

**DOKUZ EYLUL UNIVERSITY  
ENGINEERING FACULTY  
DEPARTMENT OF COMPUTER ENGINEERING**

**CME4434  
Data Warehouses and Business Intelligence**

**CREDIT RISK PREDICTION**

**By  
2018510019 DENİZ DOĞAN  
2018510059 ALPEREN TURHAN  
2019510017 ALİ SİYAR ARSLAN**

**31.12.2023**

**İZMİR**

# Joint Contributions to This Project

## **Deciding on the Dataset:**

We started by clearly defining the goals of our project. We understand the problem we aim to solve or analyze with the help of data. We considered the suitability of potential datasets for your goals.

## **Examining the Selected Dataset:**

We conducted an initial exploration of the dataset to gain a high-level understanding of its structure, size, and format.

Define the variables (features) present in the dataset. We understand the types of variables (numeric, categorical) and their potential relevance to your analysis or prediction task.

For numerical variables, we calculated and examined descriptive statistics such as mean, median, standard deviation, and quartiles.

# My Contributions to This Project

## **Analyzing Numerical and Categorical Variables:**

I performed a comprehensive analysis of both numerical and categorical variables in the dataset.

I examined the distributions and characteristics of numerical features such as age, loan amount, and duration.

I explored the frequency and patterns of categorical variables such as gender, job classification, housing status, and purpose of loan application.

## **Performing Missing Observation Analysis:**

To assess the extent of missing data, I identified and analyzed missing observations in the data set.

I have formulated strategies to handle missing values and ensure data integrity and integrity for subsequent analysis.

### **One-Hot Encoding:**

I applied one-hot encoding to categorical variables, converting them into a binary format suitable for machine learning algorithms.

By improving the algorithm's ability to capture patterns and relationships, I ensured that categorical information was included in the model training process.

### **Model Training, and Hyperparameter Optimization for XGBoost, LightGBM and Neural Networks:**

I trained and analyzed XGBoost , LightGBM and Neural networks machine learning algorithms for credit risk prediction.

I performed model training using the dataset, leveraging the power of these algorithms to learn complex patterns and relationships within the data.

I used hyperparameter optimization techniques to fine-tune the model parameters, improving the performance and prediction accuracy of the algorithms.

I leveraged neural network algorithms for credit risk prediction, highlighting the role of deep learning in financial decision-making.

I performed model training with neural networks and optimized hyperparameters to improve the model's ability to learn complex patterns in the data.

### **Feature Importance Analysis for XGBoost and LightGBM:**

I researched and analyzed the importance of features in credit risk prediction models using XGBoost and LightGBM.

I found influential variables identified that contributed significantly to the predictive power of the models.

I provided insights into the key factors affecting credit risk, helping to interpret the model results.

# Abstract

Credit risk management is a critical aspect of financial institutions' sustainability and decision-making processes. The study begins with an introduction emphasizing the global significance of credit risk management and the increasing role of machine learning in financial decision-making. The dataset used in the study comprises various characteristics and financial attributes of individuals applying for credit, including age, gender, job classification, housing status, savings, credit amount, duration, purpose, and the outcome of the credit application (positive or negative). A thorough dataset analysis explores feature distributions, statistical summaries, and a correlation matrix, providing insights into potential influential factors for credit risk prediction. The importance and use of the dataset in training machine learning algorithms are highlighted, emphasizing its role in making accurate financial decisions. The subsequent sections detail the data preparation and feature engineering processes. Data cleansing, preprocessing, and the identification of significant features are crucial steps in ensuring the dataset's suitability for modeling. Categorical variables are transformed, and numerical variables are scaled for consistency, contributing to better model learning. Feature engineering involves the selection of influential features, processing of categorical variables, creation of new features, and transformation of numerical variables. The model selection and application stage explore various machine learning algorithms, including K-Nearest Neighbors, Support Vector Classification, Neural Networks, Decision Tree, Random Forest, Gradient Boosting Machines, XGBoost, LightGBM, and CatBoost. Evaluation metrics such as accuracy, precision, recall, and F1 score are used to identify the best-performing models—XGBoost and LightGBM.

# Introduction

In the global financial system, credit risk management is a critical element that forms the foundation of financial institutions' sustainability and decision-making processes. Credit risk refers to the likelihood of borrowers being unable to repay their loans, which significantly impacts not only financial stability but also economic growth.

Presently, banks and financial institutions are developing various methods to manage the credit risk they face and maintain their financial stability. Lending decisions consider factors such as individuals' and companies' payment histories, financial standings, and many other elements.

At this juncture, the importance of machine learning techniques in financial decision-making processes is increasingly recognized. This study aims to utilize machine learning techniques in credit risk prediction. Accurate credit risk prediction can assist financial

institutions in making more informed and effective decisions while enabling more efficient utilization of resources.

### **Significance and Objectives of the Study**

The primary objective of this study is to develop more accurate and reliable models using machine learning techniques for credit risk prediction. These models could enable financial institutions to minimize risks and make more informed decisions.

