A FRAMEWORK FOR FEATURE SELECTION USING XGBOOST FOR PREDICTION BANKING RISK

AsmaaSaeed Embark[1], Riham Y. Haggag[2], Samir Aboul Fotouh Saleh[3]

[1]Departmentofinformationsystem, FacultyofCommerce&BusinessAdministration, Helwan University, Cairo, Egypt
[2]Business Information Systems department, Faculty of Commerce and Business Administration, Helwan University
[3]Department of Accounting &Information Systems, Faculty of Commerce, Mansoura University

ABSTRACT

machine learning methods have become one of the dominant approaches in an effort To find accurate predictions. Given the presence of large quantities of high dimensional data (which may come in a variety of noisy forms) and the lack of a comprehensive understanding at the molecular level, mining microarray data present strong challenges in Dealing with large amounts of data. To account for these challenges, Related Work has proposed different methods for selecting a Feature selection that can be used as an accurate Predictor, Feature selection is a very important pre-processing step of the data to discover the parts that help in the accuracy of the predictive performance of the model. This study's goal is a framework for feature selection using (XG-Boost, Linear Regression and Logistic Regression) for banking risk prediction. 350 examples and 19 characteristics make up the available Kaggle datasets that we used. The data were analyzed using SPSS, WEKA and python program. Based on our findings, this model exhibits greater effectiveness when compared to traditional feature selection methods. According to our results, XGBoost demonstrated superior discrimination capabilities when compared to alternative feature selection methods, namely Linear Regression and Logistic Regression, as evaluated by the AUC metric for forecast accuracy. The respective accuracy rates achieved by XGBoost and the two other methods were 91.43%, 89.12%, and 87%.

Keywords

Banking Risk, feature selection, XG-Boost, Linear Regression, Logistic Regression

1. INTRODUCTION

In the last century, the speed, size, and reasonableness of data have increased dramatically, and this is the result of the emergence of new technologies and establish, including the database , GPS, Internet, DNA microarray, mobile technology, and this results in big data. [1]

Big data has much more attributes or features than there are examples, which is one of its main properties. [2]

One of the most popular strategies used in microarray analysis is data mining. In the discipline of supervised learning, classifiers are developed and trained to categorize new cases according to a set of features obtained from the data as a topic. [3]

However, utilizing an excessive number of features in the classification method can be troublesome, especially if some of the features are pointless. Due to the small size of the training data, this can result in over fitting, where noise or irrelevant characteristics may have an overwhelming influence on the classification judgments. [4]

Despite data mining techniques' early success, the inclusion of a sizable number of irrelevant attributes renders such analysis relatively susceptible to the dimensionality curse.

The prediction accuracy may be increased by choosing only a portion of the qualities, and by focusing solely on the features that are pertinent to the forecast.[5]

A feature is a distinct, measurably present aspect of the process under observation. The process of choosing the most important characteristics from a given dataset is known as feature selection (FS). A machine learning model's performance can frequently be improved through FS as well. Many machine learning methods can do classification using a set of features. Figure 1 displays a general FS technique.[6]

This research aims to pald a framework for Feature Selection Using XG - Boost to Predic Banking Risk.

2. RESEARCH BACKGROUND

2.1. Machine Learning

Machine learning is teaching computers to perform tasks that go beyond standard number-crunching by learning from their environment through repeated instances. Machine learning typically involves a learning process where the computer learns from experience, or training data, in order to complete a task. This training data consists of a set of examples, each of which is described by a set of properties, or features, that can be binary, nominal, ordinal, or numeric. The performance of a machine learning model in a specific task is measured by a performance metric that is improved over time through experience. To evaluate machine learning models and algorithms, a range of statistical and mathematical methodologies are used. Once the learning process is complete, the trained model can be used to identify, predict, or cluster new examples, often using the knowledge gained during the training process, or testing data.

2.2.Technologies Used

2.3.XGBoost Technique

Gradient tree boosting is a machine learning technique that is highly effective in various applications. It has been shown that tree boosting can achieve exceptional performance on a range of commonly used classification benchmarks. In fact, tree boosting has been demonstrated to provide state-of-the-art results on several standard classification benchmarks.[7]

2.4. Linear Regression

Linear regression is a statistical approach that models the connection between one or more independent variables and a dependent variable. This technique assumes a linear correlation between these variables and endeavors to determine the most appropriate line of best fit that can describe the variation in the dependent variable based on the independent variables.[8]

2.5. Logistic Regression

Logistic regression is a statistical technique frequently employed for binary classification problems, where the objective is to forecast the likelihood of a binary result (e.g., yes or no, true or false) based on one or multiple input variables.[9]
Our thesis will utilize the previously mentioned methods to choose the most influential features for the model.

