learning, the study [17] provided a personal RMS for distance learning. The system recommends educational programs and courses according to the educational path of learners through the Internet, using a set of association rules, content filtering, and collaborative filtering to improve system performance and independence of learning and raise student competency.

Collaborative filtering models are considered one of the most highly efficient models in forecasting based on data analysis from the previous experiences of the data set [39], However, filtering systems are not free from the problem like a bias through equal opportunity between elements. One of the most common problems facing the collaborative filtering algorithm is the popularity bias problem, by filtering the most common element

with certain attributes at the expense of the least common elements. This type of bias is considered to confuse recommender systems and leads to negative outcomes in the final decision-making process. The results of the study [40] showed that bias can arise between elements in this type of model, and therefore the study suggested bias processing algorithms by integrating edge processing algorithms in collaborative filtering algorithms compared to the baselines designed for the traditional fame bias. On the other hand, the excessive influence of homogeneity with the structure of collaborative filtering is one of the most important reasons that lead to the inaccuracy of the RMS and the deterioration of its results [36], as well as the deviation and dispersion of user interactions from the standard deviation. These challenges are faced in models based on raw and unprocessed data. Therefore, the treatment of Data by statistical methods may result in clean data and have a strong impact on the performance of prediction models, the author [36] proposed a divergent cooperative filtering model to jointly capture deviant cooperative relationships with a differentiated learning structure to comprehensively capture complex relationships and influence between the elements, the study [36] was based on enhancing neural cooperative filtering using a paragraph Self-supervised learning and the results showed a significant improvement in the performance of the models. In the same context, the study [32] suggested a new adaptive prediction mechanism called genetics-based prediction. The model is based on improving the prediction tool on the principle of determining the optimal predictor for each user based on concept levels rather than the level of representation through collaborative filtering techniques. The results of the study [32] showed that the system has reduced the effect of data variance and the scattering of the elements on which the system is based. The historical behavior data of users is one of the most important

elements on which Collaborative Filter RMS are based in favoring items [42], and this is mainly based on the data source on which the system depends. The lack of historical behavior data for users affects the ability to prefer items and sometimes the accuracy of the recommendation is incorrect. The researcher [42] introduced the “Hybrid Collaborative Filter Recommendation approach with User Items Trust Records (UITR) which is a new approach that integrates user trust into CF-based methods” [42] in a harmonious way to complement classification information as study results achieved better accuracy after following the UITR approach.

The study in [39] used a collaborative filtering model to analyze human emotions through the information and results of analyzing the user's speech according to the verbal emotion in the mobile application to provide a recommendation for applications that suit the user's feeling and purpose and increase the user's satisfaction. The SVM algorithm was used, where the results showed the model's accuracy with the highest emotion recognition rate of 86.98%. The results showed that the RBF kernel achieved the best performance with an accuracy of 94.70% and a recall of 94.77% and an F scale of 94.71%. The study showed that modern advising systems have the capabilities to analyze facial gestures and speech and plans for the study to develop an intelligent system that may be used in the fields of AI to allow robots to understand human emotions and act based on feelings toward the user.

Table 1: Summary of Related Studies' Recommendation Methodologies.

Methodologies	**Classification**	**Conclusions**
Collaborative filtering [19] [42] [39] [36]	Memory-Based (User/Item-Based)	1. Fast inclusion of up-to-date information 2. get the best recommendations 3. Poor prediction due to unavailability of ratings 4. Scalability is a challenge 5. Individual user preferences can be identified
	Model-Based	1. Easy to use and fastest customization recommendations 2. The data is compacted only at the stage of creating the form 3. Suitable for systems where user preferences slowly change to the time required to build the model
	Deep-Learning	1. Capable to process non-linear data. 2. Used in complex or numerous training situations. 3. Required sufficient data.
Content-Based Filtering [35] [33]	Hybrid approaches	It allows "the system to consider collaborative data without relying on it exclusively, so it reduces the sensitivity of the system to the number of users" [35] to which a specific item applies.

2.6 Conclusion

From reviewing the literature models, it can be seen that the focus is on predicting the default probability. To the best of my knowledge, no studies in the open literature have addressed the combination of predicting banking credit default with an early intelligent RMS. Moreover, the lack of studies that pay attention to the possibility of explaining the AI model in the default prediction process, motivates us to investigate this route of research. In this study, various prediction techniques have been proposed and compared to achieve the best accuracy that enables us to predict the probability of the client's default before it occurs, as well as the intelligent RMS based on collaborative filtering.

This study aims to construct a novel early warning model based on ML to predict the probability of a customer default by analyzing all credit accounts in addition to the ability to classify the credit status to achieve high predictive performance with the conclusion of the necessary recommendations to enhance customers' ability to repay debts and avoid default by applying the best techniques in RMS using a collaborative filtering algorithm, through customer data analysis (financial and demographic), this approach will be done on a novel dataset of the Al-Ahli Bank in Palestine. The available data is related to credit and demographic record information, according to the instructions of the Palestinian Monetary Authority.

Our methods and conclusions can assist banks in establishing a credit scoring model for borrowers. My approach can assist in predicting a borrower's credit health and ability to repay by combining with the evaluation and study of comprehensive borrower information from credit assets, credit history, the profitability of the existing business, business growth rate, working capital turnover, industry development prospects, operating

status of subsidiaries, and audit data from industrial, commercial and collection departments.

The following models are included in our analysis: Logistic Regression (LogR), Random forests (RF), AdaBoost, XGBoost, Naive Bayes, Support Vector Machine (SVM), K-nearest Neighbors (KNN) along with a recommendation system using a collaborative filtering algorithm.

Chapter 3

3 The Proposed Method

3.1 Introduction

This chapter illustrates the proposed method, which aims to construct a novel early warning model based on ML to predict the probability of a customer default with the conclusion of the necessary recommendations by applying the best techniques in RMS using a collaborative filtering algorithm. It begins by forming the dataset and describing the preprocessing steps then the chapter illustrates the deployed supervised ML models that can combine the prediction of the customer's probability of default and classify the credit risk, based on the two criteria of customer classification specified in our data set. Default and Non-Default credit, then it will illustrate the technique of building the intelligent RMS based on a collaborative filtering algorithm. Finally, it illustrates the metrics used to measure model performance. The classification and regression were implemented using Python 3 with Pandas, Seaborn, Sklearn, Matplotlib, and NumPy libraries; we used a platform of intel core i7 processor 2.6 GHz with 16 GB RAM running Windows 11 Laptop for applying this approach.

3.2 Dataset

The data set used represents a sample extracted from the banking services system for credit control this dataset contains 8506 bank customer accounts with 16 different features. The necessary approvals were taken from the bank’s senior management. The data set is considered unique and extracted to implement this study. The data set contains

a mixture of good and bad credit account information for individual clients and small and medium-sized companies. These accounts have been classified by the credit control department. The data extracted in the dataset covers the period between 2000 and 2022. Credit accounts consist of a mixture of credit requests between direct and indirect facilities.

The dataset was extracted and stored for bank customers using a CSV file with 8506 rows and 21 features as shown in Table 3.1, feature No. 18 represents the credit status, the value that will be relied upon to predict the credit status of good or bad credit. This area represents the credit status decision made by the bank credit control team as they are experts in evaluating loan applications and making the appropriate decision to grant or not.

Table 3.1 Data Set Features Description

Feature Name	**Feature Description**	**Feature Type**
BRA_CODE	*Represent Customer geographic location*	*Number*
FAC_TYPE	*Represent the credit category*	*Number*
BALANCE	*Represent Customer Balance*	*Number*
ACCOUNT_OPEN_DATE	*Represent Credit Account Open Date*	*Date*
DUE_AMT	*Customer due amounts*	*Number*
SUSP_AMT	*If the customer has any outstanding interest*	*Number*
LOAN_AMOUNT	*Requested credit amount*	*Number*
TYPE_OF_DEP	*Customer type of depositor classification.*	*Number*
CUST_MGR	*Represent Customer category classification.*	*Number*
INSTALLMENT_AMOUNT	*Requested installment amount.*	*Number*
INSTALLMENT_PER	*Required credit term.*	*Number*
MIN_INT_RATE	*Credit Interest rate.*	*Number*
RET_CHQ	*Represents the number of customer rejected or returned checks.*	*Number*
DUE_INSTALLMENT	*The number of installments owed by the customer.*	*Number*
DUE_DAYS	*The date of the first unpaid due.*	*Number*
AGE	*Customer Age by years.*	*Number*
DATE_OVDN	*The date on which the customer stopped paying any dues.*	*Date*
MONTHLY_SAL	*It represents the customer's monthly salary amount in USD*	*Number*
DUE_CLASS	*Credit classification (Default, Good)*	*Character*

3.2.1 Data Exploration and Visualization

Exploratory data analysis is used to better understand the data set and uncover insights from the start or identify areas and patterns to dig deeper. Data exploration helps in analyzing the data and identifying general trends, anomalies, patterns, and relationships within the data, the goal of this stage is to explore the information contained in the Jordan Ahli Bank (JAB) credit account data that can play an important role in guiding our choices when building learning models.

To help explore the distribution of data by a classifier from which to predict a customer's credit status, a bar chart was used to compare the distribution of data according to a customer's credit rating status, which helps to show the distribution of data points or create a comparison of metric values across different subsets of the data set.

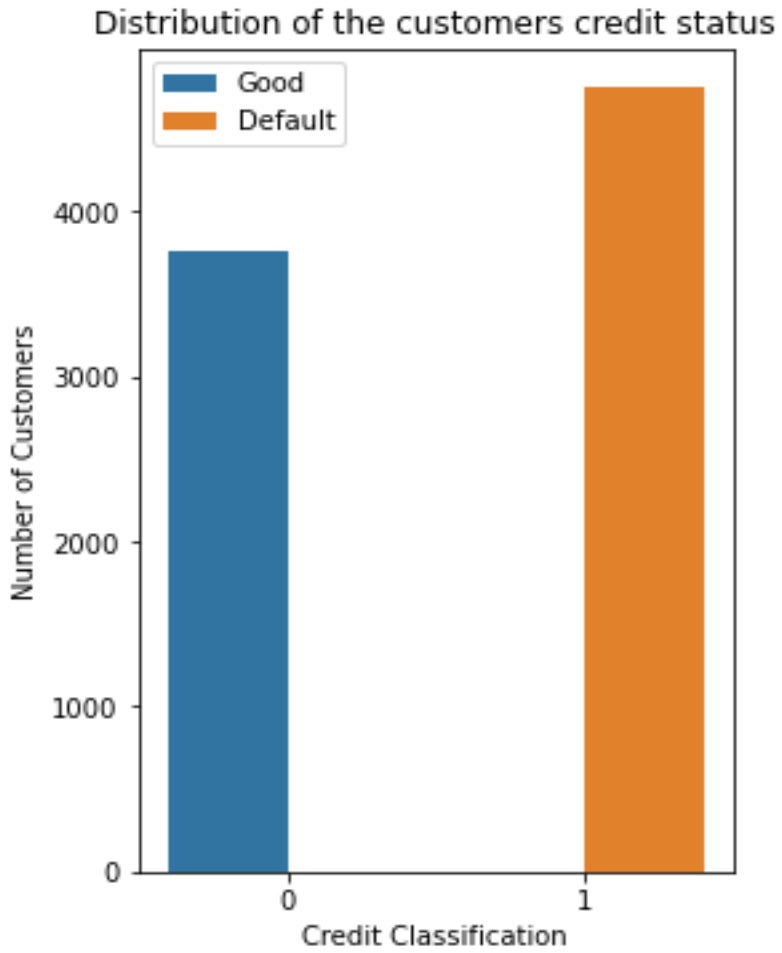


Figure 3.1: Bar Chart for Distribution of the Customer's Credit Status.

Figure 3.1 shows the distribution of data points in terms of the customer's credit status rating. We also note that the number of customers with a good credit rating is close to 800 customers out of the total data set, as well as customers with a bad rating of close to 1200 out of the total data set.

To explore the relationship between the features of the data set, a heat map matrix was used for all the elements of the data sets. The heat map helps to know the correlation between the features to know whether there is a sufficient correlation that adequately predicts customer default, Figure 3.2 shows the correlation matrix for the features of JAB customer's credit portfolio dataset.

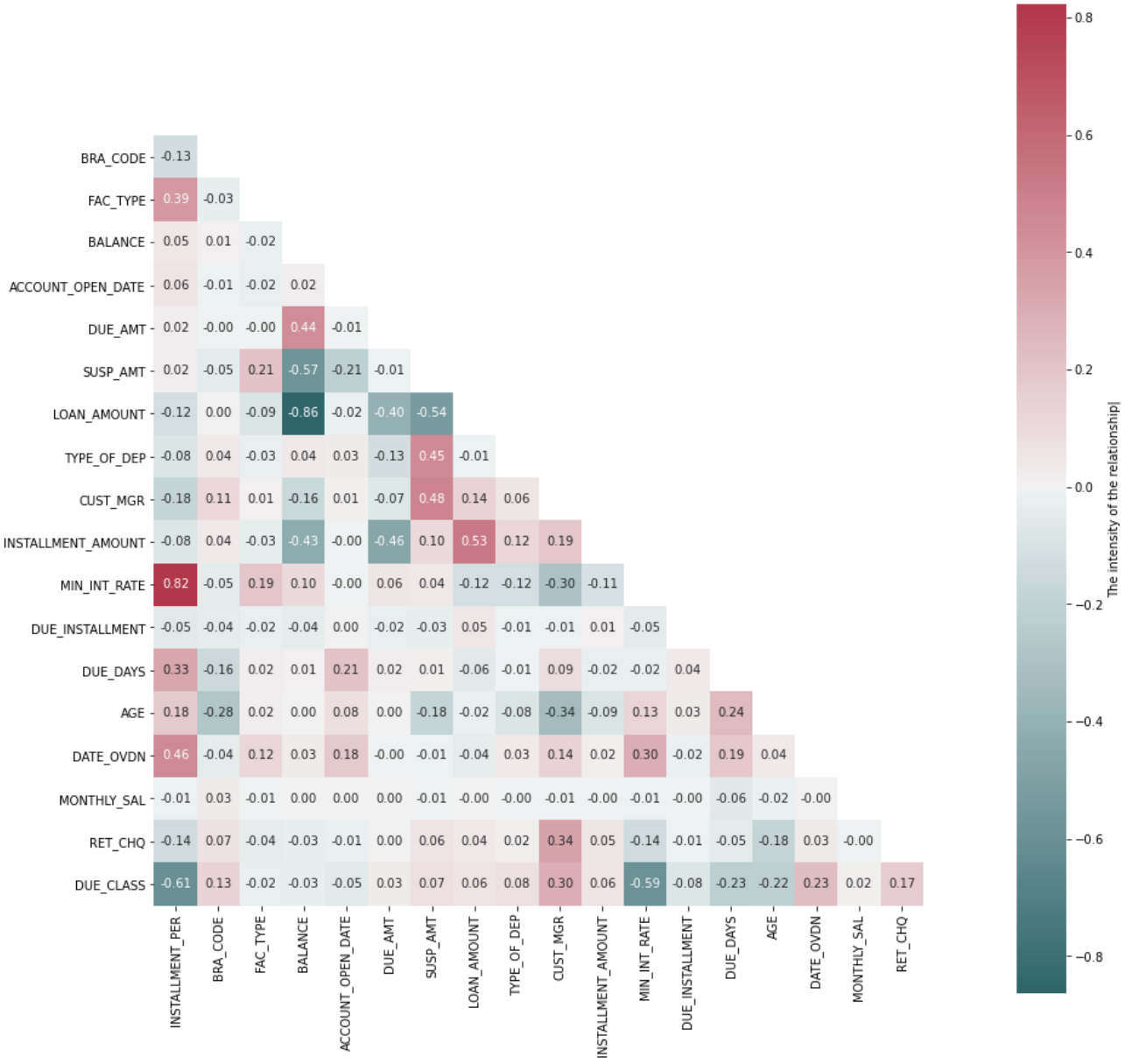


Figure 3.2: Heat Map for Measurements Locations

Figure 3.2 shows that there is a real correlation between the dataset features. By looking at the heat map, we conclude that there is a relationship between the amount of credit granted and the payments owed by the customer that causes the customer to fail to pay his due installments. On the other hand, we note that there is a strong relationship between the first

date of stopping paying and the total suspense interest resulting from the non-payment of the customer's due, which in turn also affects the classification of the customer.

To further study the relationships between the dataset features, Figure 3.3 a horizontal bar plot was used to study how each feature was related to the credit rating classifier. Through this, we determine how each variable affects the credit rating of the customer. In this figure, the type of effect is shown if it is positive or negative, where the bar in the positive influence is headed to the right and the bar in the negative influence is headed to the left.

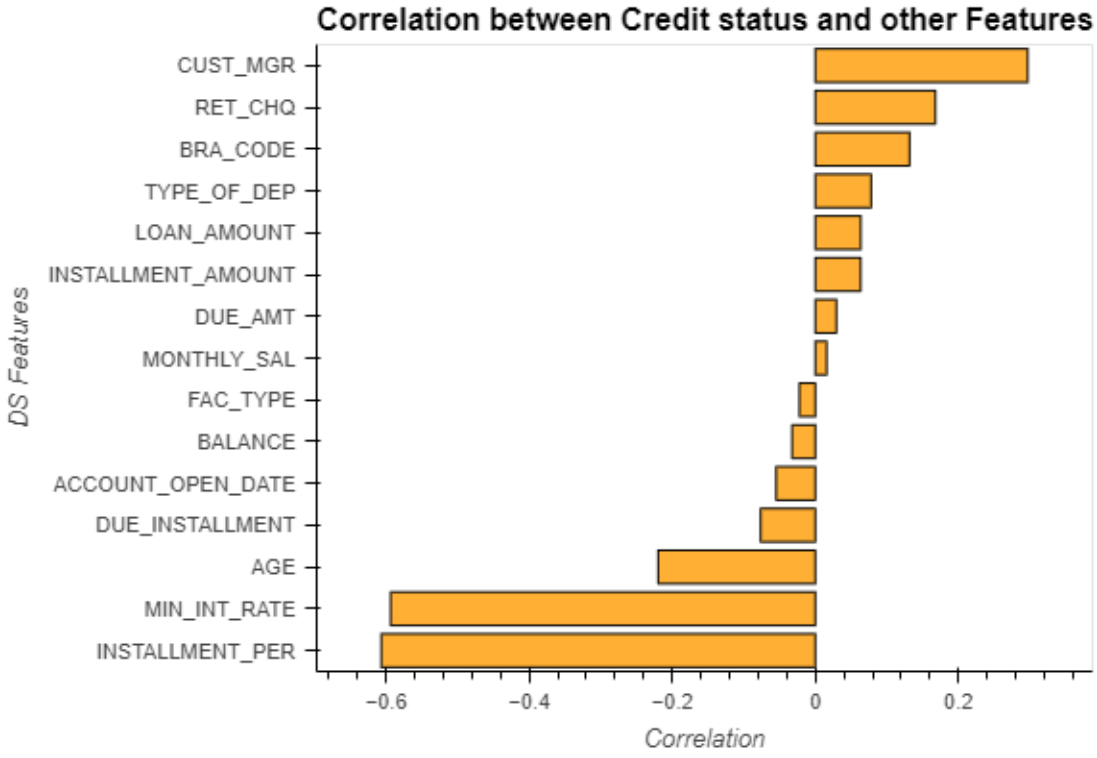


Figure 3.3: Correlation Between Credit Status and Other Features.