The focus of the study is on enhancing customer satisfaction and maintaining financial stability. Accurate credit risk prediction allows institutions to provide better services to their customers while effectively managing risks to preserve their financial stability.

### **Expected Outputs and Impacts**

This study is anticipated to have a significant impact on applications in the financial sector. The developed models and findings can guide the industry in optimizing credit processes, minimizing risks, and making more informed decisions in financial decision-making processes.

In the traditional lending process, loan institutions mainly adopt the "5C" principle (Character, Capacity, Collateral and Conditions) to evaluate borrowers' capability subjectively. This type of evaluation relies on the personal experience and knowledge of the creditors, and most of the information is obtained from the customer relationship. Such method has great limitations in the retail credit field with high customer mobility and urgent need for business expansion. Since the 1960s, credit scoring has been used to determine whether a borrower is able to apply for a loan and is willing to repay it on time. Credit scoring guided credit decisionmaking refers to implementation of mathematical models that convert relevant data collecting from customers themselves, internal systems and credit agencies into a value. In the field of retail credit, this method not only reduces the subjective judgment of the creditors, but also maximizes the value of available information, and greatly saves the cost of human resources.

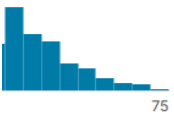

The commonly used scoring methods include the scoring method based on logistic regression algorithm, which has the advantages of: (1) simple algorithm and mature technology; (2) robust estimation of probability under given data conditions; (3) strong explanatory power of variables and models; (4) easy detection and deployment of models. But at the same time, there are also some problems in the traditional credit scoring system: (1) there are limited variables in the model and the possibility of being attacked by fraudsters; (2) the logistic regression algorithm needs to meet some assumptions, but the actual business may not meet the corresponding assumptions; (3) the differentiation ability of the logistic regression algorithm is difficult to improve. In the past 40 years, besides logistic regression algorithm, other supervised learning methods in machine learning have been developed rapidly such as neural network, nearest neighbor methods and support vector machine[1]. Logical regression can be traced back to the 1950s. Three classical implementations of decision tree, ID3, CART and C4.5, were important achievements from 1980s to 1990s. At the same time, the theory of neural network was greatly enriched and improved. In 1995, the birth of two classical algorithms, SVM and AdaBoost, marked the triumph of core technology and integrated learning algorithm respectively. Subsequently, random forest, GBDT, XGBoost and LightGBM arithmetic appeared one after another[2]. This paper explores the use of supervised machine learning method for credit risk prediction, and carries out a detailed study of the forecasting methods and forecasting effect.

# Dataset

This study focuses on a machine learning study centered around credit risk prediction. The dataset used includes various characteristics and financial attributes of individuals applying for credit. The dataset contains the following columns:

- Age: Age of individuals applying for credit.
- Sex: Gender of individuals applying for credit.
- Job: Job classification (represented with values like 0, 1, 2, 3).
- Housing: Housing status - owned, rented, or free.
- Saving accounts: Amount of savings in the bank account.
- Checking account: Status of the checking account.
- Credit amount: Amount of credit.
- Duration: Duration of credit payment.
- Purpose: Purpose of the credit application.
- Risk: Outcome of the credit application - positive or negative.

Here is our part of dataset:

Age	Sex	Job	Housing	Saving accounts	Checking account
	male 69% female 31%		own 71% rent 18% Other (108) 11%	little 60% NA 18% Other (214) 21%	NA little Other (332)
	male	2	own	NA	little
	female	2	own	little	moderate
	male	1	own	little	NA
	male	2	free	little	little
	male	2	free	little	little
	male	1	free	NA	NA
	male	2	own	quite rich	NA
	male	3	rent	little	moderate
	male	1	own	rich	NA
	male	3	own	little	moderate
	female	2	rent	little	moderate
	female	2	rent	little	little
	female	2	own	little	moderate
	male	1	own	little	little
	female	2	rent	little	little
	female	1	own	moderate	little

## Dataset Analysis

The dataset contains '1000' examples, each representing a credit application. Each feature in the dataset could play a significant role in predicting credit risk. For instance, features like payment history, income status, and credit amount can be influential in determining credit risk.