2.6. Bank Risk Management

Banks are vulnerable to different types of risks, both financial and non-financial, and it is crucial for them to regularly implement risk management strategies to handle and regulate these risks during their operations and for future contingencies. Risk management plays a crucial role in the banking and financial sector since the success or failure of banks or financial institutions depends on how they manage and mitigate risks. In banking, risk refers to the possibility of unexpected outcomes that may deviate from the expected or desired returns.[10]

Although there are various other risks related to banking, our thesis will solely concentrate on liquidity risks.

3. RELATED WORK

There are many studies on Feature Selection, where researchers have used many techniques, including a decision tree, support vector machine, logistic regression, artificial neural network, and other feature selection algorithms An overview of studies on feature selection is provided below.

Ben Jabeur Sami, et al., 2022, This paper aims to It is suggested to use the FS-XGBoost algorithm, which bases itself on feature importance selection Comparisons between FS-XGBoost and seven machine learning algorithms based on stepwise discriminant analysis, stepwise logistic regression, and partial least squares discriminant analysis, which are frequently used in bankruptcy prediction (PLS-DA). According to the results, logistic regression performs better than traditional feature selection algorithms and provides more accurate forecasts. [11]

R.Rajadevi E.M.Roopa Devi ..et al, 2021, The objective of this study is to propose the use of the black hole optimization (BHO) algorithm as an effective feature selection technique for medical diagnosis. After the relevant features are selected using the BHO algorithm from a medical dataset, they are utilized as inputs for the XGBoost classifier for classification purposes. The findings of this research suggest that the integration of the BHO algorithm and XGBoost technique helps in identifying smaller subsets of features, leading to improved diagnostic accuracy when compared to existing classifiers.[12]

Jaber Jemai, Anis Zarrad, 2023, This study proposes a feature selection engineering approach for credit risk assessment in the financial industry to determine the best features to learn from. Various feature selection methods, including univariate feature selection (UFS), recursive feature elimination (RFE), feature importance using decision trees (FIDT), and the information value (IV), were utilized. Two versions of the XGBoost classifier were employed on an open data set provided by the Lending Club platform to assess and compare the performance of these different

feature selection methods. The research indicates that all four feature selection techniques were successful in identifying the most relevant features.[13]

Nisha Arora, ..et al, 2020, The objective of this paper is to explore the application of modern data mining and machine learning techniques for accurate credit risk prediction and decision-making. To achieve this, feature selection techniques are employed to remove redundant and irrelevant attributes from the dataset. The study begins by introducing Bolasso (Bootstrap-Lasso), which effectively selects consistent and relevant features from a pool of features. The experimental results demonstrate that Bolasso provides superior stability of features compared to other algorithms, as evaluated using the Jaccard Stability Measure (JSM). Furthermore, the study finds that BS-RF (Bolasso Random Forest) exhibits higher classification accuracy, better AUC, and overall better performance than other methods, effectively improving the decision-making process for lenders.[14]

.

Pantelis Z. Lappas, Athanasios N. Yannacopoulos, 2021, The objective of this paper is to propose a strategy that combines soft computing methods with expert knowledge to improve credit scoring applications. The study highlights the usefulness of expert opinions in interpreting the predictive power of each feature in the credit dataset. To this end, a wrapper-based feature selection approach is proposed, which explores the features that contribute the most towards the classification of borrowers. The integration of expert knowledge and soft computing methods helps to strengthen the ability of interpretation and enhance the credit scoring process.[15]

Stjepan Oreski, Goran Oreski, 2014, This paper introduces a new advanced heuristic algorithm, the Hybrid Genetic Algorithm with Neural Networks (HGA-NN), which aims to identify the optimal feature subset and improve the classification accuracy and scalability in credit risk assessment. The proposed HGA-NN classifier is evaluated through experimental results and shows promising performance for feature selection and classification in retail credit risk assessment. This study suggests that the HGA-NN classifier is a valuable addition to existing data mining techniques in credit risk assessment.[16]

Elnahas, Ayat, et al., 2020, This study seeks to Also, a novel technique is suggested for choosing the best traits. These formulas were employed (SVM, KNN, NB) According to the experimental findings, the SVM classifier outperformed the KNN and NB classifiers in terms of performance. [21]

4. PROPOSED FRAMEWORK

This section thoroughly applies the general experimental methodologies. Pre-processing and feature selection are the two main parts of the experiment, and they are shown in Figure 2 below. The following description of the raw data comes first.