Through Figure No. 3.3, it is clear that the increase in the interest rate owed by the client and the higher due date is an important factor in classifying the client as a defaulter. The figure also shows that the fewer due days, the less incidence of customer default, the figure shows that there is a relationship between a customer's credit rating and the

customer's age, to dig deeper we used a Histogram to consider the nature of this relationship.

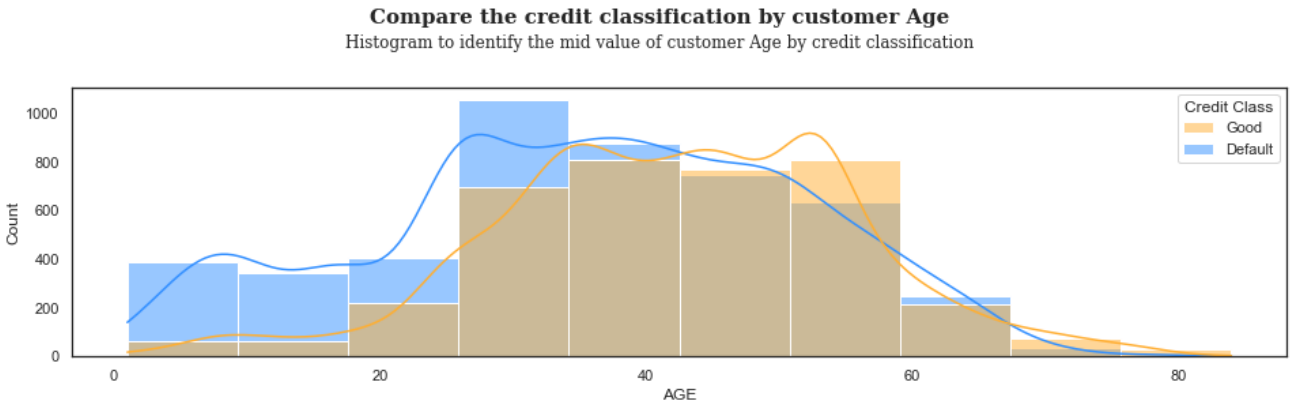


Figure 3.4: Histogram for Customer Credit Classification by Customer Age.

In Figure 3.4 Histogram is used to show frequency distributions between age versus credit rating, as it turns out that the rate of defaults decreases for customers aged between 40 to 60 years. This conclusion indicates the extent to which the client's age affects his financial stability and ability to repay the debt.

Figure number 3.5 the bar plot showing how a customer's annual income affects his credit rating, the figure shows that the higher customer's income (transferred salaries) positively affects the customer's ability to pay his debt, as the graph shows that the probability of default is high for customers with an annual income of less than $30,000, while the probability of default is lower for customers with the total annual income is more than $30,000, and we note that there is good stability in the credit status of this category of clients.

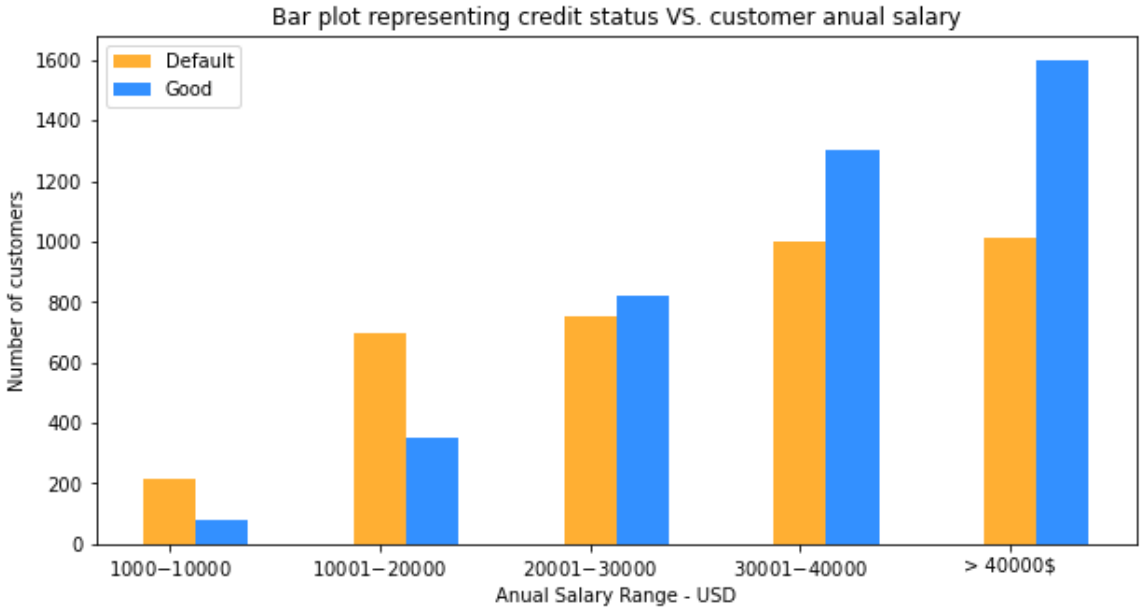


Figure 3.5: Bar Plot Credit Status Vs Customer Annual Salary.

3.3 Data Preprocessing

Data pre-processing is an important step for building data prediction models and increasing their efficiency. “Data collection methods are often loosely controlled, resulting in out-of-range or missing values. Analyzing data that has not been carefully examined to counter these problems can lead to misleading and illogical results” [43]. Oftentimes, data pre-processing is the most important stage in ML and thus, data representation and quality precedes everything before analysis is performed.

If there is a lot of correlational or redundant information or garbled and unreliable data, discovering knowledge during the training phase becomes more difficult. “Data preparation and filtering steps can take a significant amount of processing time. Data pre-processing includes data normalization” [43], Data Standardization, feature selection, optimization, transformation, feature selection, etc. the last result of data processing is the final training set. The following section will describe data preprocessing steps in detail.

3.3.1 Data Normalization

Data normalization is the process of transforming data shapes, which are usually complex, overlapping, repetitive, and random, into a solid, integrated, and interconnected fabric of data, i.e., transforming the complex structure of the data into a simple data matrix that represents all the elements and variables of the data.

Currently, “there are many data standardization methods, which can be divided into straight-line methods such as the min-max value method and the standard deviation method, broken line methods” [44] (such as the trifold line method), and curvilinear methods (such as the quasi-normal distribution). Different calibration methods will have different effects on the evaluation results of the system, and unfortunately, there is no universal rule to follow in selecting methods for data standardization [44].

In this work, the range of all features normalized to be between [0-1], the max-min normalization method was used to perform a linear transformation on the data, max-min normalization method was calculated using the equation

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

While x' is the normalized value, x is the original feature value, $x$min is the minimum value of the feature and $x$max is the maximum value of the feature.

3.3.2 Data Standardization

Standardization of the data set positively affects the performance improvement of many algorithms in ML (including gradient descent). If the data is not uniform (for example, when the features of the data set differ by an order of magnitude), then many algorithms perform poorly.

The used dataset contains categorical features such as gender and ethnicity, and some used algorithms such as the SVM that cannot work directly with categorical data, require all input variables and all output variables to be numeric. We converted the categorical data into numeric discrete data by using the integer-encoding technique.

3.3.3 Optimization and Feature Selection

Optimization and Feature selection “is very important for data scientists and ML practitioners” [21]. A good choice of features can improve the performance of the model and help us understand the basic properties and structure of the data. This has an important role in further improving the model and algorithm [21]. “Feature selection can eliminate irrelevant or redundant features, reduce the number of features, improve model accuracy, and reduce runtime. On the other hand, identifying truly relevant features simplifies the model and makes it easier for researchers to understand the data generation process” [45]. The feature selection phase begins after data pre-processing is completed, at which point meaningful features and input ML algorithms and models are selected for training.

In the actual application of ML, the number of features is often large, there may be unrelated features, and there may also be correlations between features, which easily leads to the following results:

- Analyzing the features and training the models took longer.
- Increased probability of causing a "Dimensional Disaster".
- The complexity of the model increased.
- The predictability of the model is reduced.

3.3.3.1 Significance Level and P-Value

we will choose the features based on the backward elimination method, this method is used to preserve the important features in the data set that are significant, and the insignificant features will be removed through the P-Value criterion, “The level of significance is the amount of change that the feature will affect towards the final output” [17] as the choice of these features greatly affects the dependent variable on the model and so on to reach features with a level of significance less or equal to 0.05 of the significance level. Figure number 5 shows how the backward elimination algorithm works [46].

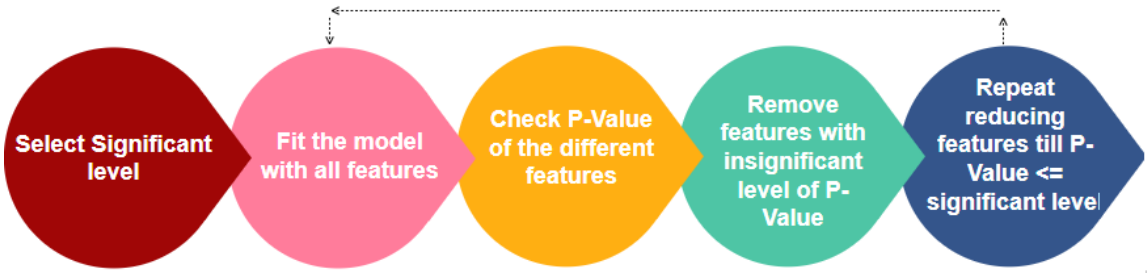


Figure 3.6. Backward Elimination Algorithm

3.3.3.2 VIF (Factor Inflation Variance) Calculates Correlation and Multicollinearity

The value of the variance inflation factor when it is greater than 5 indicates the existence of collinearity between the independent variables under the data set “the minimum possible value is 1 which indicates no collinearity” [47]. The variance inflation factor is calculated as follows:

$$\mathrm{IFj} = \frac{1}{1 - \mathrm{R}j^2}$$

R^2 = coefficient of determination of the variable X extracted from the regression of Y on the independent variables where j=1,2,…….k It is used to measure the degree of variance as an indicator of the accuracy of increasing or decreasing the variance. If the strength of the correlation between the set of independent variables increases, then R rises Thus, the

denominator decreases, the variance inflation factor increases, the variance increases, and thus the measurement accuracy decreases. the variables that showed a high correlation were eliminated by calculating VIF.

3.3.3.3 Cross Validation

Cross Validation method mainly aims to use all the data for training to ensure that the model works efficiently on all the data of the data set, to increase the model's accuracy during deployment and dealing with real data, instead of splitting the data into training data and test data which reduces the data we can use for training, we use cross-validation with a fixed number of K-Folds.

The training data is divided into K number of containers and then we iterate the following K times: each time we train the model with the data of K-1 container (all containers except container i) and then evaluate it with the data of container i. Finally, the final performance measure is the average rating scale for all iterations (i:1..K).

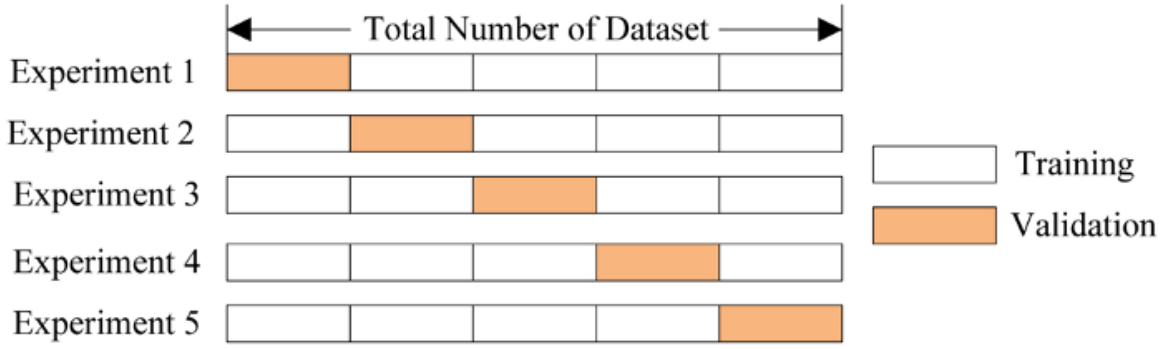


Figure 3.7. 5-Fold Cross Validation

3.3.3.4 Grid Search

In the proposed models, Grid Search will be used to reach the optimal models in terms of accuracy of prediction of customer default cases, and to build models with the best

efficiency. Through this technique, a set of different values of hyperparameters are determined, so that they are entered periodically during the training process, and their values are calculated based on training accuracy each time to determine the best hyperparameter values. This method takes a long time to train ML models due to the large number of hyperparameters involved.

My goal to use this method is the automatic tuning of the parameters, and since the data set used is rather small, which allowed the use of Grid Search for all ML models used in this thesis.

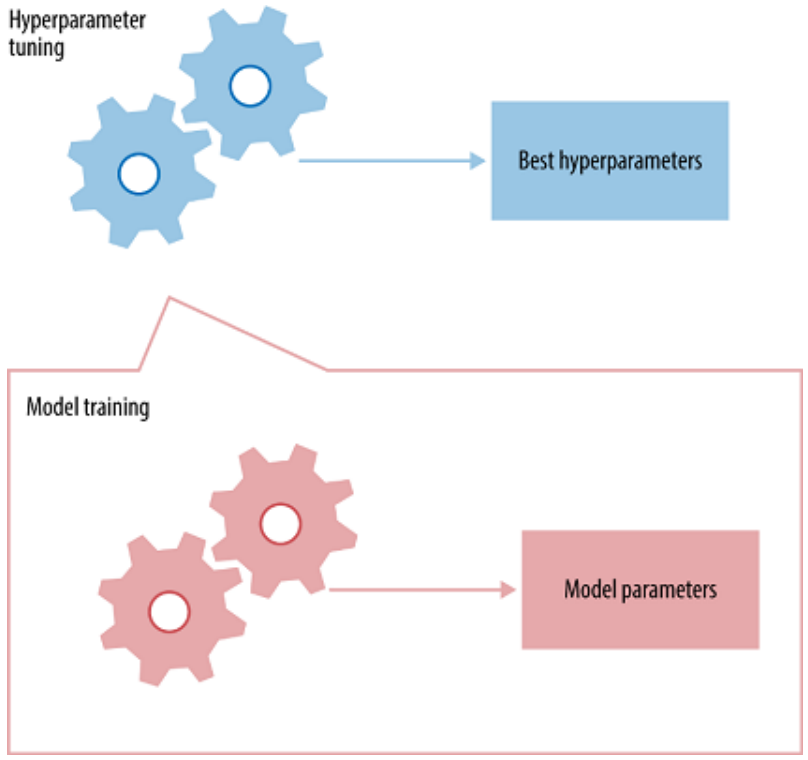


Figure 3.8. Hyperparameters Tuning

3.3.3.5 Outlier Removing

The outliers are any data point with an abnormal distance from other data points, data outliers removing can improve the result's accuracy [48], so we removed the outlier's values, which are bigger than 97% of the dataset; also, we removed the outlier's values,

which are less than 97% of remaining dataset, As we can see by the following boxplot figures 3.9, 3.10 which shows outliers for customer interest rate and customer age, these were the outliers that needed to be removed.

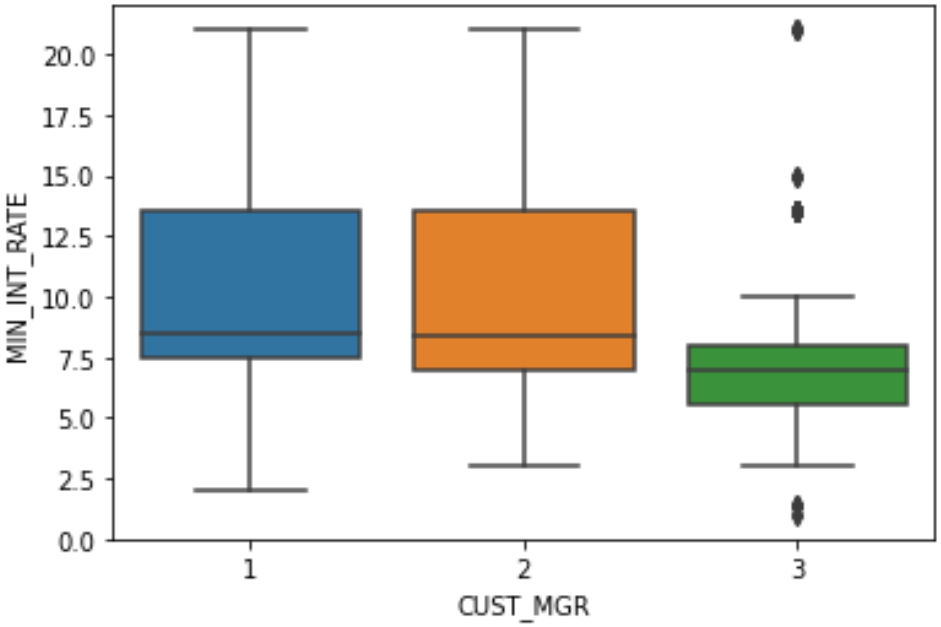


Figure 3.9. Boxplot Interest Rate Distribution.

