ABSTRACT:

In nowadays, with the development of technology, credit risk is on of the main risk currently faced by commercial bank. The significant problems experienced by banks have highlighted the critical importance of measuring and providing for credit risk. In this project, we focused on assessing credit risk using machine learning models such as Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Machines and KNN. Feature engineering techniques, including Weight of Evidence analysis, are utilized to identify and transform informative variables, improving model performance and interpretability. The study evaluates the predictive performance of each model using metrics such as accuracy, precision, recall, and F1-score. The findings of this project contribute to advancing credit risk analysis methodologies, providing insights into the effectiveness of machine learning approaches in mitigating credit risk and informing lending decisions.

Keywords: German Credit Dataset, Logistic Regression, Machine learning, WOE

1. **INTRODUCTION**

Credit risk analysis plays a pivotal role in the financial industry, serving as the cornerstone of prudent lending practices and risk management strategies. With the proliferation of credit transactions and the dynamic nature of financial markets, accurately assessing the likelihood of borrowers defaulting on their obligations has become increasingly vital for safeguarding the stability and profitability of lending institutions.

This paper embarks on a comprehensive exploration of credit risk analysis, focusing on the prediction of creditworthiness by categorizing borrowers into "good" or "bad" risk categories. The German Credit Risk Dataset serves as a fundamental resource for studying and understanding the intricacies of credit risk analysis. In the realm of finance, assessing the creditworthiness of borrowers is paramount for ensuring the stability and profitability of lending institutions. The dataset encapsulates a comprehensive array of borrower attributes, spanning demographic information, credit history, loan duration, and purpose.

By leveraging machine learning techniques and predictive analytics, this project seeks to delve deep into the dataset's intricacies to discern patterns and relationships that underlie credit risk. Through rigorous analysis and model development, the objective is to construct robust predictive models capable of accurately predicting the likelihood of bad risk for the bank. Such models hold the potential to revolutionize credit risk assessment practices, providing financial institutions with invaluable insights for making informed lending decisions while mitigating risk.

2. THEORETICAL BACKGROUND

2.1. Models

2.1.1. Logistics Regression

      Logistic Regression, a fundamental machine learning technique frequently employed for binary classification tasks, distinguishes itself from linear regression by forecasting the likelihood of an instance belonging to a specific class rather than predicting continuous values.

      In logistic regression, the output is transformed using the logistic function (or sigmoid function) to ensure that the output falls between 0 and 1, representing probabilities.

The form of the model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

      Logistic Regression is commonly applied in credit risk analysis for assessing the likelihood of a probability having bad credit risk or being unable to meet their financial obligations. It provides a transparent and interpretable framework for assessing credit risk by quantifying the probability of bad risk based on borrower characteristics.

2.1.2. Decision Trees

      A Decision Tree is a supervised machine learning algorithm that is used for both classification and regression tasks. At each step of the partitioning process, the Decision Tree selects the feature that best splits the data into homogenous subsets, typically using metrics like Gini impurity or information gain.

      Decision Trees are highly interpretable, as they can be visualized as a tree structure where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the final prediction or classification. This interpretability makes Decision Trees particularly useful in domains where understanding the reasoning behind predictions is important, such as credit risk assessment.

      In credit risk analysis, Decision Trees can be used to assess the creditworthiness of borrowers by evaluating their financial history, employment status, income level, and other relevant factors. By analyzing these features, Decision Trees can classify borrowers into different risk categories, helping lenders make informed decisions about loan approval or rejection.

2.1.3. Random Forest

Random forest is a machine learning algorithm that creates an ensemble of multiple decision trees to reach a singular, more accurate prediction or result. The "forest" it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result.

2.1.4. Gradient Boosted Machines

Gradient Boosted Machines (GBMs) are a class of machine learning algorithms that are widely used for both regression and classification tasks. They are based on the idea of sequentially training weak learners, typically decision trees, to correct the errors made by previous learners.

Numerous gradient boosting models enjoy widespread adoption, including XGBoost and CatBoost. XGBoost, short for Extreme Gradient Boosting, has gained popularity for predictive tasks since its inception by associate professor Chen at Carnegie Mellon University, USA, in 2016. In contrast, CatBoost, also referred to as Category Boosting, represents a more recent addition to the gradient boosting repertoire. This algorithm implements a strategy of grouping potential feature-split pairs at the leaf and opting for the split with minimal penalty, harnessing the benefits of balanced trees to expedite computation and mitigate overfitting.