4.1. Description of Data

This research aims to develop a system for feature selection using XG-Boost for banking risk prediction using balance sheet data. It is available on the Kaggle website [27] information gathered from the Bank of England This dataset consists of 350 instances and 19 variables. Table 1 provides comprehensive explanations of the characteristics.

Table 1 Description of the Dataset

Features	Description
Year	Year
Total Assets	Total Assets
Government debt	Denotes the Government debt
Other Government securities	Denotes the Other Government securities
Other securities	express the Other securities
Coin and bullion	It describes the coin and gold.
Notes in the Bank	In the Bank Notes
Notes In circulation	Notes In circulation
Notes in the Bank	Deposited Notes
Capital	Expresses the value of a Capital
Rest	Rest
Deposits	Expresses the value of a Deposits
o/w Public deposits	Describes an o/value w's Public deposits
o/w Special deposits	Describes the value of a special deposit in o/w
o/w Bankers deposits	Expresses the value of an o/w Bankers deposits
Other accounts	Expresses the value of Other accounts
7 day and other Bills	7 day and additional Bills
Total Liabilities	Total liabilities
Check Assets=Liabilities	Check Assets=Liabilities

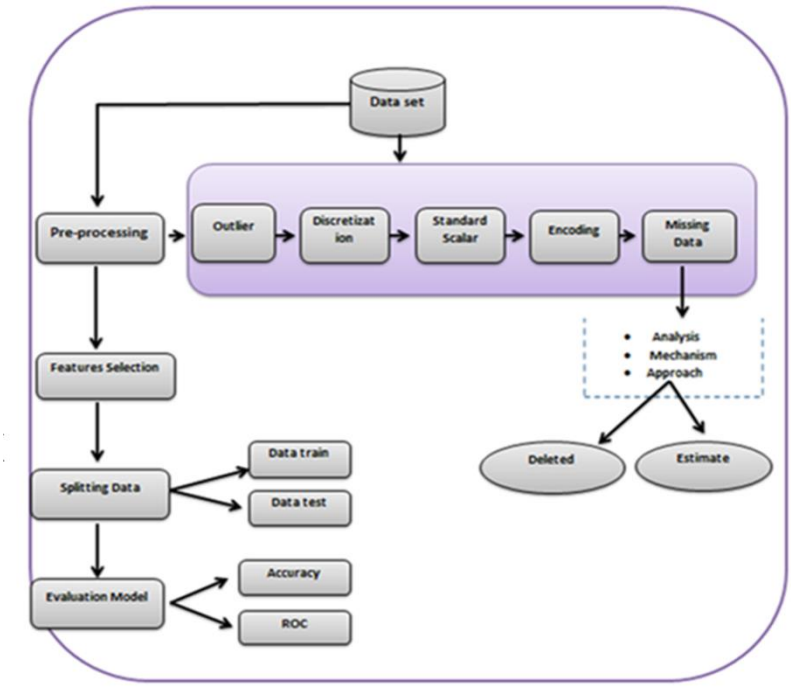

Figure 1 General layout of the suggested model

Using machine learning techniques (MLT) with the loans, deposits, and securities given in figure, we created a framework for feature selection using XG-Boost for banking risk prediction (1). data collecting phase, (2) data preprocessing before applying the MLT, (3) feature selection, (4) data

partitioning into training and testing, (5) selection of classification models, and (6) assessment phase to assess the developed model's accuracy using a machine learning technique

4.2. Data Pre-processing

In order to develop the predictive model, the raw data often contain missing values and inconsistencies. Therefore, it is necessary to preprocess the data prior to using the predictive model. The following steps were taken to achieve data preparation in this section.

4.3. Handle Missing Data

Figure 2 and Figure 3 present an analysis of the missing data in the dataset being used, which reveals the presence of numerous missing values. This analysis was conducted using the SPSS tool.

4.4. Missing Data Analysis

After removing the noise from the dataset, it was observed that a significant number of quantitative variables had missing values. In fact, missing values were present in all cases. The exact counts and percentages of these missing values are displayed in figures 2 and 3. Figure 2 highlights that 9 out of 19 variables in the dataset contain missing values, which account for 47.37% of the total variables. Furthermore, at the row level, all cases were observed to have missing values.