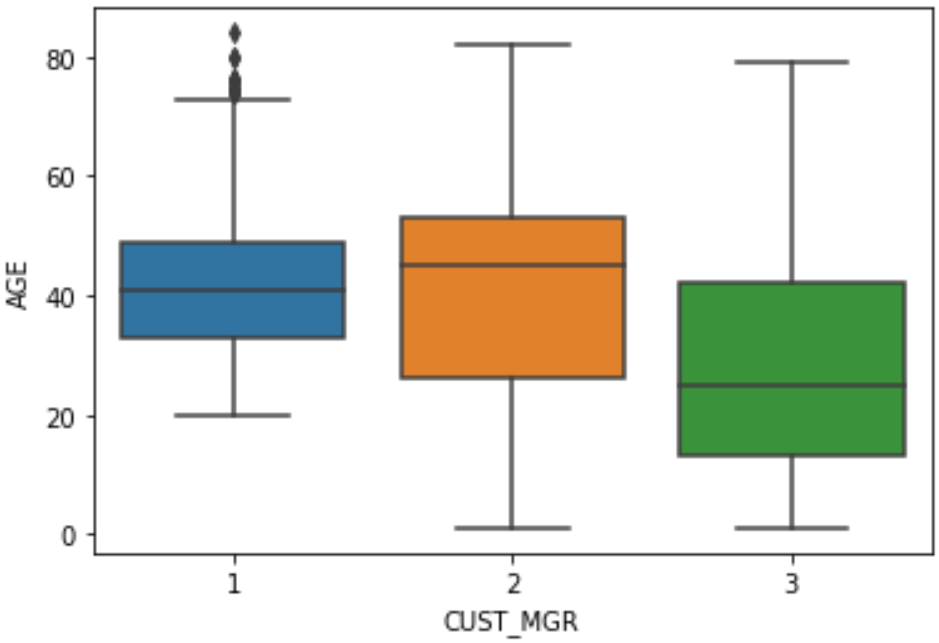


Figure 3.10. Boxplot Customer Age Distribution.

3.4 Model Development Phase

In this research, several models for forecasting and early warning of credit default cases were built through several axes. The first axis is the classification to predict the credit status in terms of good or defaultable credit through analyzing the demographic and financial data of the bank's customers, the second axis is the construction of a RMS based on the principle of early warning and the indication of the criteria that the bank should take to avoid the bank's customer entering into default situations, based on the literature and previous cases and the important default variables resulting from ML models built to predict other customers who are within the default field, we used the following ML algorithms to build our model and comparing the efficiency of each model to reach the adoption of the best one with high efficiency and accuracy in predicting and classify customers in terms of credit quality: LogR, RF, AdaBoost, XGBoost, Naive Bayes, SVM and KNN along with a RMS based on item-based collaborative filtering technique. Figure 3.11 shows the proposed method for default prediction model workflow.

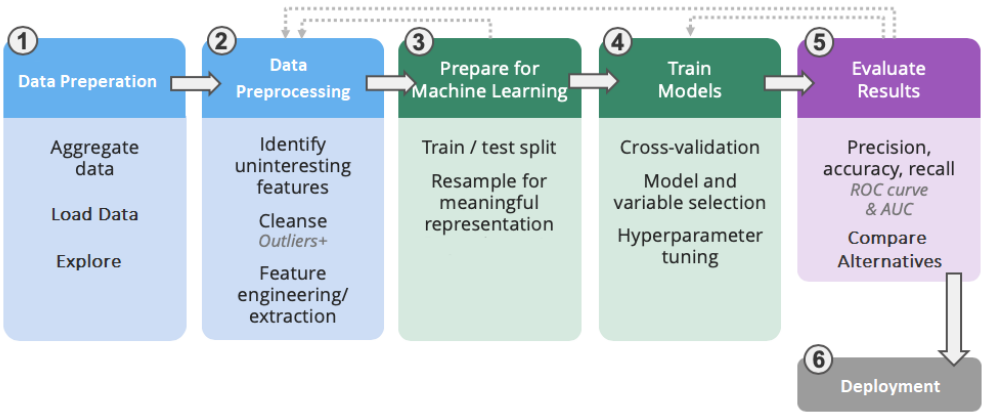


Figure 3.11: Supervised DP Model Workflow.

Several stages were passed to build a prediction model for DP.

Stage 1 Data preparation:

According to the data, determine the type of problem to which the modeling analysis in ML belongs, then data loading and exploring.

Stage 2 data Preprocessing

The quality of data processing directly determines the performance of the subsequent model. Including data cleaning, feature extraction, missing value, and data transformation.

Stage 3 & 4. Model selection

After data is processed an appropriate algorithm should be selected to train the data. The preprocessed dataset was split into 70% for training and 30% for testing, the training data set was used as data input for one of the used

Stage 5: Training Model

The process of training a ML model, feeding the algorithm with training data, enabling it to learn from them according to the division that took place in the data preparation stage. The training process provides the ability of models to process large amounts of data quickly, identify patterns, and find deviations or test correlations that are difficult for a human to do out without help.

Stage 6: Evaluate the model

Evaluation of ML models is fundamental in ML issues, it helps determine how and why ML models behave the way they do. These techniques are used to improve the performance of the model, at this stage a set of performance evaluation measures is calculated to demonstrate that the learned model has reached a good degree of learning so that it can be relied upon and used in a problem (generalized).

These metrics ultimately allow comparison of the various possible learning models (each generated using a particular algorithm), allowing the selection of the most appropriate ones.

Stage 6: Deploy the model

Deploying a ML model is the process of applying and using the model in a live work environment to use it within its intended purpose and to make practical business decisions based on the real-world impact of the data. This stage lies in its importance in the life cycle of the ML model and is considered one of the last stages of adopting the model.

3.4.1 Supervised Prediction Models

Supervised learning algorithms are the types of ML in which machines are trained using well-labeled data. The data is known as training data, and it consists of a set of training examples. Each training example contains one or more required inputs and outputs, also known as a supervisory signal. In supervised ML models, "each training example is represented by a vector, sometimes called the feature vector, and the training data is represented by a matrix. Through iterative and equivalent optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs" [8]. Supervised learning is the most popular way to train ML models. This technique relies heavily on the information provided by a pre-determined classification system.

In this chapter seven individual classifier models will apply to predict the probability of bank customers' credit default, Individual classifiers for heterogeneous ensemble learning Logistic regression, Tree-based algorithms, AdaBoost, XGBoost, KNN, and SVM.

3.4.1.1 Logistic Regression (LR)

The logistic model is one of the most important statistical models for modeling the "probability of a particular category or event such as success/failure" [49]. Whereas LR uses several predicted variables which can be numeric or categorical. This can be extended to model several classes of events. It will be assigned a probability between 0 and 1 for a feature detected during learning so that the total equals one [50]. LR is also known by other names as the Logit model or the general classifier of entropy. LR is one of the supervised ML algorithms for classification tasks, as a function of explanatory variables, the probability log changes linearly in the following equation:

$$Log(odds) = lg\left(\frac{p}{1-p}\right) = B_0 + B_1 * X_1 + B_2 * X_2 + \cdots + B_n * X_n$$

The pre-filtering variable was used by calculating the information value (IV) for each feature to filter them before entering them into the model, IV was used to reduce the variant and To determine the predictive ability of an independent continuous variable x in convergence with the dependent variable Y, it is usually used to reduce the number of variables in the LR if there are a large number of variables to reach the best results and with the least errors of predictability.

LR Advantages: -

- LR is easier to implement, and interpret and is highly effective training.
- It can be easily expanded into multiple classes (polynomial regression) and a normalized probabilistic view of predictors of frequencies.
- Very fast in classifying unknown records.
- It is very efficient in dealing with simple data and it also works well when the data set is linearly separable.

- It can interpret the parameters of the model as indicators of the importance of the feature.
- LR does not tend to overfit, but it may increase fitness in multidimensional datasets.

LR Disadvantages:-

- LR leads to overfitting If the number of observations is less than the number of features.
- It builds linear boundaries.
- Assuming linearity between the dependent variable and the independent variables.
- Non-linear problems cannot be solved using LR because there is a linear decision surface.

3.4.1.2 Decision Tree

"Decision tree learning is used to start from observations about a particular element and to conclude about the value of that element represented by the leaves of the tree, while the element itself is represented by the branches of the tree" [51]. The decision tree algorithm falls into the category of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used to solve regression and classification problems as well.

The general motivation for using a decision tree is to create a training model that can be used to predict the category.

The level of understanding of the DT algorithm is very easy compared to other classification algorithms. DT algorithm attempts to solve the problem using a tree representation. "Each inner node of the tree corresponds to an attribute, and each leaf node

corresponds to a class label" [52], Figure 3.10 represents the general structure of the DT algorithm

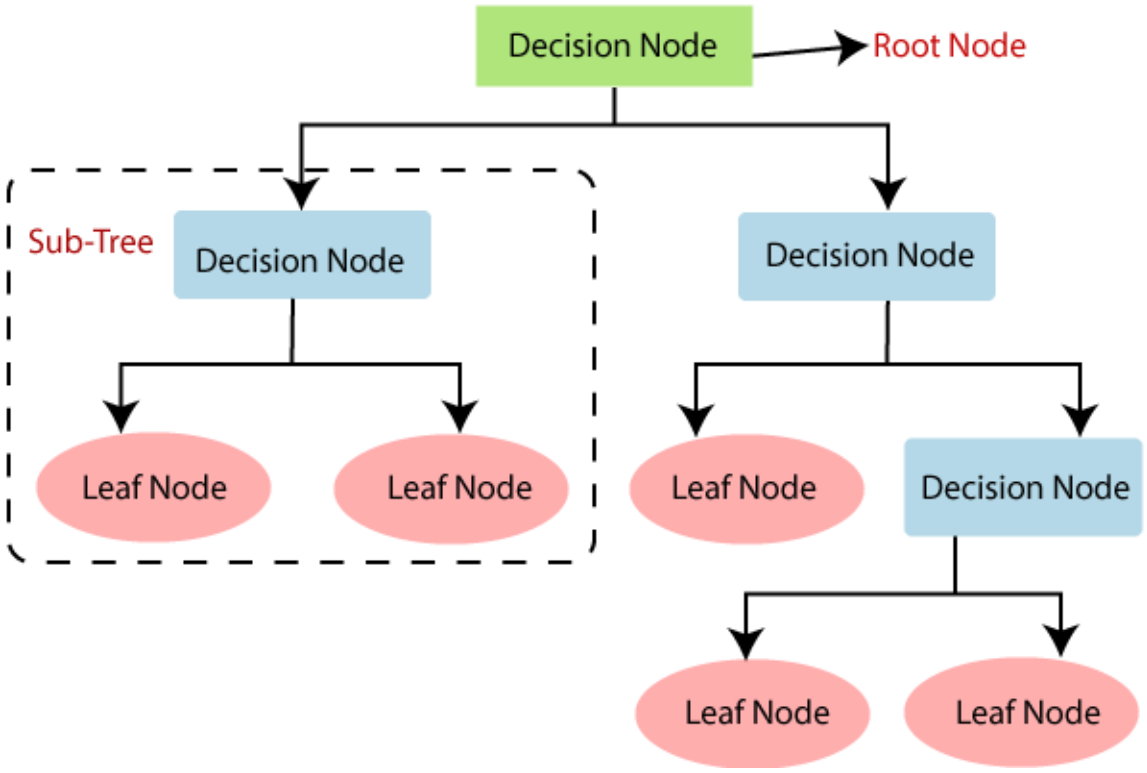


Figure 3.12 Decision Tree Architecture

DT Advantages:-

- Decision tree algorithm is easy to explain. This results in a set of rules.
- It follows the same approach that humans generally take while making decisions.
- The interpretation of a complex decision tree model can be simplified by its visualization.
- The number of hyper-parameters to set is almost null.

DT Disadvantages:-

- There is a high probability of overfitting problems in the decision tree algorithm.

- Decision tree algorithm in general, gives lower prediction accuracy for a data set compared to other ML algorithms.
- "Information gain in a decision tree with categorical variables gives a biased response for traits with the largest number of categories" [52].
- Calculations can get complicated when there are many class labels.

3.4.1.3 Random Forest

A random forest algorithm, as its name suggests, consists of a large number of individual decision trees that act as a set. Each individual tree in the random forest emerges from the class prediction and the class that gets the most votes becomes the prediction of our model as seen in figure 4.3.

The reason why the random forest model is so successful in data science fields is that a "large number of relatively uncorrelated (trees) models that act as a committee will outperform any of the single component models" [53].

"The low correlation coefficient between the models is key" [53]. Not all trees often err in the same direction, which results in higher accuracy in model performance, as the trees protect each other from individual errors, where the accuracy is greater for uncorrelated tree models, "While some trees may be wrong, many others will be right, so that the trees as a group can move in the right direction" [53]. So, the basic requirements for a random forest to perform well are that:

- There should be some actual cues in the features so that models built using these features perform better than random guessing.

- Predictions (and thus errors) made by individual trees must have a low correlation coefficient with each other.

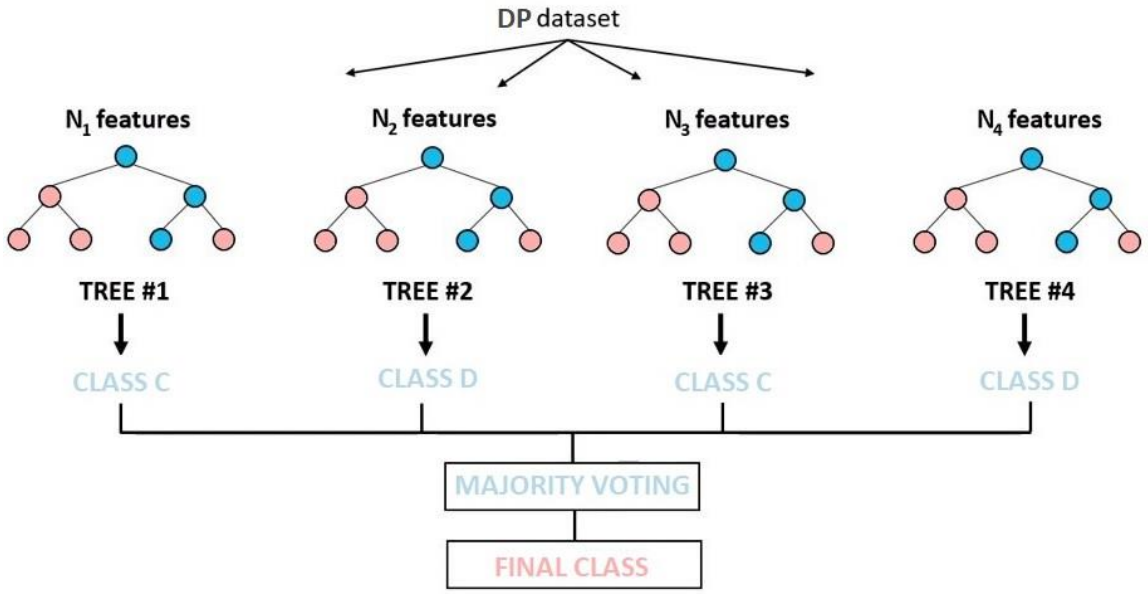


Figure 3.13: Random Forrest Architecture.

RF Advantages:-

- Improve accuracy by Reducing the overfitting.
- The algorithm is flexible for both classification and regression problems.
- "It works well with both categorical and continuous values" [53].
- It automates the missing values in the data.
- Normalizing the data is not required in this algorithm because it uses a rule-based approach.

RF Disadvantages:-

- It requires a huge amount of computational power as well as resources because it builds many trees to combine its output.
- Takes a long time to train because it combines a lot of decision trees to define a class.

3.4.1.4 AdaBoost

In ML, a boosting algorithm is a descriptive group algorithm whose main function is to reduce bias, as well as variance in supervised learning. It is also considered as a group of ML algorithms that transform weak learners into strong ones. A weak learner is defined as a classifier that is only slightly related to the true classification. In contrast, a strong learner is a classifier that is closely related to the true classification.

While boosting is not computationally restricted, most boosting algorithms consist of iteratively learning weak classifiers concerning distribution and adding them to a final strong classifier. When added, they are weighted in a way related to learners' poor accuracy. After a weak learner is added, the data weights are reset, known as "re-weighting". Wrong input data gain more weight and properly categorized examples lose weight; Thus, “future weak learners focus more on the examples that previous weak learners have misclassified” [53].

sample weights calculated by this equation:

$$w(x_i, y_i) = \frac{1}{N}, i = 1,2, \dots n$$

Where N is the total number of data points.

3.4.1.5 XGBoost

The XGBoost algorithm “is an efficient implementation of the Gradient Boosting algorithm” [2]. Due to its excellent effect and efficiency in application practice, it is widely respected by the industry, "In the XGBoost model, k number of regression trees are generated to ensure that the prediction of the tree cluster is as close to the actual value (accuracy) as possible and that the generalization power is as high as possible” [54][55].

XGBoost is like GBDT optimization. It is based on the CART tree. It consists of several weak classifiers through a negative gradient optimization framework, an integrated learning method for learning strategies, it is expressed by the following mathematical equation:

$$\hat{} = sigmoid(f1(x1) + f2(x2) + \cdots + f_m(x_m))$$

XGBoost passes all the feature split points and uses the above objective function value as the evaluation function [55]. “The specific method is that the value of the objective function after splitting is greater than the gain of the objective function of the single leaf node”. Meanwhile, to limit the growth of the tree too deeply, a threshold is added, and division is performed only when the gain is greater than the minimum.

"A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model" [54] This is shown in Figure number 3.14.

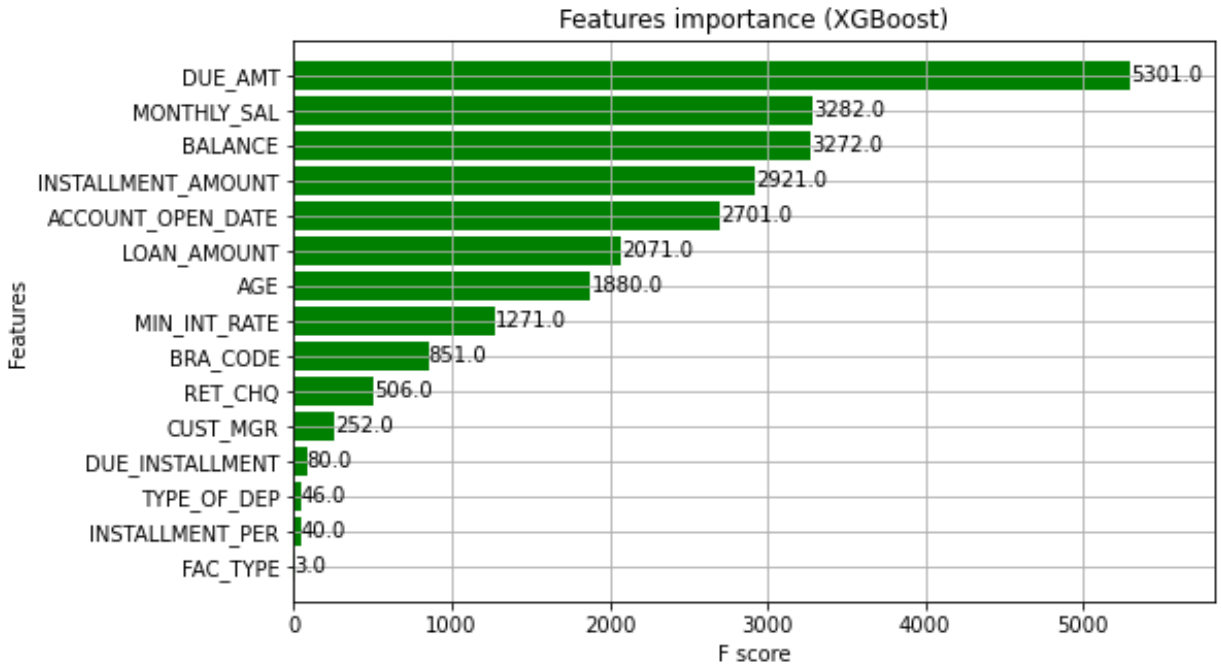


Figure 3.14 XGBoost Features Importance.