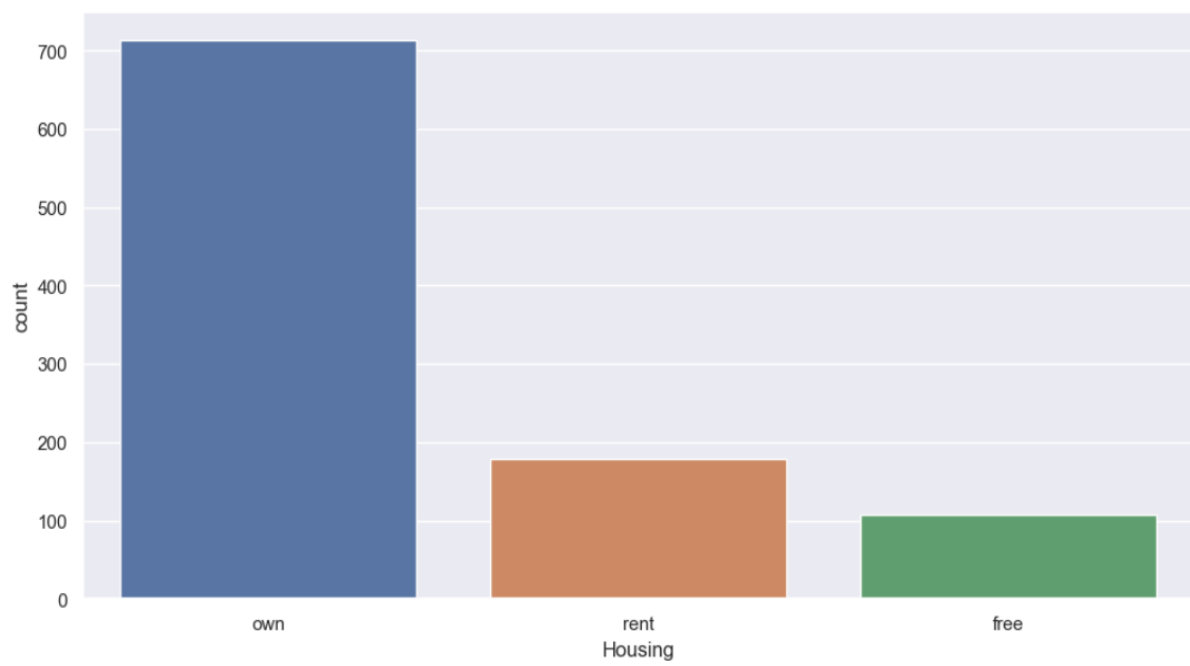
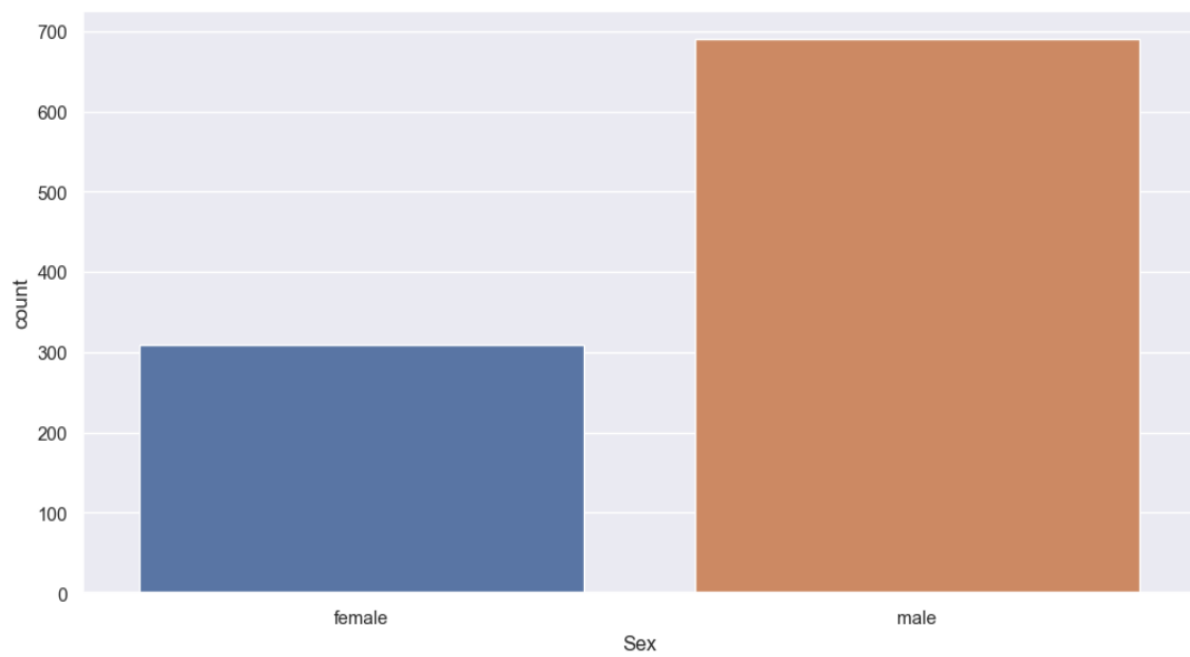
Structural analysis of the dataset, including feature distributions, statistical summaries, and representation of sample data, can provide insight into understanding the dataset and which features could be considered during the modeling process.

The dataset comprises several key columns, including Age, representing the age of credit applicants, Sex, denoting the gender of individuals applying for credit, and Job, which categorizes applicants based on job classification values (0, 1, 2, 3).