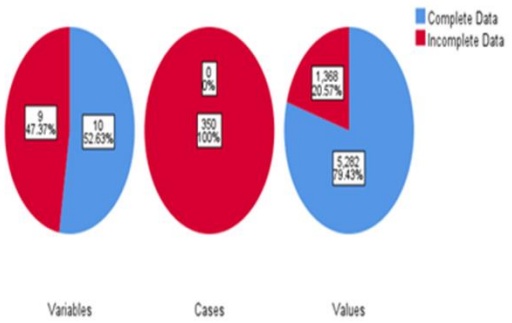

Figure 2.Summary of all missing data

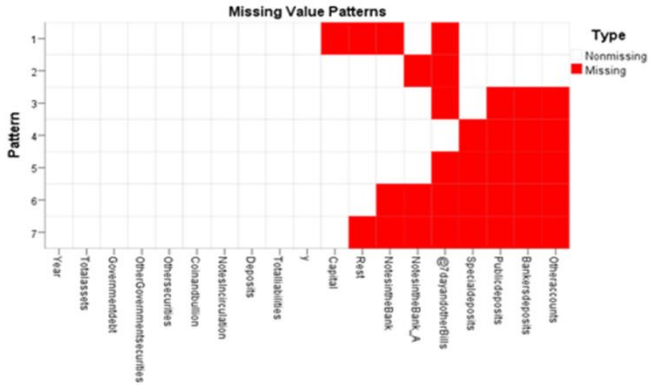

Figure3.overall patterns of missing values

The missing value patterns for the analysis variables are displayed on the patterns chart. Each pattern, which is a collection of occurrences, represents the same complete and partial data patterns. The situations shown in Figure 3 all suffer from data loss.

4.5. MCAR Test

Rubin and colleagues published a system for categorizing issues with missing data. The findings of this study lead to three concepts that are referred to as "missing data processes" and explain how the likelihood of missing data relates to the observed data. Missing values, often known as missing at random, have a consistent relationship with one or more evaluated attributes (MAR). Whereas missing completely at random (MCAR) asserts a relationship between missing data and observed data, missing not at random (MNAR) proposes a connection between the probability of missing data on a variable Y and the values of Y.In this phase, we first determine if the data was wholly randomly lost or not using Little's We identify the manner by which the data was lost and the best technique to manage the missing data based on the fact that it was fully missing from the test. [28]

Null hypotheses H0 = MCAR
Alternative hypothesis H1≠ MAR
Since the sig is greater than .05 then this indicates that the missing data (= MCAR or ≠ MAR).

4.6. Impute Missing Values

In order to finalize the previous phase, the next step involves using the multiple-imputation approach to estimate the missing values in three stages: imputation, analysis, and pooling. This method involves creating several copies of the dataset (e.g., m=3) that include different estimates of the missing values. Essentially, the imputation stage is a repeated version of the stochastic regression method. Equation 1 outlines the algorithm for the multiple-imputation method.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_k X_k + \varepsilon$$

The pooling step is given by:

$$\overline{\beta}_{MI} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\beta}_i \qquad (1)$$

To verify the accuracy of the imputation method, we conducted a mean comparison test to check the consistency between the three iterations and the mean before handling the missing data. The results were satisfactory as the averages were very similar both before and after imputation.

5. DISCRETIZATION

A crucial method of data reduction is discretization. Its primary goal is to discretize a set of continuous variables by partitioning the range of the variables into a finite number of discrete intervals and coupling each interval with a denotation label. [29]

All features in the dataset have already been partitioned into a finite number of discrete time intervals. Figure 4 and 5 depict one of the features in the dataset before and after the range of variables was divided into a finite number of intervals.

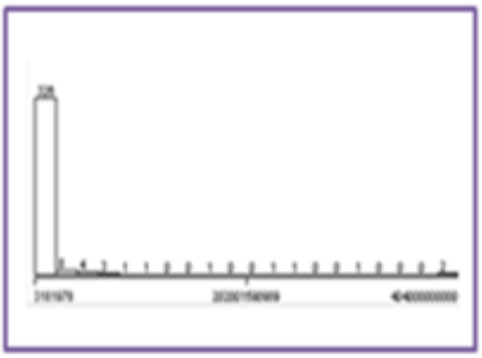

Figure 4 Total assets before discretization Figure 5. Total assets after discretization

6. STANDARDIZATION

We used standardized processing on all attributes in the data set to convert the raw data into a dimensionless index, where each index value is at the same scale level, due to the different properties of the indicators. [21]