3.4.1.6 Nive Bayes

The Naive Bayes algorithm is a classification technique based on Bayes' theorem with the imposition of independence among predictors. “The Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the existence of any other feature” [56].

The Naive Bayes algorithm is easy to create and is especially useful for very large data sets. Besides simplicity, Naive Bayes is known to outperform even other very complex classification methods.

Bayes theorem provides a method for calculating the posterior probability P(c | x) from P(c), P(x), and P(x | c).

posterior probability equation:-

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} = \frac{P(c \cap x)}{P(x)}$$

Where $P(c|x)$is the posterior probability (likelihood) of class (target) for the given predictor (attribute).

$P(c)$ is the class prior probability.

$P(x|c)$ is the probability which is the probability of expecting a given class.

$P(x)$ is the previous probability of the prediction.

3.4.1.7 KNN

The Nearest Neighbor algorithm is considered one of the most important and simplest directed ML algorithms - which works with a supervisor. The Nearest Neighbor algorithm is one of the “descriptive and predictive classification algorithms” [1]. It can deal with anomalous data [1]. The principle of work of this algorithm is based on calculating the Euclidean distance between points, where the less the distance between two points, the

greater the probability of the point belonging to each other, hence the name of the algorithm [57]. If we assume that k = 1, the algorithm will measure the distance between the target point and the nearest three points to it. If the two nearest points belong to group A and the third point alone belongs to group B, the target point will be classified on the basis that it belongs to group A.

We used the Minkowski distance to calculate the distance between each data point and the kneighbors; equation 4.2 shows the Minkowski distance

$$the \left(\sum_{i=1}^{n} | x_i - y_i |^p \right) 1/p$$

While D is Minkowski distance, n is the number of dimensions, x_i ,

y_I are the kth attributes respectively.

The output of the KNN classification algorithm is a class membership, good or default credit,with k =1.

Advantages of KNN

- Does not involve a training process KNN algorithm does not have a training process for the data, the data itself is a model that will be a reference for future prediction and for this reason it is very time efficient in terms of improvisation for stochastic modeling on the available data.
- Easy to implement KNN algorithm is very easy to implement as the only thing to calculate is the distance between different points based on different feature data and this distance can easily be calculated using either Euclidean or Manhattan distance function.
- New data can be added at any time as it will not affect the model.

Disadvantages of KNN

- This algorithm is computationally expensive.
- Variables must be normalized; otherwise, higher range variables can be biased to a class.
- It does not work well with a large data set because calculating the distances between each instance of the data would be very expensive.
- It does not work well with high dimensionality because this will complicate the calculation of the distance for each dimension.
- Sensitive to anomalies and missing data.

3.4.1.8 Support Vector Machine (SVM)

The SVM algorithm is a supervised ML algorithm that can be used in Classification or Regression problems. However, they are mostly used in matters of classification. In the SVM algorithm, "we plot each data element as a point in n dimension space (where n is the number of Features) with the value of each attribute being a given coordinate value" [8]. Then, we perform the classification by finding the hyper-plane that distinguishes the two classes well. Look at the figure below to understand what was indicated [8].

A kernel method is a tool that works on converting low-dimensional spaces and transforming them into more-dimensional spaces - converting two-dimensional levels into three- or four-dimensional levels - that is, it transforms the questions that cannot be separated into separable problems [58]. This method is mostly useful for non-linear classification problems. Simply put, it does some very complex data transformation, and then detects the process of classifying the data based on the labels or output you specified. There are many types of SVM kernels such as linear, polynomial, and Gaussian [58].

In this study, we used the linear kernel to find a function that gives the optimal results for the regression challenge as an equation.

$$f(x) = \omega . \emptyset(x) + b$$

Advantages for SMV:

- Works well with a clear separation margin.
- Effective in high-dimensional spaces.
- Effective in a multi-dimensional data.
- It uses a subset of the training points in the decision function (called support vectors), so it is also memory efficient.

Disadvantages for SMV:

- It does not work well when we have a large data set because it takes a lot of training time.
- It does not work very well when the data set has many errors and impurities, that is, the target groups are overlapping.
- SVM does not provide probability estimates directly, rather they are computed "using expensive cross-validation. It is included in the related SVC method in the Python Scikit-Learn library" [8].

3.4.2 Collaborative Filtering Recommender Systems Item Based

Collaborative filtering is one of the most effective RMS techniques similarities between different items in a data set are calculated using one of the similarity metrics, and these similarity values are used to predict the ratings of pairs of user items that are not in the data set, This algorithm is based on the principle of searching in the set of items $\{i_1,$

$i_2, i_3, \ldots, i_k\}$ allocated in the prepared matrix (collect user information) that have been built and labeled according to the historical record of defaulted customer recommendations produced by the bank and select *k* most similar items, calculating their similarity $\{s_{i1}, s_{i2}, s_{i3}, \ldots, s_{ik}\}$ with the other item by computing the cosine similarity between these vectors. and choosing the most similar one, Once the most similar item is deduced, the prediction is calculated by taking a weighted average of the target customer's on these similar items, to calculate the similarity and predict the most accurate recommendation for the customer [57].

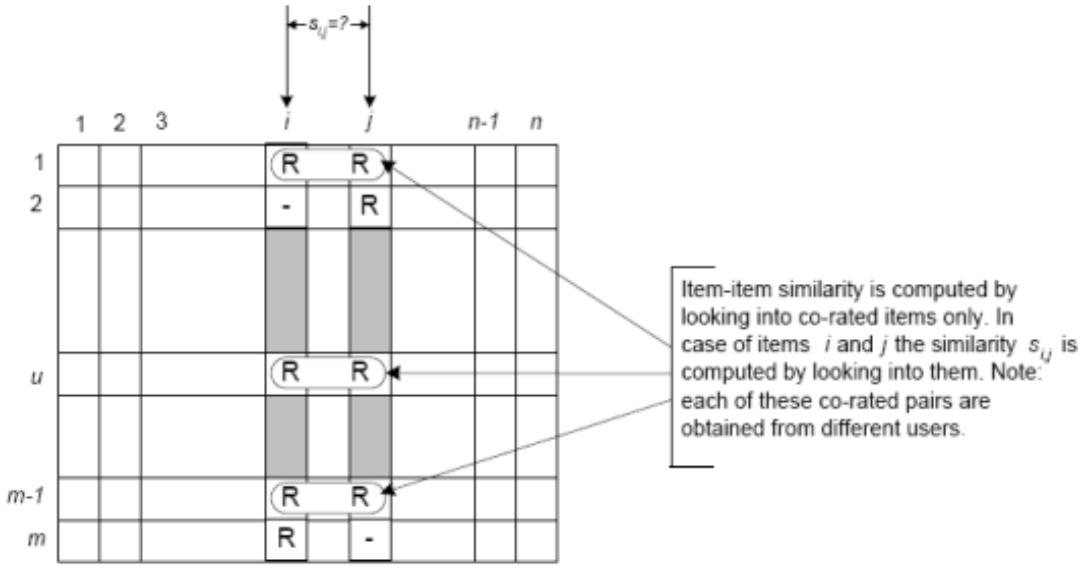


Figure 3.15 Item-Based Similarity Calculation.

In this thesis, we use item-based collaborative filtering based on the advantages we have found according to the previous literature. The collaborative filtering process is divided into two steps: first, analyze the user-item matrix to calculate the similarity, and make high-reliability recommendations.

3.4.2.1 User-Item Matrix

To build the recommendation model, we relied on the recommendation resulting from the credit decision of the bank's defaulting customer database and calculated the similarity by comparing it with this data.

The recommendations were divided as follows:

- Credit score is too low
- Debt-to-income Ratio was too high
- The income is insufficient or unstable
- Insufficient collateral
- Age does not match the type of loan
- Unpaid Dues
- Other Debts
- Address on Defaulter List
- Previously Defaulted
- Didn't meet the basic requirements

3.4.2.2 Cosine Similarity Measure

We selected cosine similarity measure item-based collaborative recommendation algorithms to produce predictions of the recommendation needed for a customer to avoid credit default.

"The cosine similarity uses cos(θ) to measure the distance between two vectors. As θ increases, cos(θ) decreases (cos(θ) = 1 when θ = 0 and cos(θ) = 0 when θ = 90). Therefore,

as the value of θ is smaller, the two vectors are considered closer (the similarity gets greater)" [59] as shown in the above equation.

$$\text{Sim (i, j)} = \frac{\sum_{u \in U}(R_{u,i} - R_u)\ \sum_{u \in U}(R_{u,i} - R_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - R_u)^2}\ \sqrt{\sum_{u \in U}(R_{u,i} - R_u)^2}}$$

The above figure illustrates the collaborative filtering workflow:

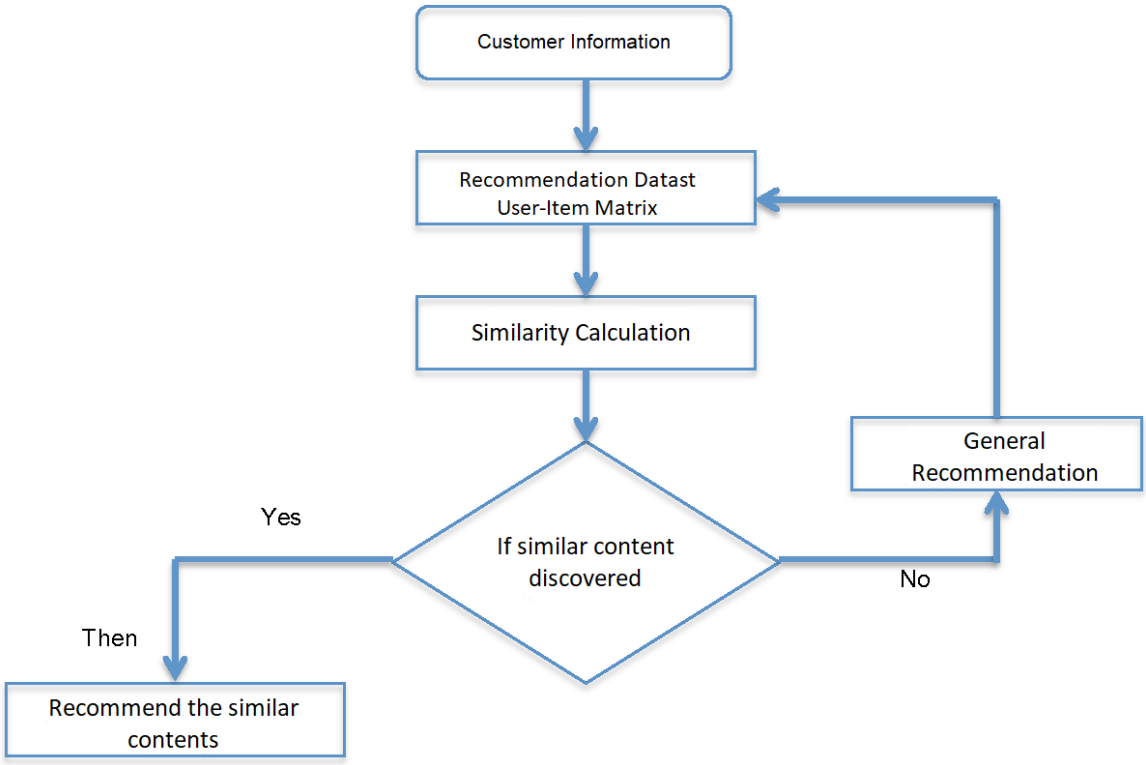


Figure 3.16 Collaborative Filtering Item-Based Workflow.

3.5 Model Evaluation

Metrics are used to measure the quality of ML model. The evaluation of ML models is essential to measure the performance ability of models to predict state. To choose the appropriate model, these models must be subjected to a test of evaluation metrics, which require testing the classification accuracy, confusion matrix, logarithmic loss, and other various evaluation metrics.

Classification accuracy is considered as the proportion of correct predictions out of the total number of classified probabilities or samples. Logarithmic loss is based on the principle of penalizing wrong classifications which is called log loss. Confusion matrix is known as probability model or error matrix. It is a specific matrix to provide a visual impact of the algorithm's performance, usually supervising the learning of the model which in turn gives us a complete description of the model's performance. [58].

It is very important to use multiple metrics to evaluate the model. This is because a model may perform well using certain metrics but may perform poorly using another of the other metrics [60].

In this thesis we will focus on the following metrics for classification: confusion matrix, true positive (TP), false positive (FP), false negative (FN), true negative (TN), accuracy, sensitivity (recall), specificity, and precision. For regression, the following were used: mean absolute error, accuracy (score coefficient), root mean squared error and mean squared error.

TP: the samples number was correctly classified as good credit.

FP: the samples number was incorrectly classified as good credit.

TN: the samples number was correctly classified as default credit.

FN: the samples number was incorrectly classified as default credit.

3.6 Accuracy

The accuracy of a classification algorithm in ML is one way to measure how often the algorithm correctly classifies a data point. Precision is “the number of correctly predicted data points out of all data points” [61]. The following is the precision equation.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

The above accuracy equation represents the number of samples that the model classified as positive (Default) which is positive in addition to the number of samples that the model classified as negative (non-default) which is negative divided by the number of samples that are true positive and true negative in addition to the false positive samples and the false negative samples as shown above. Correct negative samples or correct positive samples are a data point that the algorithm correctly categorized as true or false, respectively. On the other hand, a false positive or negative result is "a data point that the algorithm incorrectly classified" [62]. For example, if the algorithm classifies a false data point as true, it will be a false positive. Often, accuracy is used in conjunction with Precision and Recall, which are other metrics that use different ratios of true/false positives/negatives. "Together, these metrics provide a detailed look at how the algorithm ranks data points" [44].

3.7 Confusion Matrix

This matrix is called an error matrix or (Confusion Matrix): It is a well-known "matrix in the field of ML, used to test the performance of algorithms" [63]. It is a table containing information and details about the actual classes (classified by humans) and the predictive classes predicted by the classifier. Each column in the array represents the predicted class and each row represents the actual class. "The performance of the classifier is usually evaluated using the data contained in the matrix, and the size of the matrix is according to the number of categories" [63]. The following table shows the shape of the matrix and the meaning of each column and each cell in the matrix.

Table 3.2: Confusion matrix description for credit status classification

		Predicted Classes	
		Default	Non-Default
Actual Classes	Default	TP	FP
	Non-Default	FN	TN

Sensitivity or Recall: It forms the correct expectation ratio of all expected features. It is calculated using the following equation:

$$Recall\ (Sensitivity\ or\ TRP) = \frac{TP}{TP+FN}$$

F-measuring (F1 Score): it measures the balance between precision and sensitivity (recall). The F-measure is calculated using the equation:

$$f1\ score = \frac{2*Precision*Recall}{Precision+Recall}$$

Precision: The percentage of records correctly predicted as default to all records predicted in the class, and it can be described as shown in the equation

$$Precision = \frac{TP}{TP+FP}$$

Mean Absolute Error (MAE): "It is a measure of the errors between paired observations that express the same phenomenon" and it's used in regression models. "MAE is calculated as the sum of absolute errors divided by the sample size"

$$MAE = \frac{\sum_{i}^{n}=1\ |yi-xi|}{n} = \frac{\sum_{i}^{n}=1\ |e_i|}{n}$$

Mean Squared Error (MSE): is the mean of the squared errors that tells how the set of points closed from regression lines which is defined in equation

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - y\hat{}_i)$$

Root Mean Squared Error (RMSSE): is the square root of the mean of the squared errors, which is defined in the following equation where O_i are the observations, S_i predicted values of a variable, and n the number of observations available for analysis

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{i=1}^{n}(S_i - O_i)}$$

3.8 Conclusion

This chapter summarizes the methodology of credit default prediction along with a RMS based on collaborative filtering algorithm; the first step of the methodology is data preparation and exploration, the second step is data preprocessing for ML; the third step is to used different ML algorithms for predicting default accounts and then to give the recommendation to avoid default occur, These models were chosen as a result of the good performance they achieved along with the primary data processing methods that will help in the ability of these models to achieve the required with high to very high accuracy as will be described in detail in the next chapter. and finally use different performance metrics to compare between different ML algorithms performance and select the best approach.

Chapter 4

4 Results

4.1 Introduction

After the classification and regression models were deployed, we used the mentioned metrics to evaluate the efficiency of the models with different classification and regression algorithms. In this chapter, we will adjust the model variables used and compare the results in each tuning stage. I will illustrate the classification results using accuracy, confusion matrix, accuracy, recall, F1-score, for each prediction of the probability of a bank customer default, i will plot the performance of each step of hyperparameter tuning to show the best performance for each algorithm, also we will illustrate the accuracy for RMS.

The issue of evaluating ML models, calculating a set of performance evaluation measures that demonstrate that the learned model has reached a good degree of learning so that it can be relied upon and used in a problem. Ultimately, these metrics allow a comparison of the various possible learning models (each generated using a particular algorithm), allowing us to select the most appropriate one.

4.2 Classification Results

4.2.1 Confusion Matrix

The confusion matrix gives us an output matrix and describes the full performance of the model as it shows a summary of the test data for samples that were correctly and incorrectly predicted which constitutes 30% of the data set as per the controls developed in the model building step., the following figures shows the optimal confusion matrix for each model,

these results have been reached by adjusting the models' variables to achieve optimal performance.