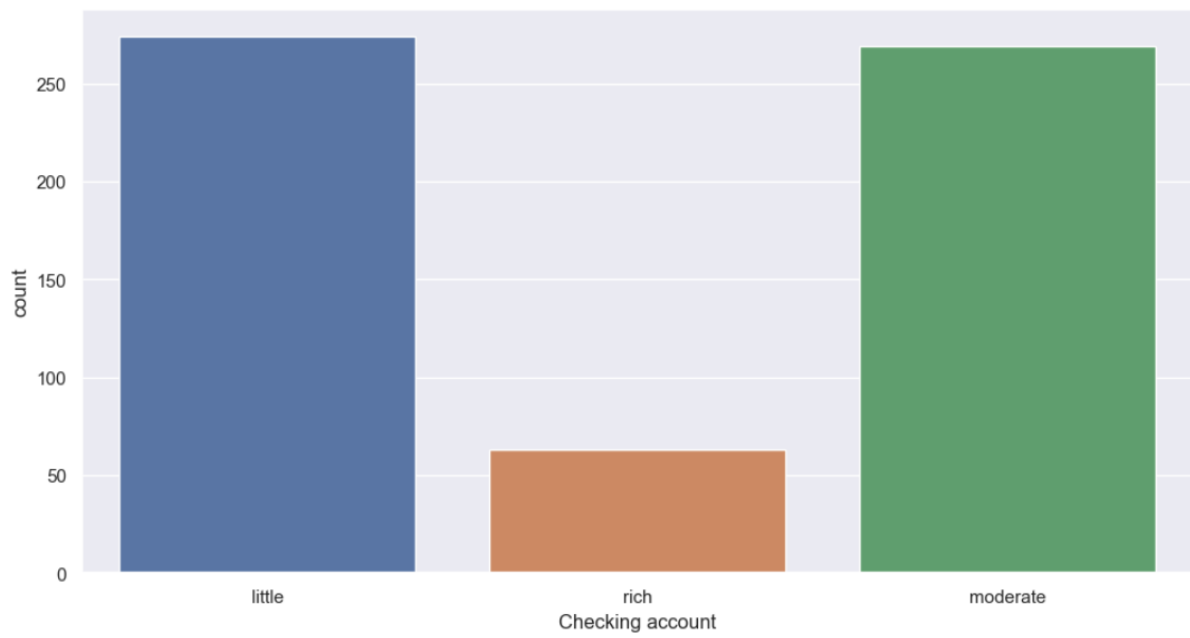
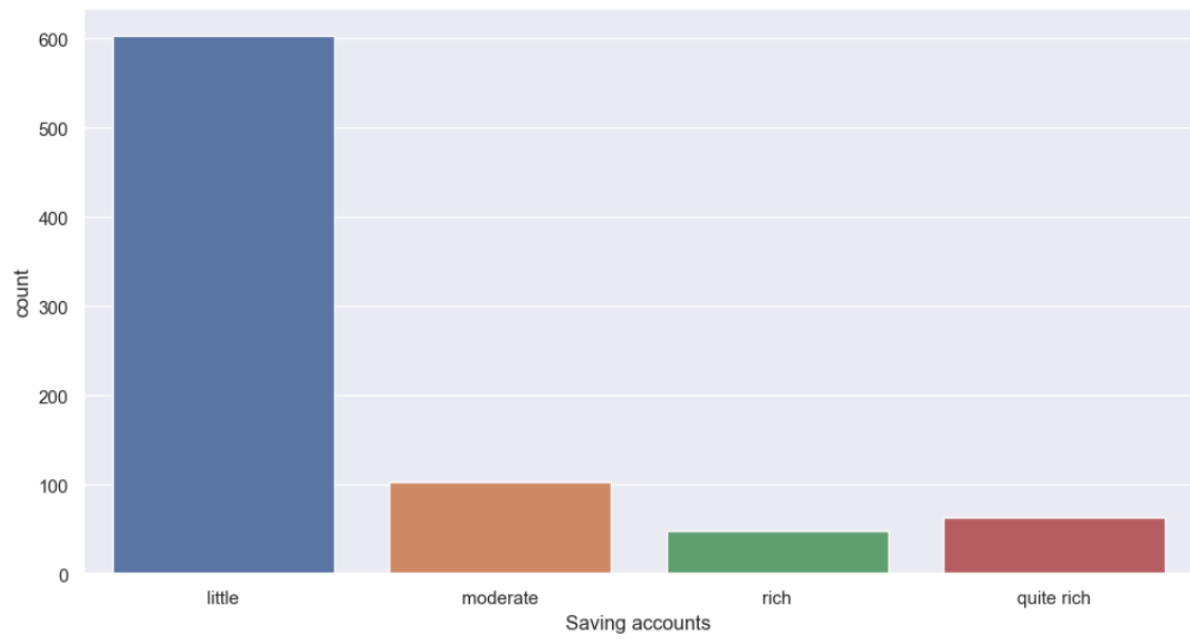
Additionally, the dataset includes Housing, delineating the housing status of applicants as owned, rented, or free. The Saving accounts column quantifies the amount of savings held in the bank account, while Checking account provides insight into the status of applicants' checking accounts. Other crucial attributes encompass Credit amount, indicating the requested credit amount, Duration, representing the duration of credit payment, and Purpose, detailing the reason behind the credit application.