7. SPLITTING DATA

To analyze the XG - Boost For model, the dataset must first be divided into two groups: training data (20% of the whole data), and test data (80% of the total data). Loans, securities, and deposits are used in this classification approach to forecast the goal. A suitable and logical model is trained using training data in order to identify features that have a significant impact on the prediction. Testing data are used to calculate the model's prediction accuracy, which can demonstrate the model's usefulness and efficiency. [30]

8. FEATURE SELECTION METHODS

The model training step may have suffered as a result of the original characteristics in certain situations being noisy and redundant. As a result, using a reliable feature extraction technique is necessary for good categorization operations. [31]

Numerous predictor variables frequently have an impact on the area of mathematical applications of the outcome. It is also advantageous to take note of their variable importance, which shows the importance of each input variable to the total output of the model. [32]

The best feature subset can be chosen between output learning and model simplicity depending on the trade-off (i.e., fewer features).Without compromising accuracy, redundant and unneeded variables can be removed. Improved interpretability, streamlined modeling, quicker learning, and stronger generalizations are all benefits of feature selection.[33]

In this study, three commonly used FS methods are applied and evaluated: XGboost, Linear Regression, and Logistic Regression We will discuss them in detail below.

8.1. Feature Selection with XGBoost

The gradient-boosting algorithms in the XGBoost library have been enhanced for use with current data science tools and issues. It receives a boost and is packaged in a useful library.

The fact that XGBoost is extremely scalable and parallelizable, rapid to execute, and often outperforms other algorithms are some of its main advantages.[11]

In this section dealt with xgboost Clarifying the feature importance of the influencing the model and clarifying the accuracy of this algorithm

```
In [10]: #Split data into training features and labels
         x = dataset.iloc[:, :-1].values
         y = dataset.iloc[:, 18].values
         from sklearn.model_selection import train_test_split
         # split data into train and test sets
         seed = 7
         test_size = 0.1
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=test_size, random_state=seed)

In [11]: #import XGBoost classifier and accuracy
         from xgboost import XGBClassifier
         from sklearn.metrics import accuracy_score

         #instantiate model and train
         model = XGBClassifier(learning_rate = 0.05, n_estimators=300, max_depth=5)
         model.fit(x_train, y_train)

         # make predictions for test set
         y_pred = model.predict(x_test)
         predictions = [round(value) for value in y_pred]

         accuracy = accuracy_score(y_test, predictions)
         print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

Figure 6. Training and testing of a model

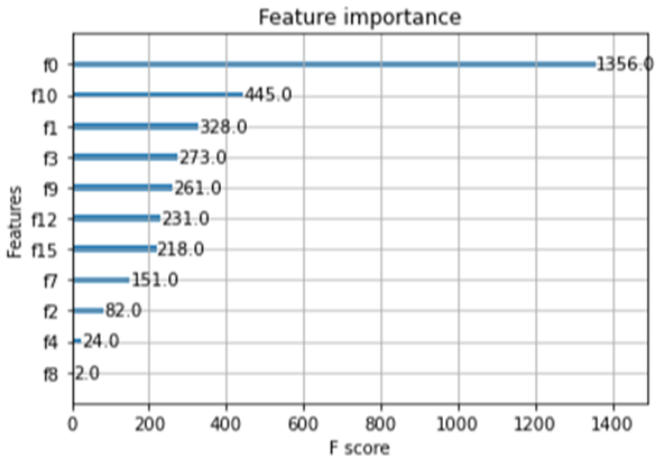

Figure 7.featureimportanc

Figure (7) shows the most important features affecting the model.

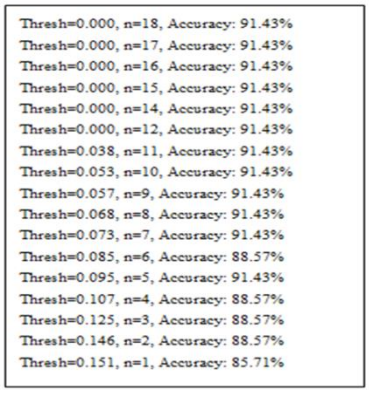

Figure 8.feature selection with xgboost

This piece of code iteratively removes features in order of significance to train and test the model, while also keeping track of the model's correctness. This makes it simple to delete functionality without having to rely solely on trial and error. Permission selected features through xgboost they 7 Features and xgboost accuracy is 91.43% .

8.2. Feature Selection with Linear Regression

The associations between several independent variables and a dependent variable are examined via many linear regression. The selection of features is the critical step that determines how well the model performs. If we take into account all the factors, even the best model will perform the worst. Thus, we will use linear regression to pick features. We evaluated the data, trained the model, and tested it.

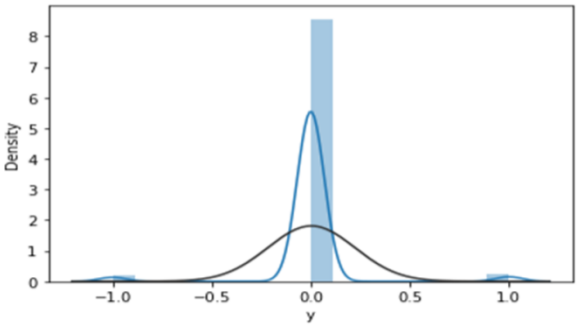

Figure 9.data normality

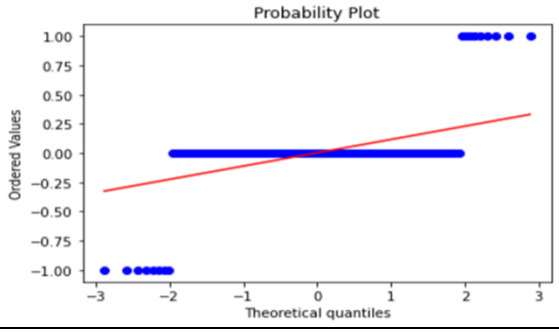

Figure 10.data probability

Figure 9 and 10 show the nature of the data and the extent of its distortion. Here we will find out how the data is related to each other.

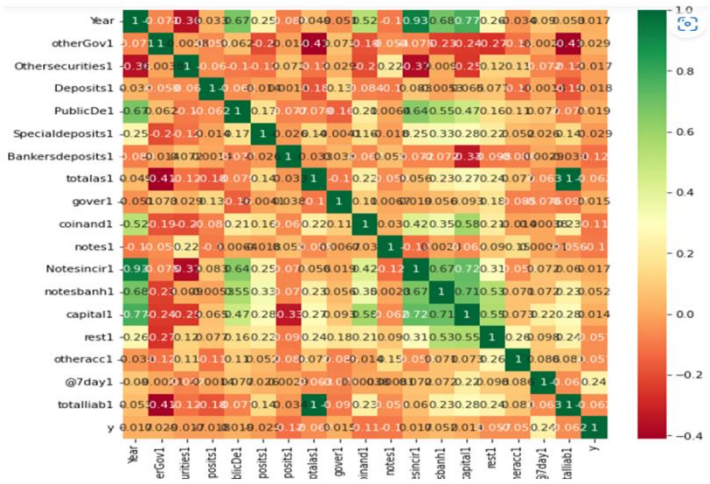

Figure 11.data correlation

Figure 11 clearly shows that the correlation coefficient is between -1 and 1. There is a strong positive association when it is close to 1.

Lastly, coefficients close to 0 indicate that there is no linear association. When the coefficient is close to -1, it indicates that there is a strong negative correlation.