Logistic Regression:-

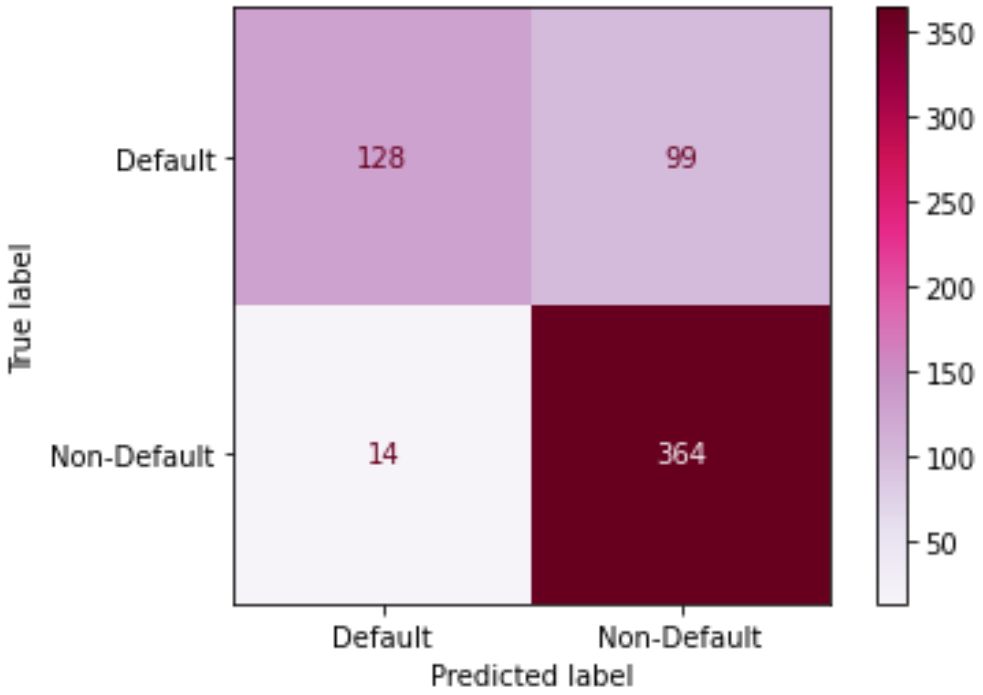


Figure 4.1 Logistic Regression Confusion Matrix

In figure 4.1 a confusion matrix for LR performance shows that 128 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 364 customers were correctly classified as default credit (TN), we also note that 99 credit customers were incorrectly classified as good credit (FP), also 14 credit customers were incorrectly classified as default credit (FN).

Decision Tree:-

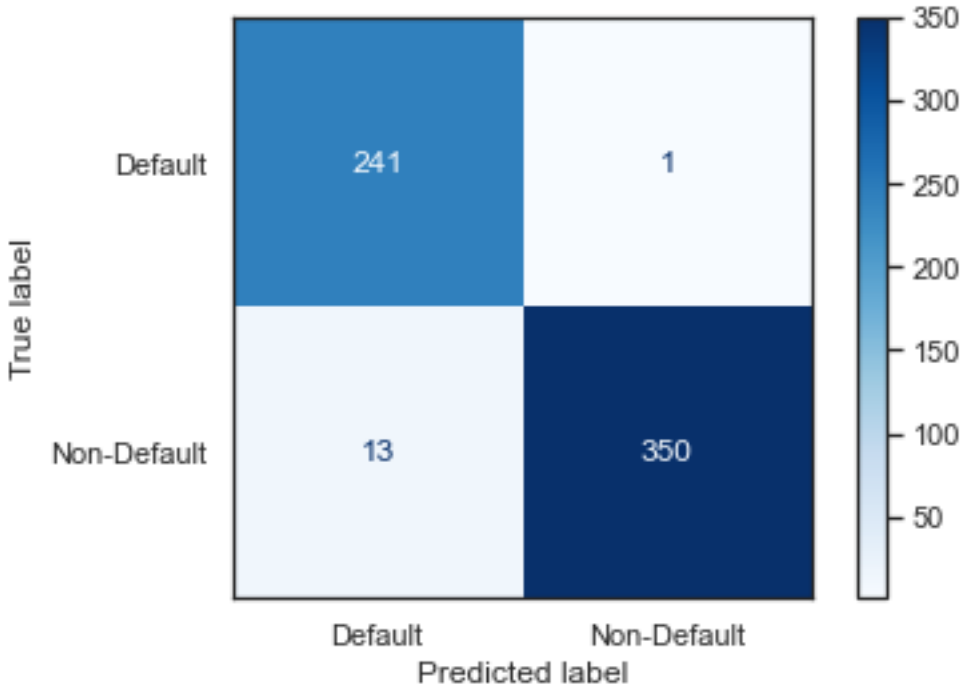


Figure 4.2 Decision Tree Confusion Matrix

Figure 4.2 a confusion matrix for DT performance shows that 241 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 350 customers were correctly classified as default credit (TN), we also note that 1 credit customers were incorrectly classified as good credit (FP), also 13 credit customers were incorrectly classified as default credit (FN).

Random Forest :-

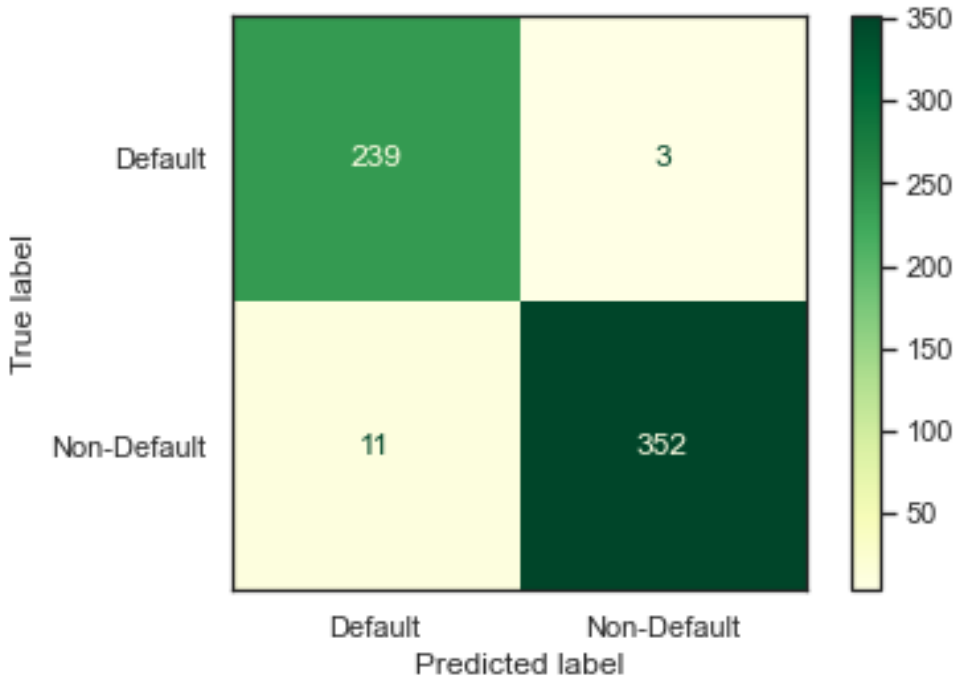


Figure 4.3 Random Forest Confusion Matrix

Figure 4.3 a confusion matrix for RF performance shows that 239 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 352 customers were correctly classified as default credit (TN), we also note that 3 credit customers were incorrectly classified as good credit (FP), also 11 credit customers were incorrectly classified as default credit (FN).

Adaptive Boosting (AdaBoost) :-

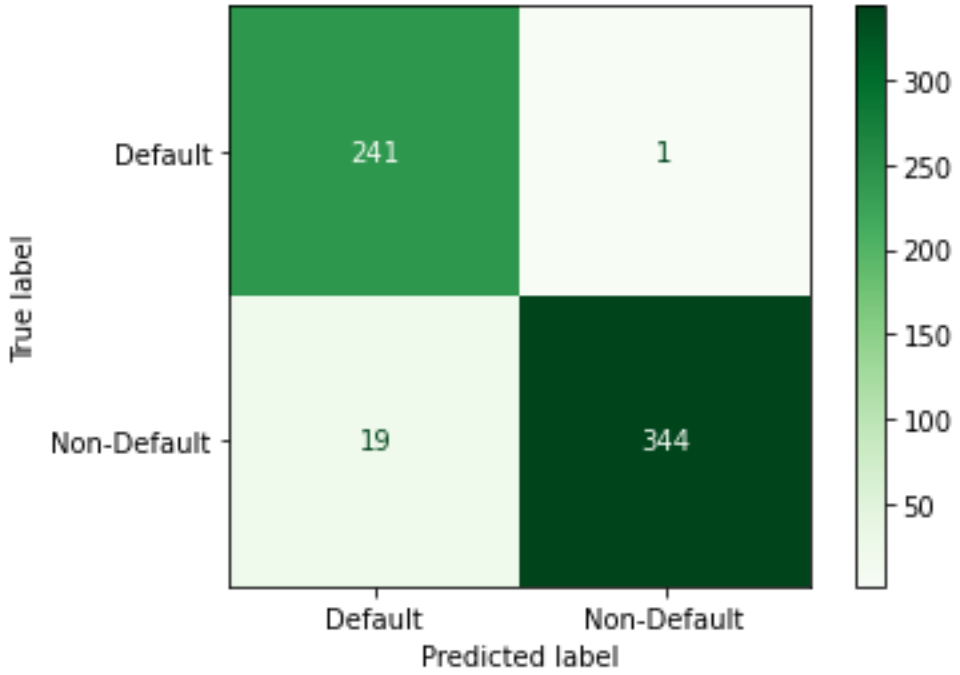


Figure 4.4 AdaBoost Confusion Matrix

Figure 4.4 a confusion matrix for AdaBoost performance shows that 241 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 344 customers were correctly classified as default credit (TN), we also note that 1 credit customer was incorrectly classified as good credit (FP), also 19 credit customers were incorrectly classified as default credit (FN).

eXtreme Gradient Boosting (XGBoost):-

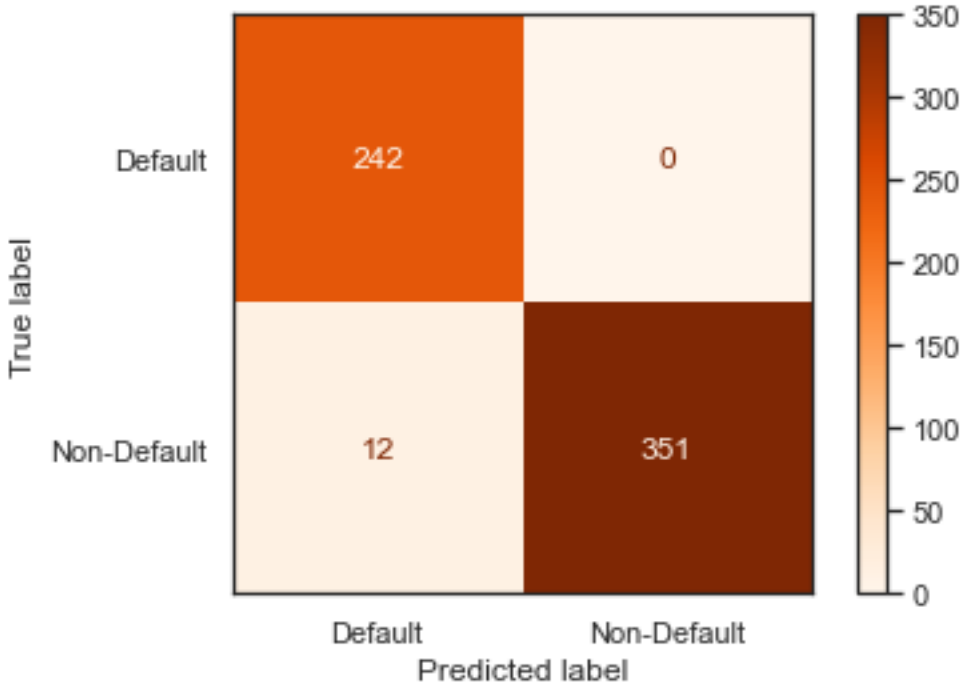


Figure 4.5 XGBoost Confusion Matrix

Figure 4.5 a confusion matrix for AdaBoost performance shows that 242 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 351 customers were correctly classified as default credit (TN), we also note that no credit customers were incorrectly classified as good credit (FP), where 12 credit customers were incorrectly classified as default credit (FN).

K-nearest neighbors (KNN):-

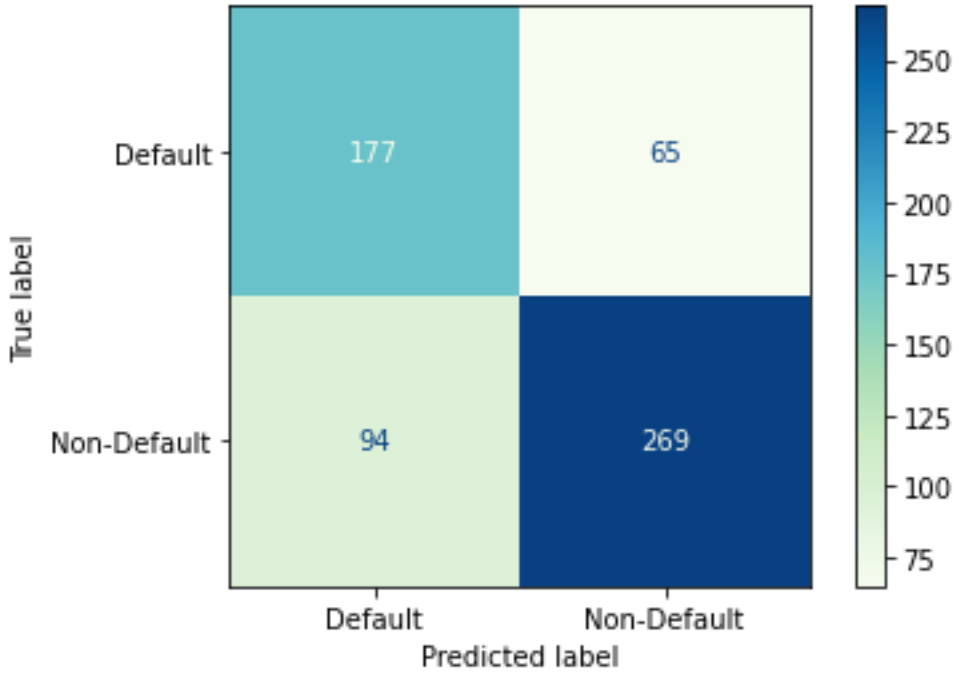


Figure 4.6 KNN Confusion Matrix

Figure 4.6 a confusion matrix for AdaBoost performance shows that 177 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 269 customers were correctly classified as default credit (TN), we also note that 65 credit customers were incorrectly classified as good credit (FP), also 94 credit customers were incorrectly classified as default credit (FN).

Support Vector Machine (SVM):-

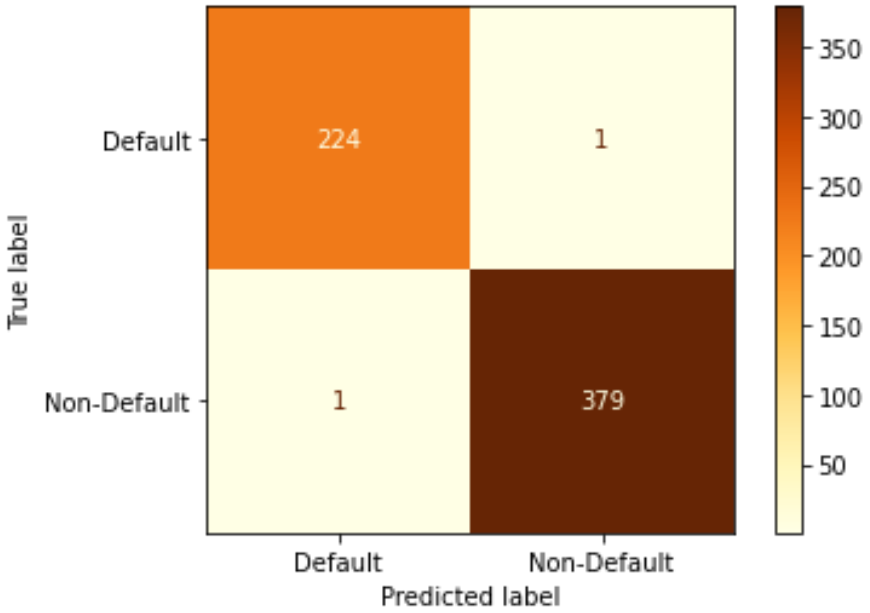


Figure 4.7 SVM Confusion Matrix

Figure 4.7 a confusion matrix for SVM performance shows that 224 customers from the testing dataset of bank credit sample were correctly classified as good credit (TP) whereas 379 customers were correctly classified as default credit (TN), we also note that 1 credit customer was incorrectly classified as good credit (FP), also 1 credit customers were incorrectly classified as default credit (FN).

4.2.2 Accuracy

In this part of the study, we will review the calculation of accuracy measures for each model (ratio of correct prediction from the correct number of the testing set) by evaluating learning models using the **Cross Validation** method along with a **grid search** method which was discussed in methodology.

The accuracy for the Decision Tree Classifier was tested using a grid search to optimize model parameters to reach the best accuracy, best accuracy was reached by using the

entropy attribute measure to measure the impurity or uncertainty in a group of observations, also the maximum depth of tree was set to be equal 5.

4.2.2.1 Decision Tree Classifier

The accuracy for the Decision Tree Classifier was tested using a grid search method to optimize model parameters to reach the best model performance, best accuracy was reached by using the entropy attribute criterion to measure the impurity or uncertainty in a group of observations, also the maximum depth of tree was set to be equal 5.

DT model has been trained using 5 cross-validations, along with the grid search method, the accuracy for each classification algorithm with different parameters was tested by total fittings of 450 fits to reach the optimal algorithm parameters.

DT Optimal algorithm parameter by Grid Search method:-

ccp_alpa = 0.1, criterion = entropy, max_depth = 5, min_samples_split = 2, min_samples_leaf=1, random_state=42.