The final column, Risk, serves as the outcome indicator for the credit application, distinguishing between positive and negative results. Through an exploration of these variables, the study aims to develop robust models that enhance credit risk prediction capabilities, contributing to more informed decision-making in the financial sector.

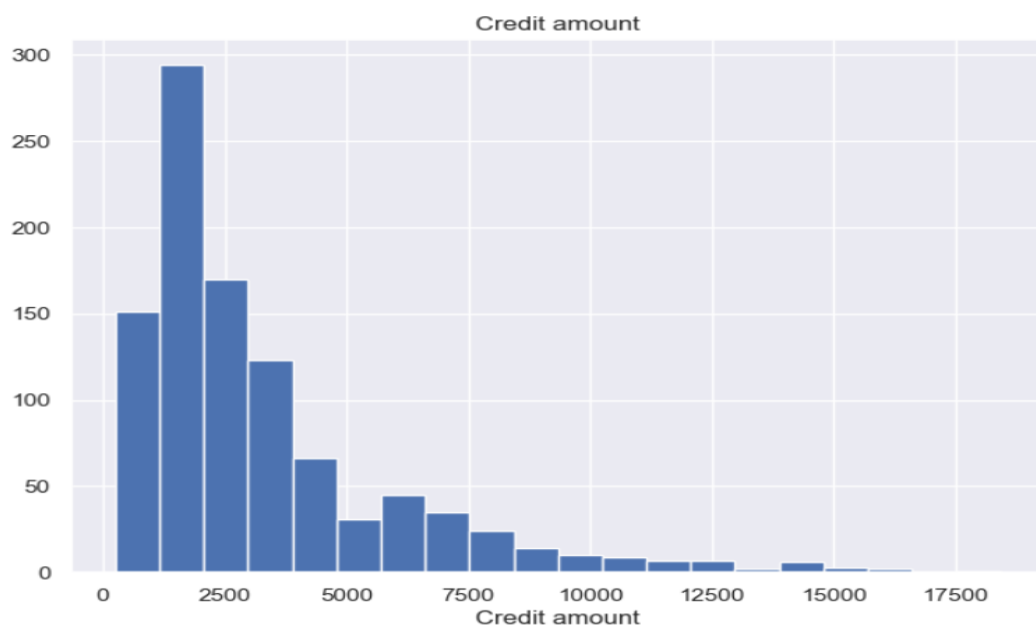
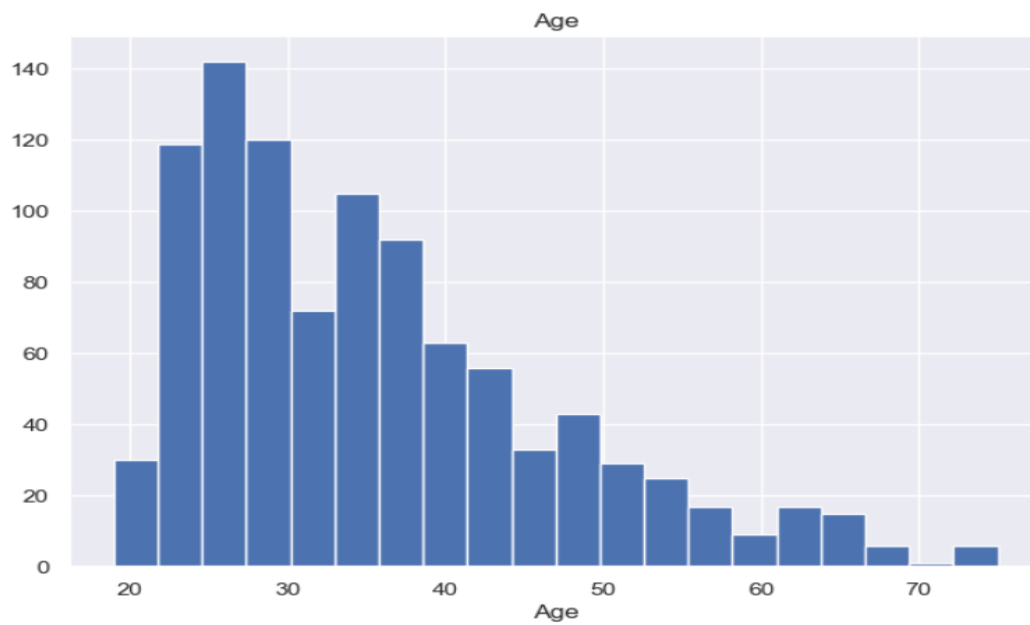
### Categorical Variables:

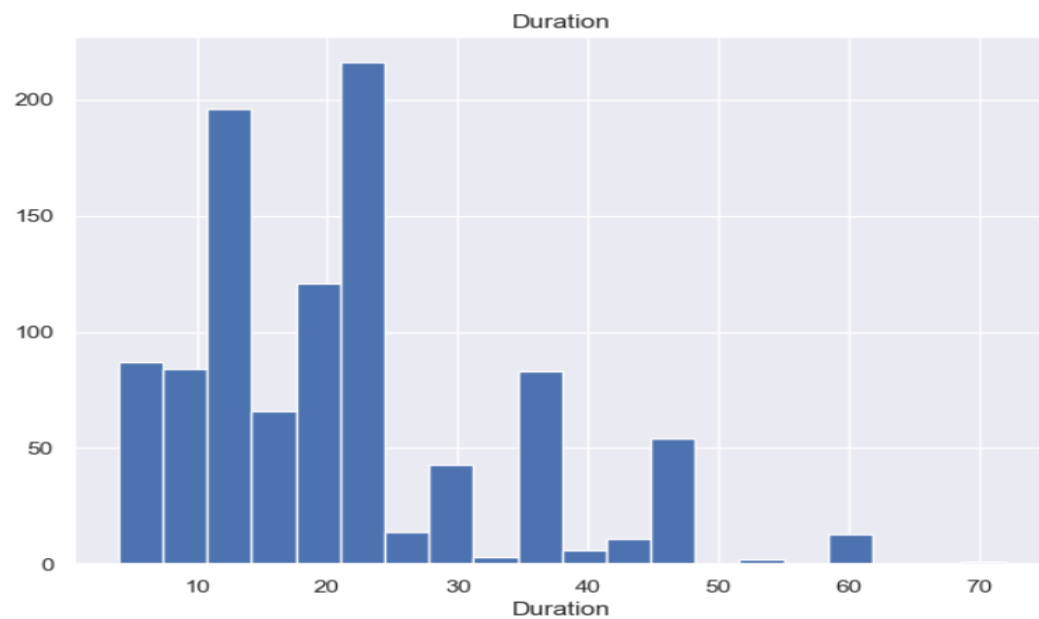




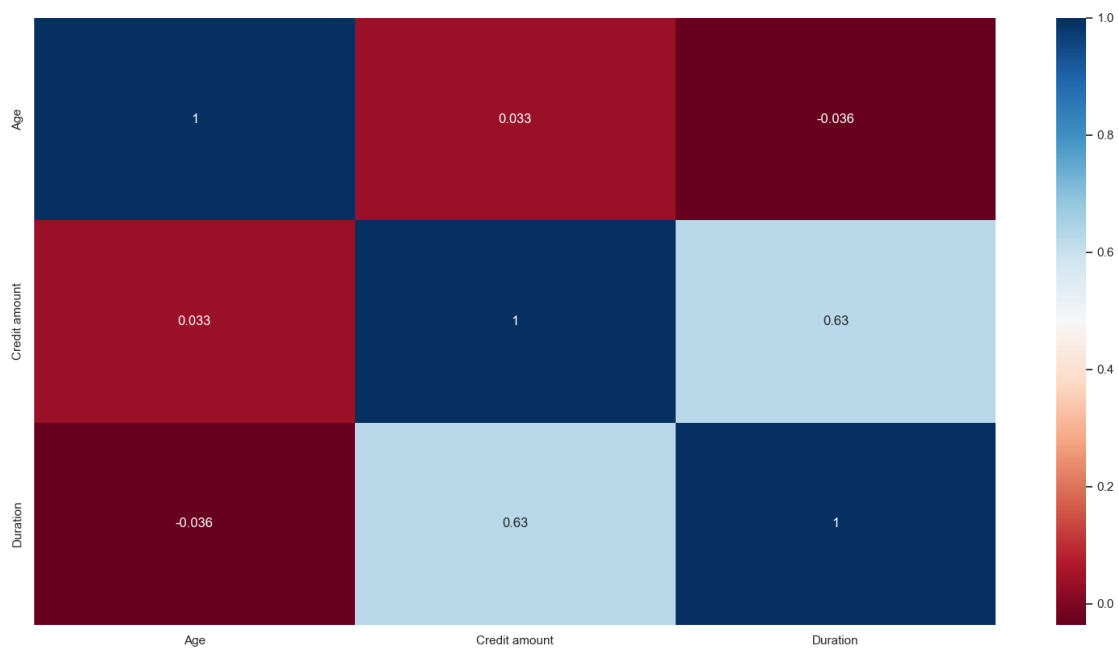


## Numerical Variables:





Corelation Matrix:



## Importance and Use of Data

This dataset serves as the fundamental data source for training machine learning algorithms and predicting credit risk. In the later stages of the study, feature engineering and modeling steps carried out on the dataset aim to achieve more accurate and reliable results for credit risk prediction.