```
In [29]: from sklearn.feature_selection import RFE
         from sklearn.linear_model import LinearRegression
         # drop the 1 column in X
         RFE_regressor = LinearRegression()
         #Initializing RFE model
         rfe = RFE(RFE_regressor)# random number(2)
         #Transforming data using RFE
         X_rfe = rfe.fit_transform(x,y)
         #Fitting the data to model
         RFE_regressor.fit(x,y)
         print(rfe.support_)
         print(rfe.ranking_)

         [False False False False False False  True  True  True  True False False
           True  True  True False  True  True]
         [10  6  9  4  8  5  1  1  1  1  2  7  1  1  1  3  1  1]
```

Figure 12.fearure selection with linear regression

Figure shows 12use RFE (Recursive Feature Elimination) to find the characteristics that are most connected with the target variable. It then ranks all the variables, with 1 being the most significant. It also offers its assistance, with True representing a relevant feature and False representing an irrelevant feature.
After the training and testing phase of the model it turned out to be accurate linear regression Accuracy: 89.12% And was chosen Num Features 8 out of 18 features.

8.3. Feature Selection with Logistic Regression

We will use Logistic Regression this is to select the most important features that affect the model. To choose the best features, employ RFE in conjunction with the Logistic Regression classifier.

```
In [49]: # Import your necessary dependencies
         from sklearn.feature_selection import RFE
         from sklearn.linear_model import LogisticRegression
         import matplotlib.pyplot as plt

In [55]: # Feature extraction
         model = LogisticRegression()
         rfe = RFE(model)
         fit = rfe.fit(x, y)
         print("Num Features: %s" % (fit.n_features_))
         print("Selected Features: %s" % (fit.support_))
         print("Feature Ranking: %s" % (fit.ranking_))