2.1.5. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) presents itself as a viable approach for risk credit analysis due to its simplicity and adaptability. In the context of credit risk assessment, KNN can be utilized to classify borrowers into different risk categories based on the similarity of their credit profiles to those of known borrowers with labeled credit outcomes. By considering the attributes of nearest neighbors, KNN provides a straightforward mechanism for determining the creditworthiness of potential borrowers. Furthermore, KNN's ability to handle both numerical and categorical data makes it suitable for analyzing diverse sets of borrower characteristics typically present in credit datasets. However, challenges such as determining the optimal value of K and managing computational complexity may arise when applying KNN to large-scale credit datasets. Nonetheless, with proper parameter tuning and preprocessing techniques, KNN can serve as a valuable tool for risk credit analysis, aiding financial institutions in making informed lending decisions while effectively managing credit risk.

2.2. WeightedW of Evidence (WOE)

In the context of credit risk analysis, Weight of Evidence (WoE) is a statistical technique used to assess the strength of the relationship between a predictor variable (e.g., a borrower's credit history) and the target variable (e.g.,Risk). The WoE is calculated by comparing the distribution of the predictor variable for different levels of the target variable (e.g., Risk good vs. Risk bad).

The formula for WOE is:

$$WOE = \ln\left(\frac{Distribution\ of\ Goods}{Distribution\ of\ Bads}\right)$$

When dealing with continuous variables, WOE involves partitioning the variable's range into intervals and computing the natural logarithm of the ratio between the distribution of good outcomes (such as Risk good) and bad outcomes (like Risk bad) within each interval. This calculation yields a measure of how well each interval predicts the event of interest. Positive WOE values indicate higher likelihoods of the event, negative values suggest lower likelihoods, and zero signifies no predictive power.

When applied to categorical variables, WOE involves calculating the natural logarithm of the ratio between the proportion of good outcomes and the proportion of bad outcomes within each category or level of the variable. This process quantifies the predictive power of each category in relation to the event of interest. Positive WOE values indicate categories associated with a higher likelihood of the event, negative values signify lower likelihoods, and zero suggests no predictive power

One of the main advantages of WOE is its ability to handle both continuous and categorical variables, making it versatile for various types of data commonly encountered in credit analysis. Additionally, WOE provides a straightforward interpretation of the relationship between predictor variables and the target variable by quantifying the strength of association through natural logarithms. This interpretability allows stakeholders to understand the impact of each variable on the risk prediction, aiding in decision-making processes. Moreover, WOE is robust to outliers and missing values since it relies on binning techniques, which can enhance the stability of the analysis and mitigate the influence of extreme values.

However, WOE analysis also has its limitations. One disadvantage is the potential loss of information during the binning process, particularly for continuous variables. Dividing continuous variables into discrete intervals may lead to information loss and reduced model sensitivity to subtle changes in the data. Additionally, the choice of the number and width of bins can introduce subjectivity into the analysis, affecting the robustness and generalizability of the results. Furthermore, WOE may not capture

complex nonlinear relationships between variables, limiting its predictive power compared to more flexible modeling techniques. Lastly, the interpretation of WOE values may be challenging when dealing with a large number of predictor variables, as it requires careful consideration of multiple factors simultaneously.

By applying Weight of Evidence analysis, lenders and risk analysts can gain insights into which borrower characteristics are most strongly associated with credit risk, helping them make more informed decisions about loan approval and risk management.

3. DATA

3.1. Dataset

The German Credit Dataset stands as a foundational resource in the field of credit risk analysis. Compiled by Professor Dr. Hans Hofmann of the University of Hamburg, this dataset provides a snapshot of credit applicants and their associated attributes and available from the UCI Machine Learning Repository [UCI Machine Learning Repository]. The original dataset contains 1000 entries with 20 categorial/symbolic attributes ranging from demographic information to financial metrics, the dataset serves as a cornerstone for developing predictive models and informing lending decisions. We used a smaller sub dataset from Kaggle. An overview over the dataset's characteristics is available in this Table 2 below.

| Attribute	Type	Details
Age	numeric	Client's age
Sex	character	male, female
Job	numeric	0: unskilled and non-resident, 1: unskilled and resident, 2: skilled, 3: highly skilled
Housing	character	own, rent, free
Saving account	character	little, moderate, rich, quite rich, unknown
Checking account	character	little, moderate, rich, unknown
Duration	numeric	loan terms (month)
Purpose	character	car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others
Credit amount	numeric	**Outcome**, credit limit (DM)
Risk	character	**Outcome**, good, bad

Table 1: Dataset's characteristics

The primary target variable in the dataset is the "Risk" attribute, which indicates whether a credit applicant is classified as a "good" or "bad" credit risk. This attribute is typically used for binary classification tasks.