This dataset contains essential information for making financial decisions, and its proper processing and analysis hold critical importance for predicting credit risk accurately.

# Data Preparation

This study encompasses the preparation phase for applying various machine learning algorithms for credit risk prediction. The following steps were undertaken in the data preparation process:

## **Acquisition and Understanding of the Dataset**

At the beginning of the study, a dataset intended for credit risk prediction was obtained. This structured dataset contains characteristics related to credit applications (such as age, gender, income status, credit amount, previous payment history, etc.). The size, features, and general characteristics of the dataset were thoroughly examined.

## **Data Cleansing and Preprocessing**

Cleansing operations were performed on the dataset. Missing data points, outliers, and inconsistent data were identified and rectified by applying necessary procedures to clean the dataset. Additionally, data formats and encodings were checked, and transformations were applied as needed.

## **Feature Selection and Identification of Significant Features**

An analysis of the dataset's features was conducted to identify features that could be influential in credit risk prediction. Features were determined based on correlation analysis and the distribution of features to ascertain which ones were more decisive for the model and predictive performance.

## **Processing of Categorical Variables**

Categorical variables were transformed into a format suitable for modeling. Techniques like one-hot encoding or label encoding were used to convert categorical variables into numerical values, making them compatible with the model.

## **Scaling of Numerical Variables**

The scales of numerical variables in the dataset were standardized or normalized. This process ensured balance among variables with different scales, enabling more consistent learning by the models.

## **Outcome of Data Preparation**

The data preparation process resulted in the creation of a prepared dataset suitable for model training. The steps involved in cleaning, transforming, and feature selection facilitated better learning by the model, leading to more accurate predictions for credit risk.

# Feature Engineering

This study involves feature engineering steps aimed at processing, transforming, and adapting features in the dataset for the modeling process. The following steps were undertaken:

## **Selection of Features and Identification of Significant Ones**

An analysis of dataset features was conducted to determine which features were more influential for credit risk prediction. Through correlation analysis, feature variance, and their relationship with the target variable, significant features were identified and incorporated into the modeling process.

## **Processing of Categorical Variables**

Categorical variables were prepared for the modeling process. Techniques like one-hot encoding or label encoding were applied to numericalize categorical variables and integrate them into the model.

## **Creation of New Features**

New features were generated based on existing features. For instance, a new feature such as the ratio of income to credit amount was created, providing the model with more information. These new features aided the model in making more accurate credit risk predictions.

## **Transformation of Numerical Variables**

The distributions or scales of numerical variables were adjusted for the model. Particularly, techniques like normalization or standardization were employed to ensure that variables were on the same scale.

## **Outcome of Feature Engineering**

The feature engineering steps ensured that features in the dataset were made compatible with the modeling process. This process contributed to better model learning and enabled more accurate credit risk predictions.

Some new feature we added in our Project:

```

"NEW_monthly_repayment" = "Credit amount" / "Duration"

"NEW_Age*Job" = "Age" * "Job"

"NEW_Housing*Job" =
    "Housing" = "free" -> 1
    "Housing" = "rent" -> 2    * ("Job" + 1)
    "Housing" = "own" -> 3

"NEW_CatAge" =
    0 < "Age" < 25 -> "Young"
    25 < "Age" < 40 -> "Adult"
    40 < "Age" < 65 -> "Middle-Age"
    65 < "Age" < 100 -> "Old"

```

## Model Selection and Application

This stage involved selecting, training, and evaluating different machine learning algorithms suitable for credit risk prediction. The following steps were taken:

### Selection of Models

Various machine learning models were attempted for credit risk prediction within the scope of this study. These included algorithms like K-Nearest Neighbors (KNN), Support Vector Classification (SVC), Neural Networks, Decision Tree, Random Forest, Gradient Boosting Machines (GBM), XGBoost, LightGBM, and CatBoost. Each algorithm was analyzed for its specific advantages, disadvantages, and performance in credit risk prediction.

### Training and Validation of Models

The selected models were trained on the training dataset and then evaluated using a validation dataset to assess their performance. Model parameters were tuned, and hyperparameter optimization was performed. Additionally, the performance of each model was evaluated using metrics such as accuracy, precision, recall, and F1 score.

### Identification and Application of the Best Model

Based on the evaluations, the best-performing model or models were identified. Models like XGBoost and LightGBM exhibited higher accuracy and overall performance in credit risk prediction compared to others and were subsequently selected for application on real-world data.

## Application of the Model and Performance Analysis

The chosen models were applied to real-world data. The success of predictions made with real data and the practical performance of the model were evaluated. The model's performance on real-world data was analyzed in terms of reliability.

## Results and Importance of Model Selection

The model selection and application were critical steps in achieving accurate and reliable results in credit risk prediction. This stage was essential to determine which algorithms were more suitable and to analyze their effectiveness when applied to real datasets.