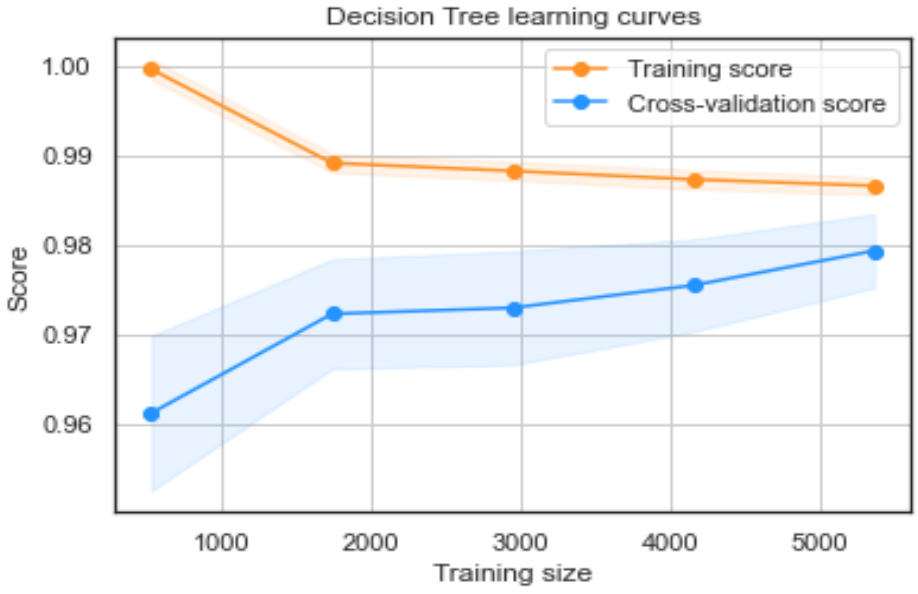


Figure 4.8 Decision Tree Cross-validation Accuracy Score.

Figure number 4.8 show the decision tree learning curve, during the training and testing process, each point represents the accuracy of the evaluation stage.

4.2.2.2 Random Forest

The techniques used for optimizing Random Forest include, hyperparameter tuning, cross validations, and averaging multiple classifiers to get better accuracy, the best performance of RF is reached with a number of trees = 300.

RF model has been trained using 5 cross-validations, along with the hyperparameter tuning (grid search method), the accuracy for each classification algorithm with different parameters was tested by total fittings of 1080 fits to reach the optimal algorithm parameters.

RF Optimal algorithm Parameter by Grid Search Method:-

bootstrap=False, max_features=3, n_estimators=300.

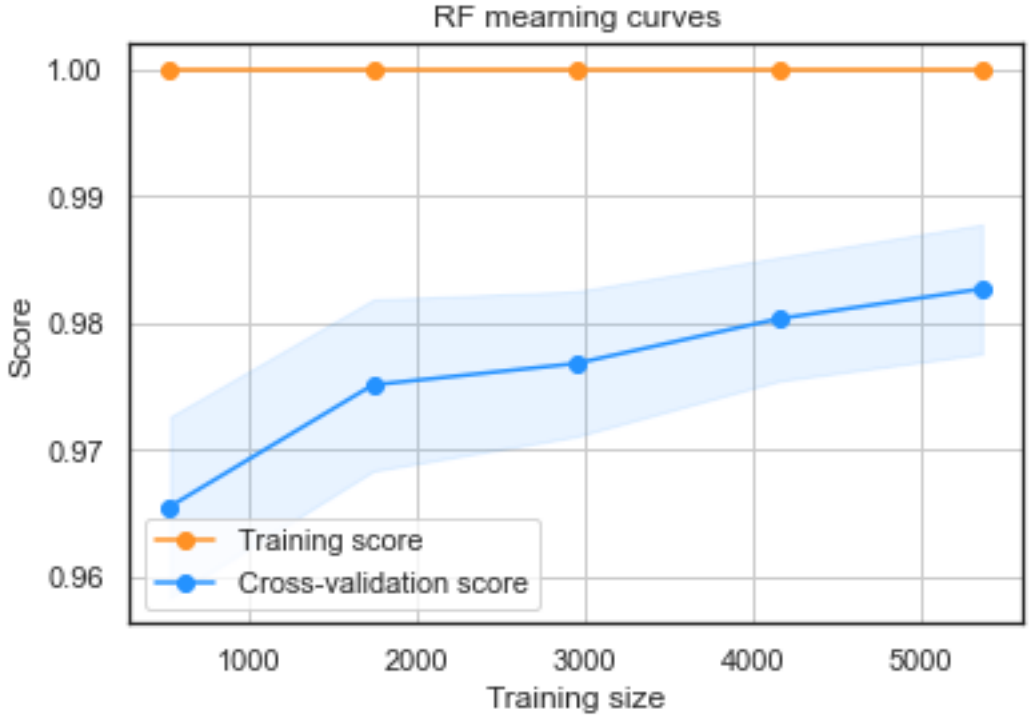


Figure 4.9 Random Forrest Cross Validation Accuracy Score

Figure number 4.9 show the RF learning curve, during training and testing process, each point represents the accuracy of the evaluation stage.

4.2.2.3 AdaBoost Classifier

AdaBoost is a type of decision tree, distinguished by the learning rate that controls the contribution of each model to the ensemble prediction. AdaBoost Accuracy was tested by tuning the hyperparameter, the best accuracy was gained using learning rate of 0.0001, tree max depth = 5, and entropy base estimator.

AdaBoost model has been trained using 10 cross-validations, along with the hyperparameter tuning (grid search method), the accuracy for each classification algorithm with different parameters was tested by total fittings of 1120 fits to reach the optimal algorithm parameters.

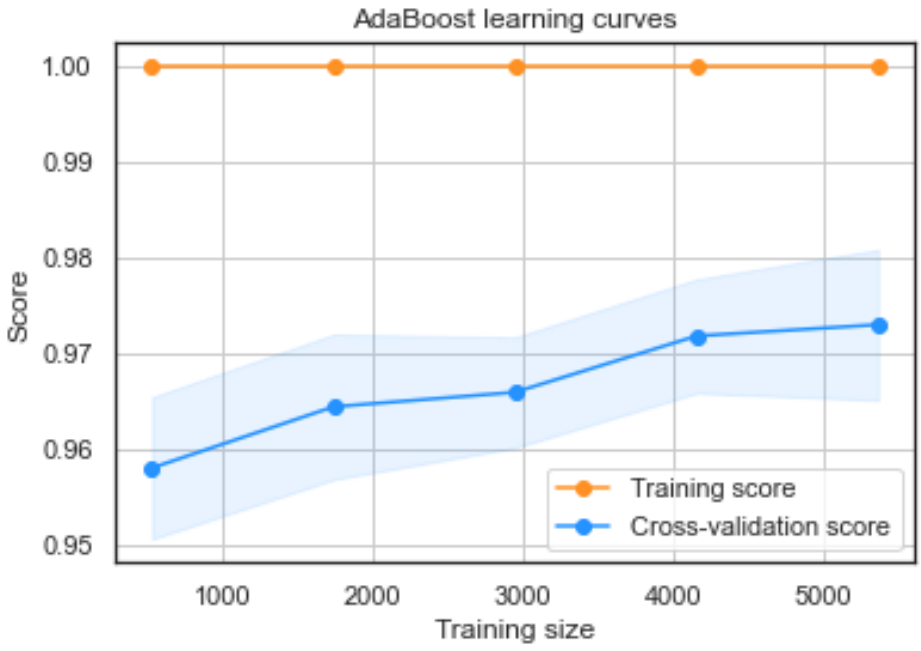


Figure 4.10 AdaBoost Cross-Validation Accuracy Score

Figure number 4.10 show the decision tree learning curve, during the training and testing process, each point represents the accuracy of the evaluation stage.

4.2.2.4 XGBoost Classifier

XGBoost Accuracy was tested by grid search method to fit the model with the most optimal parameter that gained the highest accuracy, by using hyperparameter optimization we reach to use deviance loss function, gradient descent optimization algorithm, learning rate = 0.1, maximum depth = 2, min sample leaf = 100 and n_estimatores = 100.

XGBBoost model has been trained using 10 cross-validations, along with the hyperparameter tuning (grid search method), the accuracy for each classification algorithm with different parameters was tested by total fittings of 1080 fits to reach the optimal algorithm parameters.

Figure number 4.11 show the decision tree learning curve, during the training and testing process, each point represents the accuracy of the evaluation stage.

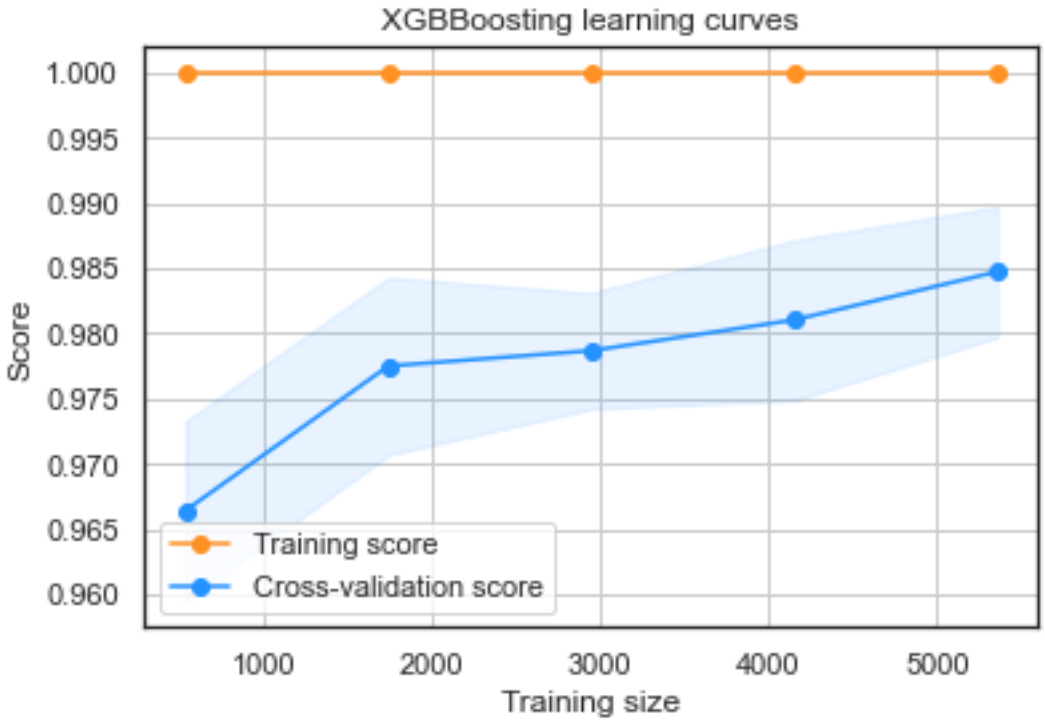


Figure 4.11 XGBoost Cross-Validation Accuracy Score

Figure number 4.11 show the XGBoost learning curve, during the training and testing process, each point represents the accuracy of the evaluation stage.

4.2.2.5 KNN Classifier

KNN model has been trained using 5 cross-validations, by using the grid search the accuracy for each classification algorithm with different parameters was tested by total fittings of 128 fits to reach the optimal algorithm parameters, the highest accuracy has been reached by setting n_neighbors = 8, p = 1, Distance metric = 'Minkowski', leaf_size = 1.

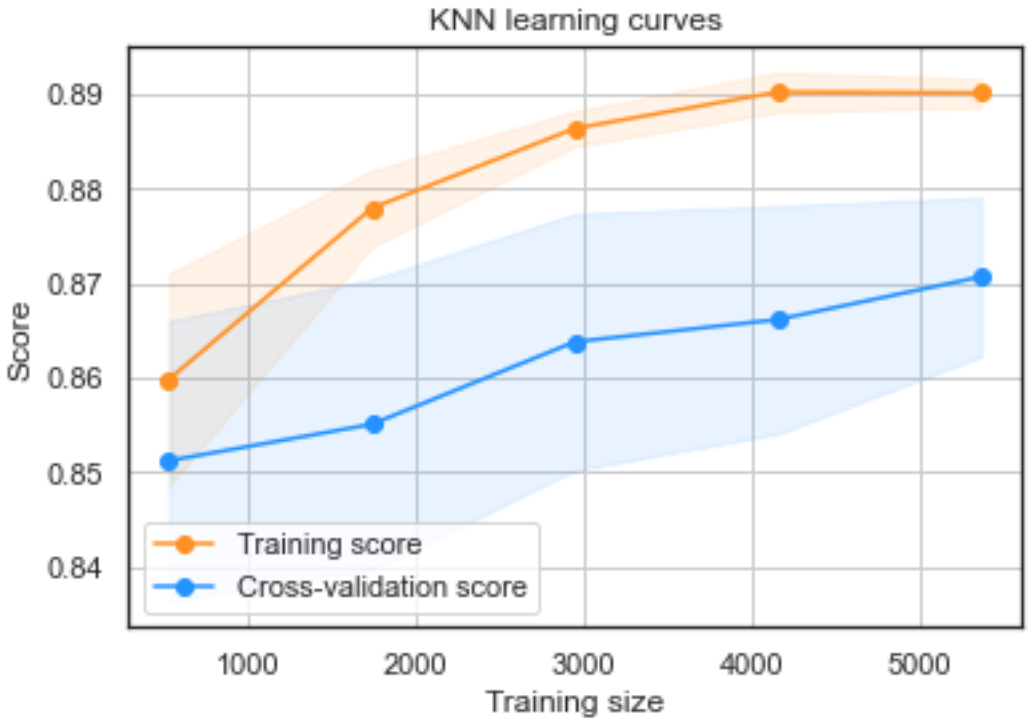


Figure 4.12 KNN Cross-Validation Accuracy Score

Figure number 4.12 show the KNN learning curve, during the training and testing process, each point represents the accuracy of the evaluation stage.

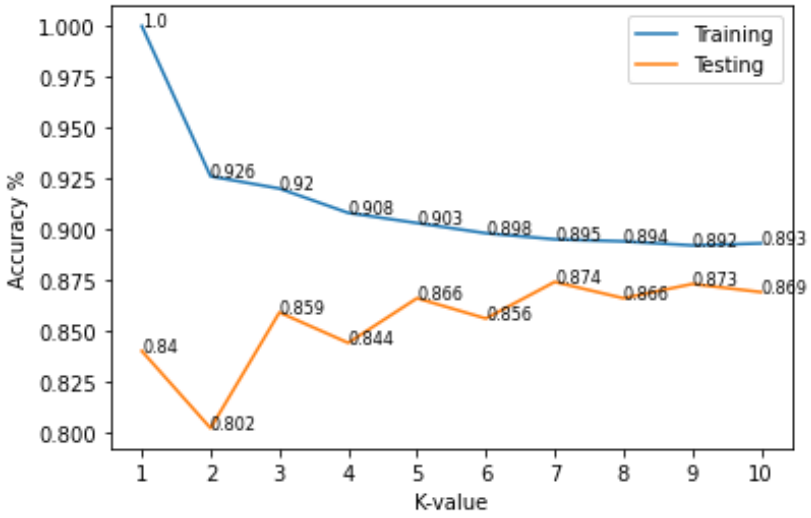


Figure 4.13: Accuracy Measures Using Different K Numbers and Different Distance Metrics Algorithms

Figure 4.13 show the results for the KNN algorithm while using four different distance metrics algorithms which are Euclidean, Manhattan, and Minkowski, the results also show the accuracy using a different number of K neighbors. The best accuracy results were gained using the Minkowski algorithm and while using 7 K.

4.2.2.6 SVM Classifier

SVM Accuracy was tested thought using 10 cross-validations, along with the hyperparameter tuning (grid search method), the accuracy for each classification algorithm with different parameters was tested by total fittings of 1080 fits to reach the optimal algorithm parameters.

Hyperparameter tuning helped to improve the performance of the model, as the most influential variables on the prediction accuracy were adjusted by setting regularization parameter (C) to be = 300, gamma = 0.10, and kernel = ‘RBF.

Figure 4.14 SVM Cross-Validation Accuracy Score

Gamma is a parameter that comes with the RBF kernel. It implicitly defines how far the influence of a single training example reaches. The larger gamma value, the fewer support vectors, and the smaller the gamma value, the higher the support vectors. The number of support vectors affects training speed and prediction. The specific effect is shown in figure 4.15 where the figure shows that the best accuracy of the model's performance in predicting the probability of the client's default is when the gamma value is equal to 0.10, where we notice a decrease in the accuracy of the model at a low gamma coefficient.

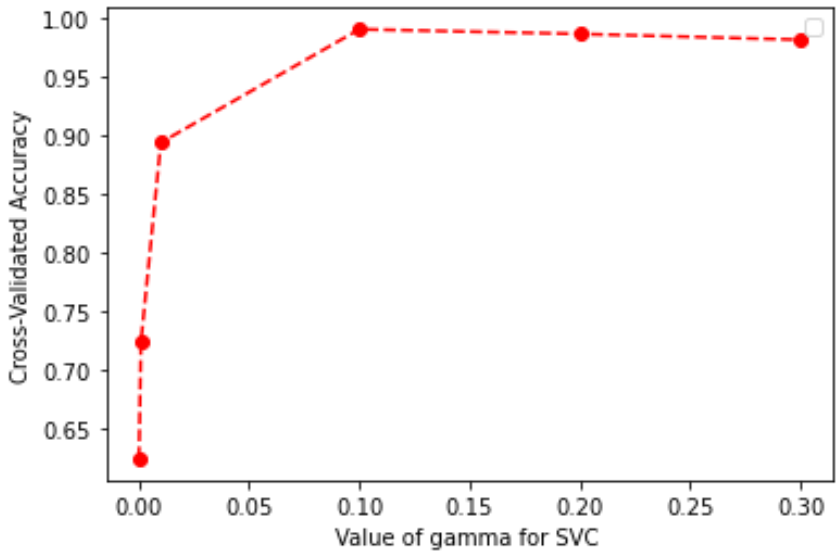


Figure 4.15 Accuracy Measures Using Different Gamma Values For SVM.

4.2.2.6 Logistic Regression

LR is one of the algorithms that rely heavily on the quality of the data entered by the learner, to ensure that the model reaches the greatest ability to predict the probability of the customer default, we calculated information value (IV) that's it is very useful in the preliminary filtering of variables before including them in LR model. The efficiency of the LR model is affected by the initial processing of the data, and an important step before the model is installed is to reduce the variables through the use of IV, IV used to extract the best information value and to exclude the variables that have little effect on the performance of ML model in terms of prediction. Table 4.1 list the information value that has been calculated for all features of the DP dataset and added together to measure the predictive power for each feature. We got the top IV values for due amount as 2.88, and due installment as 0.97 meaning it is a suspicious or too good predictor, it was also found that some variables have no effect and are considered as weak predictors to classify events

like customer classification (CUST_MGR), Credit Balance (Balance), credit open date (OPEN_DATE) where IV < 0.5.