3.2. Exploratory Data Analysis

In this section, I present the results of EDA on the German Credit Dataset, focusing on summarizing key attributes and visualizing data distributions.

A crucial aspect of this analysis involves understanding the target variable, typically labeled "Risk", which categorizes individuals' risk into "good" or "bad" credit risks. With Good Credit Risk, these applicants are deemed more likely to repay their loans according to the terms. They might be labeled as "Good"; with Bad Credit Risk, these applicants are considered less likely to repay their loans in full or on time so they might be labeled as "Bad". Firstly, I will explore its distribution through visualizations with bar charts, revealing the proportion of applicants falling into each credit risk category. It showed the distribution of Risk with 700 good risk and 300 bad risk. Secondly, cross-tabulations will be constructed to identify potential initial patterns

between the target variable and other features in the dataset. For instance, I analyzed how credit risk is distributed across different genders, age groups, or loan purposes. This initial investigation into the target variable lays the groundwork for further exploration of how various factors might influence creditworthiness.
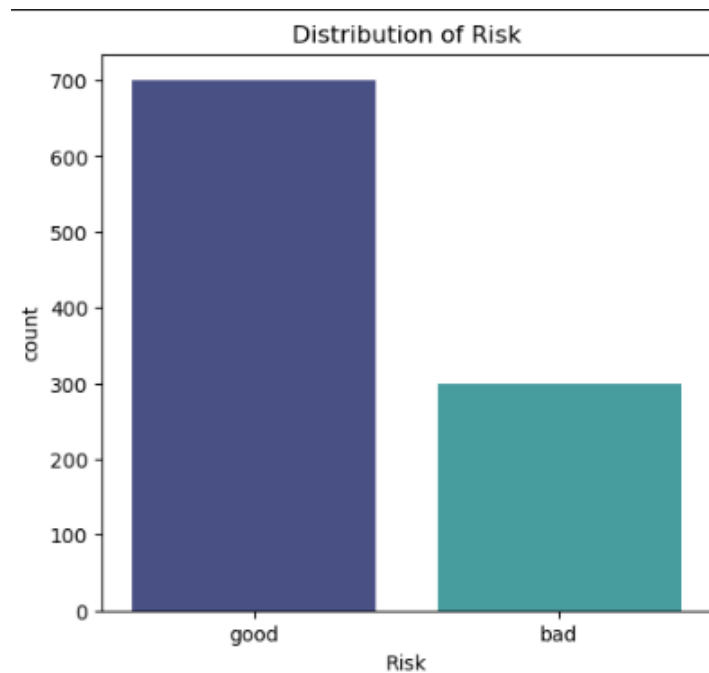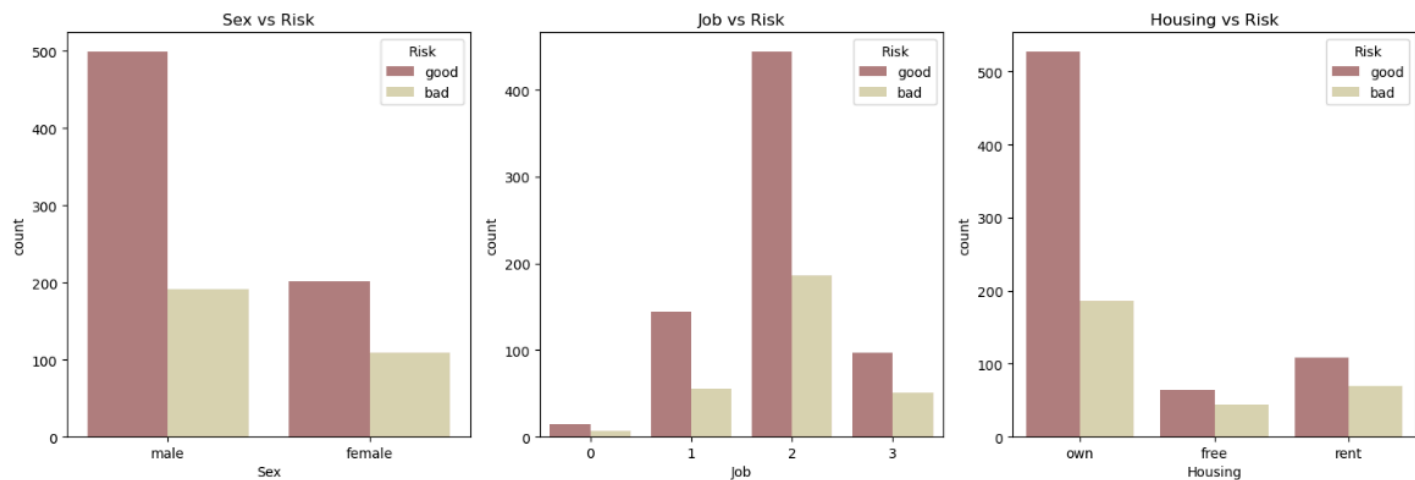