Num Features: 9
Selected Features: [False False  True False  True False False False False  True  True  True
 False  True  True  True  True False]
Feature Ranking: [10  7  1  6  1  2  4  5  9  1  1  1  3  1  1  1  1  8]
```

Figure 13.fearure selection with Logistic Regression

The figure shows 13 uses RFE (Recursive Feature Elimination) to find the

characteristics that are most connected with the target variable. It then ranks all the variables, with 1 being the most significant. It also offers its assistance, with True representing a relevant feature and False representing an irrelevant feature.

After the training and testing phase of the model, it turned out to be accurate Logistic Regression Accuracy: 87% And was chosen Num Features 9out of 18 features.

8.4.The Evaluation Model and Outcomes of Experiments

To begin evaluating the three machine learning models, it is required to split the dataset into two groups: training data (20%) and test data (80%), each of which accounts for 20% of the total data. As we previously discussed.

The test data will then be entered into the three suggested models to determine their performance.

We calculated the accuracy of the models using precisions, the findings showed that the xgboost had achieved a high accuracy of 91.43%, as shown in Table (2).

$$ACC = True\ Positives + True\ Negative\ /\ Total\ population$$ [34]

Table 2: Comparison Techniques results

TECHNIQUES	RESULTES
Xgboost	91.43%
Linear Regression	89.12%
Logistic Regression	87%

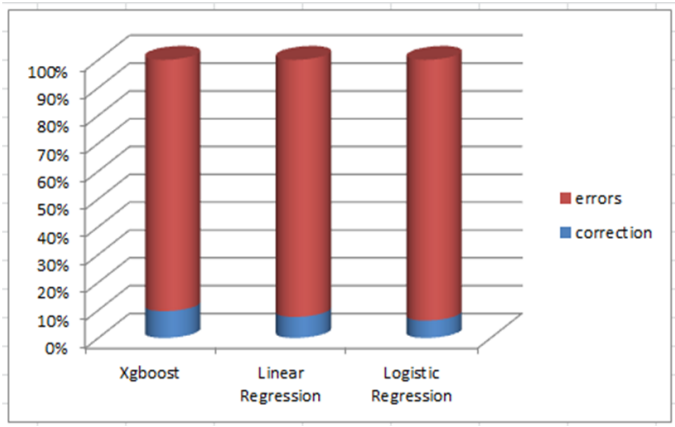

Figure 14. Comparison Techniques results

Previous studies have explored various algorithms for feature selection in credit risk assessment, with the third study being the only one to include XGBoost alongside other algorithms. The results indicated that XGBoost was equally accurate compared to the other algorithms. The first study used both XGBoost and logistic regression for feature selection, with the research revealing the superiority of logistic regression over XGBoost. In the seventh study, feature selection was performed using SVM, KNN, and NB, with the results demonstrating that SVM was the most effective. In contrast, our thesis utilized XGBoost, linear regression, and logistic regression for feature selection, and the findings highlighted the superiority of XGBoost over all other algorithms.

9. CONCLUSIONS

In our novel concept, feature selection techniques and new banking risk methodologies are merged (XGBoost - Linear Regression - Logistic Regression). Our results show that our approach outperforms traditional feature selection techniques. The method's impressive ability to tell apart different observations would seem to confirm its efficacy. Our results demonstrate that, using AUC as a measure of forecast accuracy, XGBoost has a higher discrimination power when compared to other feature selection algorithms, such as Linear Regression and Logistic Regression. Our estimates show that improving the AUC requires a careful examination of various feature selection techniques. Faster processing times and reduced complexity are typically the results of a reduction in the number of banking risk indicators.

REFERENCES

[1] FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, 1996, 17.3: 37.

[2] BEN-BASSAT, M. Pattern recognition and reduction of dimensionality. Handbook of Statistics-II, PR Krishnaiah and LN Kanal, eds, 1982, 773-791.
[3] SIEDLECKI, Wojciech; SKLANSKY, Jack. On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence, 1988, 2.02: 197-220.
[4] JIAWEI, Han; KAMBER, Micheline. Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, 2001, 5.
[5] Cannas, Laura Maria. "A framework for feature selection in high-dimensional domains." (2013).
[6] Shah, S. A., et al. "A comparative study of feature selection approaches: 2016-2020." International journal of scientific and engineering research 11.2 (2020): 469.
[7] Li, Ping. "An empirical evaluation of four algorithms for multi-class classification: Mart, abc-mart, robust logitboost, and abc-logitboost." arXiv preprint arXiv:1001.1020 (2010).
[8] Belgiu, M., &Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.
[9] Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage.