VAR_NAME	IV
INSTALLMENT_PER	0.221444
BRA_CODE	0.224593
FAC_TYPE	0.133231
BALANCE	0.002998
ACCOUNT_OPEN_DATE	0.000443
DUE_AMT	2.888973
LOAN_AMOUNT	0.055914
TYPE_OF_DEP	0.083917
CUST_MGR	0.000611
INSTALLMENT_AMOUN	0.00859
MIN_INT_RATE	0.210799
DUE_INSTALLMENT	0.970706
DUE_DAYS	0.157156
DATE_OVDN	0.005357
SUSP_AMT	0.13795
AGE	0.161506
MONTHLY_SALARY	0.032893
RET_CHQ	0.242746

Table 4.1 DP features Information Value.

An insignificant feature has been eliminated to reach features with a level of significance less or equal to 0.05 of the significance level by calculating the P-value for each feature where the following features have been removed 'SUSP_AMT', 'DUE_DAYS', 'DATE_OVDN'.

RF model has been trained using 5 cross-validations, along with the hyperparameter tuning (grid search method), the accuracy for each classification algorithm with different parameters was tested by total fittings of 4800 fits to reach the optimal algorithm parameters.

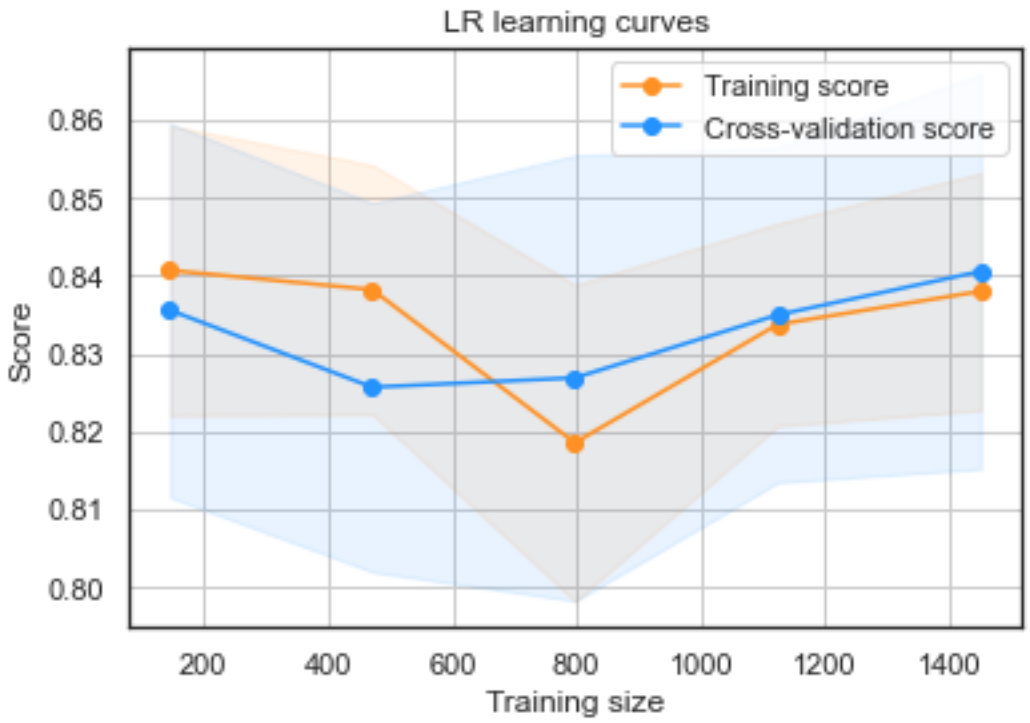


Figure 4.16 Logistic Regression Cross-Validation Accuracy Score.

Hyperparameter tuning helped to improve the performance of the model, as the most influential variables on the prediction accuracy were adjusted by setting regularization parameter (C) to be = 1.4, Figure number 4.17 shows the best score of grid search evaluation accuracy for training and testing LR model.

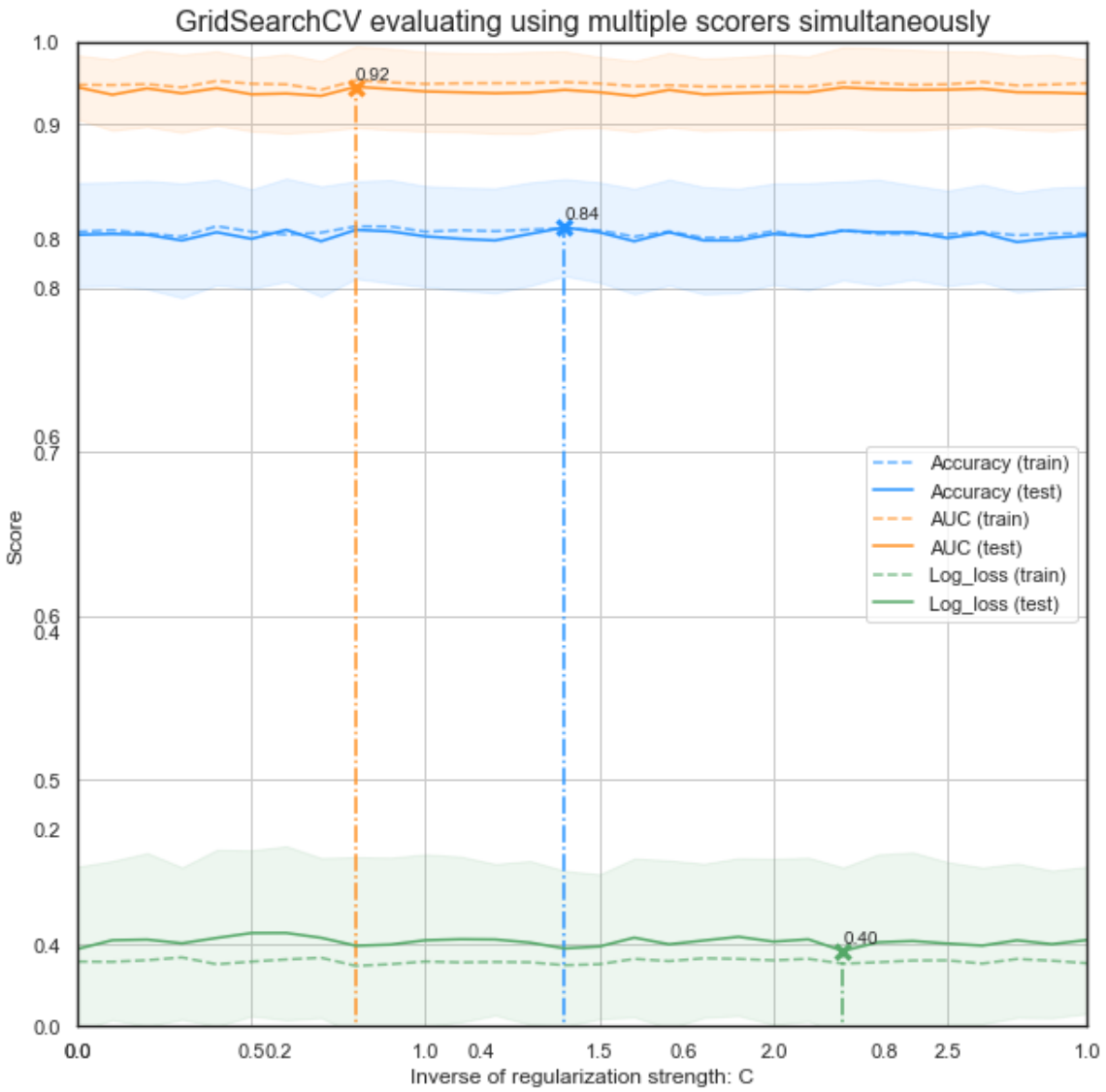


Figure 4.17 LR Grid Search Evaluation Using Multiple Scores Simultaneously.

4.2.3 Models Comparison and Overall Performance

The AUC-ROC curve is a performance measure for classification problems under different threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It tells us how many models can be categorized. The higher the AUC value, the better the model, as well as the ability of the model to distinguish customers with high default risk from good customers., A good model has an AUC value close to 1,

which means that it has good separability. The AUC of a bad model is close to 0, which means it has the worst measure of separation. When the AUC is 0.5, it means that the model has no ability to separate bad credit from good for customers.

Based on the above analysis, the AUC score is used as an evaluation criterion for the prediction PD model in this study. In addition, to accurately assess the probability of credit default, some assessment metrics of the ML algorithm such as SPE, TRN, SEN, ROC, and AUC are included to calculate the final prediction result as it's shown in Figure number 4.18.

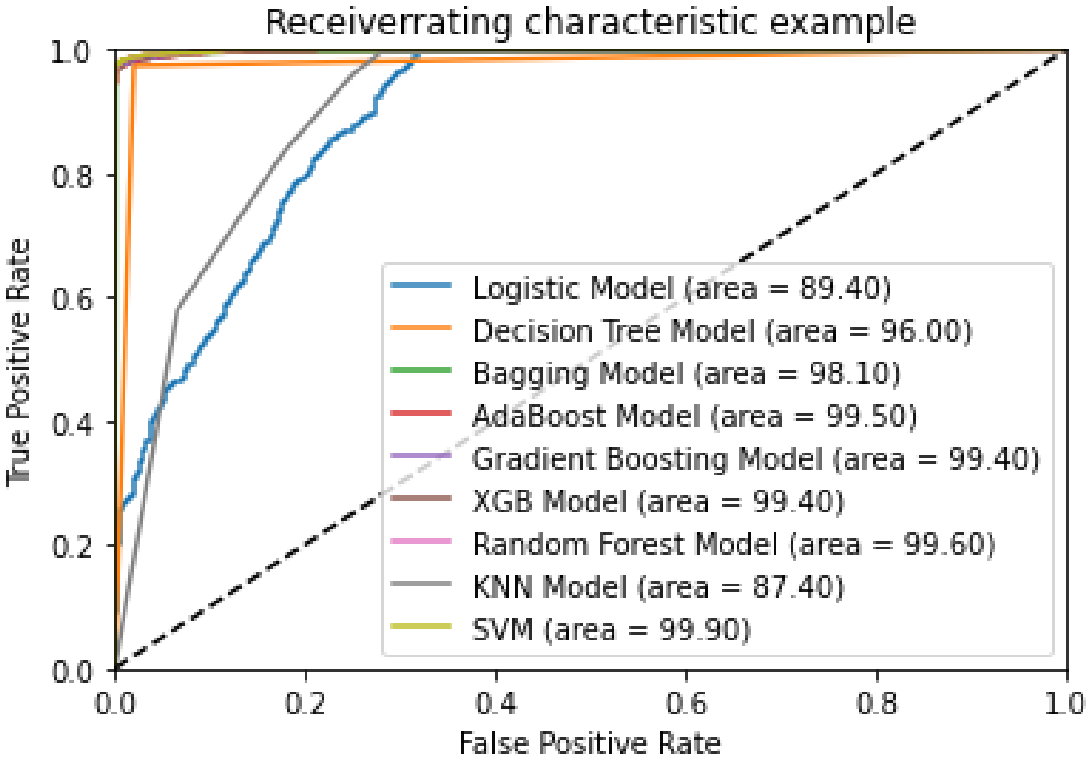


Figure 4.18 ROC Curve Evaluation

Table No. 4.2 Summarizes the Performance Results of the ML Models Used in this Thesis in Terms of Each Model Performance Results, The Results are Shown in Terms of AUC, Precision, Recall, and F1 Score.

Model	AUC	Precision	Recall	F1 Score
LR	89.4%	84.50%	88%	87%
DT	96.0%	97%	98%	97%
RF	99.6%	97%	98%	98%
Bagging	98.1%	96%	97%	97%
AdaBoost	99.5%	97%	98%	97%
Gradient Boost	99.4%	97%	98%	97%
XGBoost	99.4%	97%	98%	98%
KNN	87.4%	87%	88%	87%
SVM	99.9%	99%	99%	99%

Table 4.2 Default Prediction Model Performance Evaluation.

4.3 General Discussion

By digging deeper into the previous literature, many studies [2] [1] [64] [14] [31] [10] dealt directly with the research of credit risks and the provision of automated solutions that would maintain a credit portfolio free from default, by predicting default cases before they occur using ML, Some studies were confined to a limited type of credit accounts or to a specific segment of customers such as individuals, corporate, and SME's, which raised caution by finding solutions to provide a prediction for various types of credit accounts in one solution. Likewise, I concluded by reviewing the literature that there is an urgent need to provide a guiding recommendation to improve the credit status of the customer with positive indicators that predict the possibility of customer default. This goal was achieved by building an early recommendation system which enhances the ability of this study to complement previous studies and fill this gap.

In this section, I will reflect on the overall results and evaluation results for the ML algorithm, several ML algorithms have been applied to predict the probability of customer

default, and the results of these models were examined, as the preliminary results were somewhat satisfactory, but the challenges we faced were how to improve the performance of ML models and choose the most appropriate ones. Several methods were discussed in the field of data science, which we learned within the course plan at the Arab American University. One of the most important of these methods is data pre-processing, which greatly affected the results of ML models in terms of performance and accuracy in predicting customer default, and the results were amazing in greatly increasing the accuracy of prediction. Among the most important stages of data pre-processing that were used for processing are data normalization, data scaling, and feature selection.

We did not stop at this point. We also implemented parameter adjustment processes by in-depth and studying the extent to which each parameter affects the performance of the models by using data visualization to display a graph of the best set of parameters that have obtained the highest prediction accuracy. This was done using Grid Search technology. We conclude through this chapter that the algorithms of ML have a great ability to predict, but relying on it alone is not enough. Enhancing the efficiency of model performance requires strengthening data pre-processing processes, especially if there is a lot of irrelevant or redundant information or distorted and unreliable data. This helps in discovering knowledge during the training phase and raises the efficiency and performance of the models in predicting with high accuracy, In addition, the features that were used for this study had a great ability to indicate and predict the probability of customer default with high accuracy, the most important of these features are the amounts owed by the customer, the number of rejected checks, monthly income, installments amount and the granted credit amount.

The ML models used performed well in terms of prediction accuracy, and the overall prediction accuracy of SVM is 99.90%, which indicates that the SVM model is superior to other models in terms of accuracy of prediction of the probability of customer default as well as achieving the best performance in terms of error. This indicates that all the methods used show the potential for accurately predicting the default risk of borrowers. To go deeper into testing the performance of different ML models, the accuracy, recall, and F1 values of the five models were calculated and summarized. Most of the models achieved over 94% performance for each metric. This also shows that all models reported in this thesis have good capabilities for customer default prediction. The SVM model has the best performance, which corresponds to the highest accuracy over all other models that helps the banking sector to rely on machine learning models to maintain a clean credit portfolio free from default.

at the end, this study presented a comprehensive solution for predicting the customer's failure to pay his indebtedness and early detection of default cases, in addition to an automated recommendation system that allows the credit-control departments to correct the customer's credit status to maintain good credit.

This study contributes to strengthening the previous literature in terms of researching a unique data set, achieving high prediction accuracy, and providing an integrated system for predicting and recommending immediate treatment of the customer's credit situation.

Chapter 5

5 Conclusion and Future Work

Financial regulatory institutions and banks must be able to make reliable predictions about borrowers' default to control credit risks and protect their customers, as lending institutions and banks in Palestine suffered from a widespread bad economic situation in the past three years, which led to an increase in the inability of customers to pay their debts. In addition, ML approaches are powerful in improving the ability to predict risks and mitigate information asymmetries. To facilitate our research, payment delays or payment failures are arbitrarily re-encoded as "1" and "0" for good credit within the study datasets, as the classification variable was labeled by credit control experts at Jordan Ahli Bank, many algorithms were used to predict the probability of customer default including Logistic regression, KNN, SVM, Random Forest and decision trees of all kinds. The study proved that decision trees have a lower average error than other algorithms, and the SVM model proved more efficient in terms of accuracy and outperformed other models with an accuracy rate of 99.9% and the lowest rate of classification error obtained by the rest of other models.

I did not stop at this point. I also carried out this research by establishing an automatic RMS based on the prediction results extracts to provide an automatic recommendation to the credit experts at the bank, aiming to correct the customer's situation in terms of the reasons that have been achieved or that may be achieved and that affect the customer's default on payment. The RMS presents the reasons that prevent the customer from being able to repay and the need to adhere to them, which would calculate the customer's credit

position in terms of the degree of credit risk and maintain a clean credit portfolio free of default cases.

As future work, we see building Graphical User Interface (GUI) applications and linking them with ML models and the RMS to provide an actual reality and real experience to predict the possibility of client default, by extracting the necessary periodic reports for decision makers and credit experts in banks, for early warning to discover the possibility of default and provide the necessary guardianship A preventive measure to avoid credit default accounts.

5.1 Knowledge Contributions and Implications for Practice.

Recently, the sector of financial institutions and banks is witnessing an upward growth in the use of AI applications, as it has become a natural evolution for banks, as digital banking services pave the way for more advanced technologies related to ML, data mining, and analytics that can identify patterns and irregularities in an attempt to bring about the necessary improvements in terms of all banking fields, especially credit risk analysis and detection of customer default cases before they occur.

In this regard, the field of scientific research is developing in finding appropriate banking solutions for leadership and enhancing the status of digital development and the use of AI as a major reliable part of banking services. Through a deep review of the literature, this study complements the previous literature in the field of predicting cases of bank customers defaulting using ML algorithms that are unique to an exclusive data set for this thesis, This study fills the silent gap centered in the shortcomings of the previous literature by providing early banking solutions to avoid default cases before they occur through the smart

recommendation system that provides automated solutions centered on giving the ideal recommendation to correct the customer’s credit status on which to rely on in the field of credit review and maintaining a clean credit portfolio free of default cases .

5.2 Limitation

The main objective of this study is to predict bank customer defaults before they occur and provide the necessary automatic recommendations to correct the customer's credit situation to reduce dependence on human resources to a minimum and rely on artificial intelligence capabilities in data analysis and predicting with high accuracy and reliability. However, the scope of this research faces some limitations, the most important of which is obtaining a sufficient amount of data within the datasets to enhance the ability of machine learning models to train sufficiently to ensure that the models are able to predict all credit situations. The dataset belongs to the credit control system of Jordan Ahli Bank, which is managed manually by the credit team of the bank which causes nullable and misleading data. This affected the quality of the data and the efficiency of the machine learning models in making predictions in the initial stages of model building. Many stages of preliminary data processing have been implemented to overcome this problem and increase the efficiency of machine learning models. Some challenges were represented in selecting the most appropriate machine learning models for the extracted data set. The supervised machine learning models were selected due to the presence of a labeled dataset to ensure the best accuracy and efficiency for model performance, hyperparameter tuning technique was used.