KNN Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.715	0.731
test_f1	0.346	0.361
test_roc_auc	0.641	0.662
test_precision	0.632	0.791
test_recall	0.241	0.237

Support Vector Classification (SVC) Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.714	0.718
test_f1	0.227	0.246
test_roc_auc	0.676	0.670
test_precision	0.610	0.630
test_recall	0.143	0.156

Neural Networks Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.692	0.702
test_f1	0.368	0.289
test_roc_auc	0.666	0.665
test_precision	0.483	0.519
test_recall	0.303	0.203

Decision Tree Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.610	0.709
test_f1	0.363	0.264
test_roc_auc	0.542	0.613
test_precision	0.355	0.551
test_recall	0.373	0.183

Random Forest Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.703	0.723
test_f1	0.354	0.309
test_roc_auc	0.653	0.675
test_precision	0.511	0.625
test_recall	0.273	0.210



Gradient Boosting Machines (GBM) Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.718	0.726
test_f1	0.292	0.349
test_roc_auc	0.664	0.706
test_precision	0.623	0.669
test_recall	0.196	0.237

XGBoost Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.714	0.752
test_f1	0.351	0.416
test_roc_auc	0.679	0.706
test_precision	0.606	0.669
test_recall	0.236	0.315

LightGBM Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.731	0.794
test_f1	0.351	0.486
test_roc_auc	0.698	0.715
test_precision	0.634	0.697
test_recall	0.278	0.368

CatBoost Results:

	before hyper parameter analysis	after hyper parameter analysis
test_accuracy	0.707	0.721
test_f1	0.338	0.351
test_roc_auc	0.675	0.698
test_precision	0.526	0.634
test_recall	0.253	0.278

# Conclusion

This study, aimed at applying machine learning models in credit risk prediction and evaluating their performance, has yielded the following summarized findings:

## Model Performance

Various machine learning algorithms were employed to create and evaluate credit risk prediction models within this study. Evaluations based on metrics like accuracy, precision, recall, and F1 score revealed that XGBoost and LightGBM models outperformed others. These models handled complexities in the dataset better, resulting in more accurate predictions. In this paper, the credit risk prediction using machine learning is studied. By comparing the model discrimination, model interpretability and model stability of logistic regression model and XGBoost model, it can be seen that the model discrimination and model stability of XGBoost model are significantly higher than that of logistic regression model, which can effectively improve the identification ability of personal fast credit risk.

## Key Findings

Detailed analyses were conducted on the features of the best-performing models. Particularly, the influence of features like income status, payment history, and credit amount in credit risk prediction was examined. Additionally, the study explored scenarios where the model made accurate predictions and situations where its performance was comparatively lower.

## Implications for Application

Important insights into the practical application of the obtained models were gained. Emphasis was placed on how these models could be utilized for credit risk assessment in financial institutions to make more robust decisions. Demonstrating successful performance on real-world data, the models could strengthen financial institutions' risk management processes.

## Limitations of the Study

Several limitations were encountered during the study process. Specifically, factors like missing specific attributes in the used dataset or having limited data from certain periods could affect the generalizability of the results obtained. Moreover, instances were observed where the model exhibited lower performance under specific conditions.

## Future Directions and Recommendations

To provide guidance for future studies, recommendations have been put forth. Particularly, suggestions such as working with more comprehensive datasets, incorporating different attributes and time intervals, could contribute to creating more robust models for credit risk prediction.

## Importance of the Results Section

The results section plays a crucial role in showcasing the study's value and the significance of the obtained findings. It emphasizes how the findings can contribute to making more reliable decisions in credit risk prediction and ensuring financial stability.

## References

The sources and cited studies used during the study process are listed below:

- Li, Y. (2019, August). Credit risk prediction based on machine learning methods. In 2019 14th International Conference on Computer Science & Education (ICCSE) (pp. 1011-1013). IEEE.
- Khemakhem, S., & Boujelbene, Y. (2015). Credit risk prediction: A comparative study between discriminant analysis and the neural network approach. *Accounting and Management Information Systems*, 14(1), 60.
- Chen, S., Wang, Q., & Liu, S. (2019, June). Credit risk prediction in peer-to-peer lending with ensemble learning framework. In 2019 Chinese Control and Decision Conference (CCDC) (pp. 4373-4377). IEEE.
- Lin, T. C. (2019). Artificial intelligence, finance, and the law. *Fordham L. Rev.*, 88, 531.
- Naresh Kumar, M., & Sree Hari Rao, V. (2015). A new methodology for estimating internal credit risk and bankruptcy prediction under Basel II Regime. *Computational Economics*, 46(1), 83-102.
- Hand, D., H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA. 2001.
- Shan Liang, Qiao Yang. *Dataing Risk Control-Credit Score Modeling Course*, Electronic Industry Press. 2018
- Zhou Zhihua. *Machine learning [D]*. Tsinghua University Press. 2015.

These resources encompass studies and research conducted on credit risk prediction, machine learning methodologies, credit assessment processes, and significant works in the domain of financial decision-making. These references form the basis of the study and provide support within the existing literature on the subject.