[10] Ahmed, Ayaz, and Henna Ahsan. "Contribution of services sector in the economy of Pakistan." (2011).
[11] Ben Jabeur, Sami, NicolaeStef, and Pedro Carmona. "Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering." Computational Economics (2022): 1-27.
[12] Rajadevi, R., Devi, E. R., Shanthakumari, R., Latha, R. S., Anitha, N., &Devipriya, R. (2021, January). Feature selection for predicting heart disease using black hole optimization algorithm and xgboost classifier. In 2021 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-7). IEEE.
[13] Jemai, J., &Zarrad, A. (2023). Feature Selection Engineering for Credit Risk Assessment in Retail Banking. Information, 14(3), 200.
[14] Arora, N., &Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. Applied Soft Computing, 86, 105936
[15] Lappas, P. Z., &Yannacopoulos, A. N. (2021). A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment. Applied Soft Computing, 107, 107391.
[16] Oreski, S., &Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.
[17] Peng, Hua, Yicheng Lin, and Mingzheng Wu. "Bank Financial Risk Prediction Model Based on Big Data." Scientific Programming 2022 (2022).
[18] Yao, Gang, et al. "Enterprise credit risk prediction using supply chain information: A decision tree ensemble model based on the differential sampling rate, Synthetic Minority Oversampling Technique and AdaBoost." Expert Systems (2022): e12953.
[19] Javeed, Ashir, et al. "Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification." Mobile Information Systems 2020 (2020).
[20] Alonso, Andrés, and Jose Manuel Carbo. "Machine learning in credit risk: Measuring the dilemma between prediction and supervisory cost." (2020).
[21] Elnahas, Ayat, et al. "Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison, and Proposal." The Egyptian Journal of Language Engineering 7.2 (2020): 1-19.
[22] Chantar, Hamouda, et al. "Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification." Neural Computing and Applications 32.16 (2020): 12201-12220.
[23] Selvi, S., and M. Chandrasekaran. "Framework to forecast environment changes by optimized predictive modelling based on rough set and Elman neural network." Soft Computing 24.14 (2019): 10467-10480.
[24] Vecoven, Nicolas. "Master thesis: Feature selection with deep neural networks." (2017).
[25] Vege, Sri Harsha. "Ensemble of feature selection techniques for high dimensional data." (2012).
[26] vanGinkel, Joost R., et al. "Rebutting existing misconceptions about multiple imputation as a method for handling missing data." Journal of Personality Assessment 102.3 (2020): 297-308.https://www.kaggle.com/sohier/the-bank-of-englands-balance-sheet?select=balance.xlsx
[27] Enders, Craig K. Applied missing data analysis. Guilford press, 2010.

[28] Eekhout, Iris, et al. "Missing data in a multi-item instrument were best handled by multiple imputation at the item score level." Journal of clinical epidemiology 67.3 (2014): 335-342.
[29] Mohamad, Ismail Bin, and DaudaUsman. "Standardization and its effects on K-means clustering algorithm." Research Journal of Applied Sciences, Engineering and Technology 6.17 (2013): 3299-3303.
[30] Yadav, Sanjay, and SanyamShukla. "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification." 2016 IEEE 6th International conference on advanced computing (IACC). IEEE, 2016.

[31] Yu, Jialin, et al. "PredGly: predicting lysine glycation sites for Homo sapiens based on XGboost feature optimization." Bioinformatics 35.16 (2019): 2749-2756.
[32] Jones, Stewart. "Corporate bankruptcy prediction: a high dimensional analysis." Review of Accounting Studies 22.3 (2017): 1366-1422.
[33] García, Salvador, et al. "Big data preprocessing: methods and prospects." Big Data Analytics 1.1 (2016): 1-22.
[34] Miguéis, Vera L., Ana S. Camanho, and José Borges. "Predicting direct marketing response in banking: comparison of class imbalance methods." Service Business 11.4 (2017): 831-849.
[35] Gupta, Chelsi. Feature selection and analysis for standard machine learning classification of audio beehive samples. Diss. Utah State University, 2019.
[36] Cannas, Laura Maria. "A framework for feature selection in high-dimensional domains." (2012).