References

A. K. Tiwari, “Machine Learning Application in Loan Default Prediction,” *JournalNX - A Multidiscip. Peer Rev. J.*, vol. 4, no. 5, pp. 1–5, 2018, doi: 2581 -4230.

J. Xu, Z. Lu, and Y. Xie, “Loan default prediction of Chinese P2P market: a machine learning methodology,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–19, 2021, doi: 10.1038/s41598-021-98361-6.

F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” *Expert Syst. Appl.*, vol. 83, pp. 405–417, 2017, doi: 10.1016/j.eswa.2017.04.006.

A. Adam Ding, S. Tian, Y. Yu, and H. Guo, “A class of discrete transformation survival models with application to default probability prediction,” *J. Am. Stat. Assoc.*, vol. 107, no. 499, pp. 990–1003, 2012, doi: 10.1080/01621459.2012.682806.

G. Teles, J. J. P. C. Rodrigues, R. A. L. Rabêlo, and S. A. Kozlov, “Comparative study of support vector machines and random forests machine learning algorithms on credit operation,” *Softw. - Pract. Exp.*, vol. 51, no. 12, pp. 2492–2500, 2021, doi: 10.1002/spe.2842.

H. P. Tserng, P. Chen, W. Huang, M. Cheng, and Q. Hung, “PREDICTION OF DEFAULT PROBABILITY FOR CONSTRUCTION FIRMS USING THE LOGIT MODEL,” vol. 20, no. 2, pp. 247–255, 2014, doi: 10.3846/13923730.2013.801886.

H. Kim, H. Cho, and D. Ryu, “Corporate Default Predictions Using Machine Learning : Literature Review,” pp. 1–11, 2020.

Y. P. Huang and M. F. Yen, “A new perspective of performance comparison among machine learning algorithms for financial distress prediction,” *Appl. Soft Comput. J.*, vol. 83, p. 105663, 2019, doi: 10.1016/j.asoc.2019.105663.

E. O. Ogundimu, “Prediction of default probability by using statistical models for rare events,” *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 182, no. 4, pp. 1143–1162, 2019, doi: 10.1111/rssa.12467.

A. Motwani, P. Chaurasiya, and G. Bajaj, “Predicting Credit Worthiness of Bank Customer with Machine Learning over Cloud,” *Int. J. Comput. Sci. Eng.*, vol. 6, no. 7, pp. 1471–1477, 2018, doi: 10.26438/ijcse/v6i7.14711477.

A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis, “Predicting bank insolvencies using machine learning techniques,” *Int. J. Forecast.*, vol. 36, no. 3, pp. 1092–1113, 2020, doi: 10.1016/j.ijforecast.2019.11.005.

Y. Son, H. Byun, and J. Lee, “Nonparametric machine learning models for predicting the credit default swaps: An empirical study,” *Expert Syst. Appl.*, vol. 58, pp. 210–220, 2016, doi: 10.1016/j.eswa.2016.03.049.

A. R. Provenzano *et al.*, “Machine Learning approach for Credit Scoring,” pp. 1–28, 2020, [Online]. Available: http://arxiv.org/abs/2008.01687.

Y. Li and W. Chen, “Entropy method of constructing a combined model for improving loan default prediction: A case study in China,” *J. Oper. Res. Soc.*, vol. 72, no. 5,

pp. 1099–1109, 2021, doi: 10.1080/01605682.2019.1702905.

D. Duffie, L. Saita, and K. Wang, “Multi-period corporate default prediction with stochastic covariates,” *J. financ. econ.*, vol. 83, no. 3, pp. 635–665, 2007, doi: 10.1016/j.jfineco.2005.10.011.

S. Birla, K. Kohli, and A. Dutta, “Machine Learning on imbalanced data in Credit Risk,” *7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016*, 2016, doi: 10.1109/IEMCON.2016.7746326.

J. Xiao, M. Wang, B. Jiang, and J. Li, “A personalized recommendation system with combinational algorithm for online learning,” *J. Ambient Intell. Humaniz. Comput.*, vol. 9, no. 3, pp. 667–677, 2018, doi: 10.1007/s12652-017-0466-8.

X. Xin, Z. Liu, C. Y. Lin, H. Huang, X. Wei, and P. Guo, “Cross-domain collaborative filtering with review text,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2015-Janua, no. Ijcai, pp. 1827–1834, 2015.

C. Iwendi, E. Ibeke, H. Eggoni, S. Velagala, and G. Srivastava, “Pointer-Based Item-to-Item Collaborative Filtering Recommendation System Using a Machine Learning Model,” *Int. J. Inf. Technol. Decis. Mak.*, vol. 21, no. 1, pp. 463–484, 2022, doi: 10.1142/S0219622021500619.

M. C. Aniceto, F. Barboza, and H. Kimura, “Machine learning predictivity applied to consumer creditworthiness,” *Futur. Bus. J.*, vol. 6, no. 1, pp. 1–14, 2020, doi: 10.1186/s43093-020-00041-w.

P. Gogas, T. Papadimitriou, and A. Agrapetidou, “Forecasting bank failures and stress testing: A machine learning approach,” *Int. J. Forecast.*, vol. 34, no. 3, pp. 440–455, 2018, doi: 10.1016/j.ijforecast.2018.01.009.

J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, “Expert Systems with Applications Consumer credit risk : Individual probability estimates using machine learning,” *Expert Syst. Appl.*, vol. 40, no. 13, pp. 5125–5131, 2013, doi: 10.1016/j.eswa.2013.03.019.

S. Bhatore, L. Mohan, and Y. R. Reddy, “Machine learning techniques for credit risk evaluation: a systematic literature review,” *J. Bank. Financ. Technol.*, vol. 4, no. 1, pp. 111–138, 2020, doi: 10.1007/s42786-020-00020-3.

L. Coenen, W. Verbeke, and T. Guns, “Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods,” *J. Oper. Res. Soc.*, vol. 0, no. 0, pp. 1–16, 2020, doi: 10.1080/01605682.2020.1865847.

A. E. Khandani, A. J. Kim, and A. W. Lo, “Consumer credit-risk models via machine-learning algorithms,” *J. Bank. Financ.*, vol. 34, no. 11, pp. 2767–2787, 2010, doi: 10.1016/j.jbankfin.2010.06.001.

P. M. Addo, D. Guegan, and B. Hassani, “Credit risk analysis using machine and deep learning models,” *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

Y. Wang, Y. Zhang, Y. Lu, and X. Yu, “A Comparative Assessment of Credit Risk Model Based on Machine Learning ——a case study of bank loan data,” *Procedia Comput. Sci.*, vol. 174, pp. 141–149, 2020, doi: 10.1016/j.procs.2020.06.069.

J. R. de Castro Vieira, F. Barboza, V. A. Sobreiro, and H. Kimura, “Machine learning models for credit analysis improvements: Predicting low-income families’ default,” *Appl. Soft Comput. J.*, vol. 83, p. 105640, 2019, doi: 10.1016/j.asoc.2019.105640.

P. Bracke, A. Datta, C. Jung, and S. Sen, “Machine Learning Explainability in Finance: An Application to Default Risk Analysis,” *SSRN Electron. J.*, no. 816, 2019, doi: 10.2139/ssrn.3435104.

M. Modina and F. Pietrovito, “A default prediction model for Italian SMEs: the relevance of the capital structure,” *Appl. Financ. Econ.*, vol. 24, no. 23, pp. 1537–1554, 2014, doi: 10.1080/09603107.2014.927566.

J. Zhou, W. Li, J. Wang, S. Ding, and C. Xia, “Default prediction in P2P lending from high-dimensional data based on machine learning,” *Physica A*, vol. 534, p. 122370, 2019, doi: 10.1016/j.physa.2019.122370.

B. Alhijawi, G. Al-Naymat, N. Obeid, and A. Awajan, “Novel predictive model to improve the accuracy of collaborative filtering recommender systems,” *Inf. Syst.*, vol. 96, p. 101670, 2021, doi: 10.1016/j.is.2020.101670.

O. Oyebode and R. Orji, “A hybrid recommender system for product sales in a banking environment,” *J. Bank. Financ. Technol.*, vol. 4, no. 1, pp. 15–25, 2020, doi: 10.1007/s42786-019-00014-w.

A. Salam and F. Fathurrahmad, “Student Final Project Recommendation System Model Using Case-Based Reasoning (CBR) Method,” *J. Mantik*, vol. 5, no. 36, pp. 1535–1542, 2021, [Online]. Available:

http://iocscience.org/ejournal/index.php/mantik/article/view/1641%0Ahttps://iocscience.org/ejournal/index.php/mantik/article/download/1641/1190.

E. Hernández-Nieves, G. Hernández, A. B. Gil-González, S. Rodríguez-González, and J. M. Corchado, "Fog computing architecture for personalized recommendation of banking products," *Expert Syst. Appl.*, vol. 140, 2020, doi: 10.1016/j.eswa.2019.112900.

L. Xia, C. Huang, Y. Xu, J. Zhao, D. Yin, and J. X. Huang, "Hypergraph Contrastive Collaborative Filtering," pp. 70–79, 2022, doi: 10.1145/3477495.3532058.

I. C. Hsu and C. C. Chang, "Integrating machine learning and open data into social Chatbot for filtering information rumor," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 1, pp. 1023–1037, 2021, doi: 10.1007/s12652-020-02119-3.

A. Hossain, Z. Tasnim, and ..., "A recommender system for adaptive examination preparation using pearson correlation collaborative filtering," *Int. ...*, vol. 2, no. 1, pp. 30–43, 2021, [Online]. Available: https://researchlakejournals.com/index.php/AAIML/article/view/55%0Ahttps://researchlakejournals.com/index.php/AAIML/article/download/55/48.

T. Y. Kim, H. Ko, S. H. Kim, and H. Da Kim, "Modeling of recommendation system based on emotional information and collaborative filtering," *Sensors*, vol. 21, no. 6, pp. 1–25, 2021, doi: 10.3390/s21061997.

Z. Zhu, Y. He, X. Zhao, Y. Zhang, J. Wang, and J. Caverlee, "Popularity-Opportunity Bias in Collaborative Filtering," *WSDM 2021 - Proc. 14th ACM Int. Conf. Web Search Data Min.*, pp. 85–93, 2021, doi: 10.1145/3437963.3441820.

M. A. Salahli, T. Gasimzade, F. Alasgarova, and A. Guliyev, “The Use of Predictive Models in Intelligent Recommendation Systems,” *Procedia Comput. Sci.*, vol. 102, no. August, pp. 515–519, 2016, doi: 10.1016/j.procs.2016.09.436.

F. Wang, H. Zhu, G. Srivastava, S. Li, M. R. Khosravi, and L. Qi, “Robust Collaborative Filtering Recommendation With User-Item-Trust Records,” *IEEE Trans. Comput. Soc. Syst.*, no. March, 2021, doi: 10.1109/TCSS.2021.3064213.

J. Ranjan and C. Foropon, “Big Data Analytics in Building the Competitive Intelligence of Organizations,” *Int. J. Inf. Manage.*, vol. 56, no. August 2020, p. 102231, 2021, doi: 10.1016/j.ijinfomgt.2020.102231.

I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, “A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain,” *Mathematics*, vol. 10, no. 11, pp. 1–18, 2022, doi: 10.3390/math10111942.

R. A. Khurma, I. Aljarah, A. Sharieh, M. A. Elaziz, R. Damaševičius, and T. Krilavičius, “A Review of the Modification Strategies of the Nature Inspired Algorithms for Feature Selection Problem,” *Mathematics*, vol. 10, no. 3, pp. 1–45, 2022, doi: 10.3390/math10030464.

F. H. Hazboun, “Decision Support System Terms of SME’s Credit Lending Based on Machine Learning Approach,” *2022 Int. Conf. Sci. Inf. Technol. Smart Adm. ICSINTESA 2022*, no. November 2022, pp. 75–82, 2022, doi: 10.1109/ICSINTESA56431.2022.10041593.

F. Ciampi and N. Gordini, “Small Enterprise Default Prediction Modeling through Artificial Neural Networks: An Empirical Analysis of Italian Small Enterprises,” *J. Small Bus. Manag.*, vol. 51, no. 1, pp. 23–45, 2013, doi: 10.1111/j.1540-627X.2012.00376.x.

Z. Cheng, E. Zhu, S. Wang, P. Zhang, and W. Li, “Unsupervised Outlier Detection via Transformation Invariant Autoencoder,” *IEEE Access*, vol. 9, pp. 43991–44002, 2021, doi: 10.1109/ACCESS.2021.3065838.

K. B. Hansen and C. Borch, “The absorption and multiplication of uncertainty in machine-learning-driven finance,” *Br. J. Sociol.*, vol. 72, no. 4, pp. 1015–1029, 2021, doi: 10.1111/1468-4446.12880.

M. S. Ullal, I. T. Hawaldar, R. Soni, and M. Nadeem, “The Role of Machine Learning in Digital Marketing,” *SAGE Open*, vol. 11, no. 4, 2021, doi: 10.1177/21582440211050394.

E. Dumitrescu, S. Hué, C. Hurlin, and S. Tokpavi, “Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects,” *Eur. J. Oper. Res.*, vol. 297, no. 3, pp. 1178–1192, 2022, doi: 10.1016/j.ejor.2021.06.053.

M. N. Omri and W. Mribah, “Towards an Intelligent Machine Learning-based Business Approach,” *Int. J. Intell. Syst. Appl.*, vol. 14, no. 1, pp. 1–23, 2022, doi: 10.5815/ijisa.2022.01.01.

V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, “Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model,” *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, 2021, doi: 10.26599/BDMA.2020.9020016.

T. Top, “Machine learning in oil and gas exploration: Machine learning era,” *CC BY Attrib. 4.0 Int.*, no. March, p. 38, 2022, doi: https://doi.org/10.31223/X5WP8D.

K. Żbikowski and P. Antosiuk, “A machine learning, bias-free approach for predicting business success using Crunchbase data,” *Inf. Process. Manag.*, vol. 58, no. 4, p. 102555, 2021, doi: 10.1016/j.ipm.2021.102555.

E. Donnellan, S. Aslan, G. M. Fastrich, and K. Murayama, *How Are Curiosity and Interest Different? Naïve Bayes Classification of People’s Beliefs*, vol. 34, no. 1. Educational Psychology Review, 2022.

S. Zhang, J. Li, and Y. Li, “Reachable Distance Function for KNN Classification,” *IEEE Trans. Knowl. Data Eng.*, vol. PP, pp. 1–15, 2022, doi: 10.1109/TKDE.2022.3185149.

K. V. Kamran, B. Feizizadeh, B. Khorrami, and Y. Ebadi, “A comparative approach of support vector machine kernel functions for GIS-based landslide susceptibility mapping,” *Appl. Geomatics*, vol. 13, no. 4, pp. 837–851, 2021, doi: 10.1007/s12518-021-00393-0.

T. C. T. Tran, L. P. Phan, and H. X. Huynh, “Energy-based Collaborative Filtering Recommendation,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 7, pp. 557–562, 2022, doi: 10.14569/IJACSA.2022.0130766.

J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, “Evaluating the quality of machine learning explanations: A survey on methods and metrics,” *Electron.*, vol. 10, no. 5, pp. 1–19, 2021, doi: 10.3390/electronics10050593.

I. C. Yeh and C. hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 1, pp. 2473–2480, 2009, doi: 10.1016/j.eswa.2007.12.020.

C. F. Oduoza and M. H. Xiong, “A decision support system framework to process customer order enquiries in SMEs,” *Int. J. Adv. Manuf. Technol.*, vol. 42, no. 3–4, pp. 398–407, 2021, doi: 10.1007/s00170-008-1596-0.

C. S. Hong and T. G. Oh, “TPR-TNR plot for confusion matrix,” *Commun. Stat. Appl. Methods*, vol. 28, no. 2, pp. 161–169, 2021, doi: 10.29220/CSAM.2021.28.2.161.

P. G. Fonseca and H. D. Lopes, “Calibration of Machine Learning Classifiers for Probability of Default Modelling,” 2017, [Online]. Available: http://arxiv.org/abs/1710.08901.

الملخص

تحاط مؤسسات الإقراض والبنوك ببيئة أعمال متعددة المخاطر، وأهمها تلك المتعلقة بتخلف العملاء وعدم وفائهم بالتزاماتهم تجاه التزاماتهم، مما يؤدي إلى عدم قدرتهم على مواصلة أنشطتهم ومن ثم الفشل والإفلاس. يتمثل التحدي الرئيسي في تحديد أهم المؤشرات والمتغيرات التي تؤدي إلى تعثر الائتمان لعميل البنك، وبناء نموذج تنبؤ افتراضي يمكنه التنبؤ بشكل فعال ودقيق باحتمالية التخلف عن السداد قبل حدوثه جنبًا إلى جنب مع نظام التوصية. هذا ضروري لاتخاذ القرارات والتدابير اللازمة التي تمنع العملاء من التعثر في المستقبل والحصول على محفظة ائتمانية خالية إلى حد ما من الحسابات المتعثرة. أجريت هذه الدراسة على بيانات مقدمة من البنك الأهلي الأردني في فلسطين ل 80506 عميل بنكي من العملاء المنتظمين والعملاء المتعثرين. مجموعة البيانات مخصصة لهذه الدراسة حصريًا وتمت الموافقة عليها من قبل إدارة البنك. في هذه الدراسة ، تم تنفيذ خوارزميات مختلفة للتعلم الآلي لأن الاحتمالية قيمة متوقعة، حيث تجاوزت دقة الطرق الأكثر استخدامًا 90٪ ، وتفوق SVM على نماذج التصنيف الأخرى. نقترح نظام توصية يعتمد على تقنية تصفية تعاونية قائمة على العناصر (والتي تستخدم بيانات تفضيل العميل من الخوارزمية) والتي تعمل جنبًا إلى جنب مع خوارزميات التعلم الآلي لتقديم التوصية اللازمة لتجنب تعثر العميل. توضح النتائج التي توصلنا إليها أن تقنيات التعلم الالي مهمة لاكتشاف حالات التعثر والتنبؤ بها قبل حدوثها. ستساعد منهجيتنا ونتائجنا المؤسسات المالية والبنوك والمقترضين على حد سواء لمواجهة تحديات منح الائتمان وتجنب الكوارث المالية والحفاظ على محفظة ائتمانية نظيفة، من خلال معالجة مختلف المخاطر المالية وتحسين التصنيف الائتماني.