Figure 1: Distribution of Risk

The age histogram displays that the highest age distribution occurs above 20 years old and under 30, then decreases significantly with increasing age. Below the age of 20, the data frequency is very low or rare. Based on the data collected from the Duration feature, it can be seen that the majority of the data falls within the range of 10 - 30 months. As we know, the longer the customer's loan duration, the greater the profit for the bank. Therefore, our target is to analyze customers who fall into the category of ≥ 30 months. Lastly, based on the data collected on the Credit_Amount, it shows that the highest distribution is observed for loan amounts ranging from 1,250 to 2,500 credit. Then, the distribution significantly decreases as the loan amount increases.
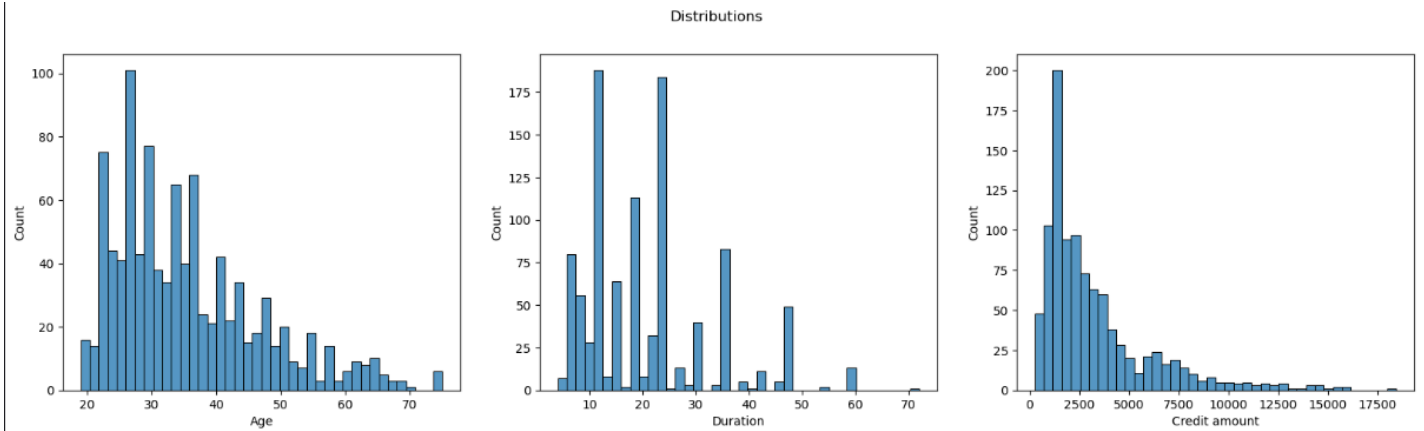
Figure 2: Distribution of Age, Duration and Credit amount

Visualizing the distribution of applicant characteristics across credit risk categories provides valuable insights. By analyzing these count plots, we can see if there are any imbalances or patterns related to creditworthiness. According to these given count plots, it reveals a higher proportion of males with good credit risk compared to females, besides, there is a connection between the amount of job (e.g., 0, 1, 2, 3) and a likelihood of being classified as a bad credit risk. Similarly, the plots indicate relationships between financial situations (savings accounts, checking accounts), and whether specific loan purposes are more commonly associated with good or bad credit risk.
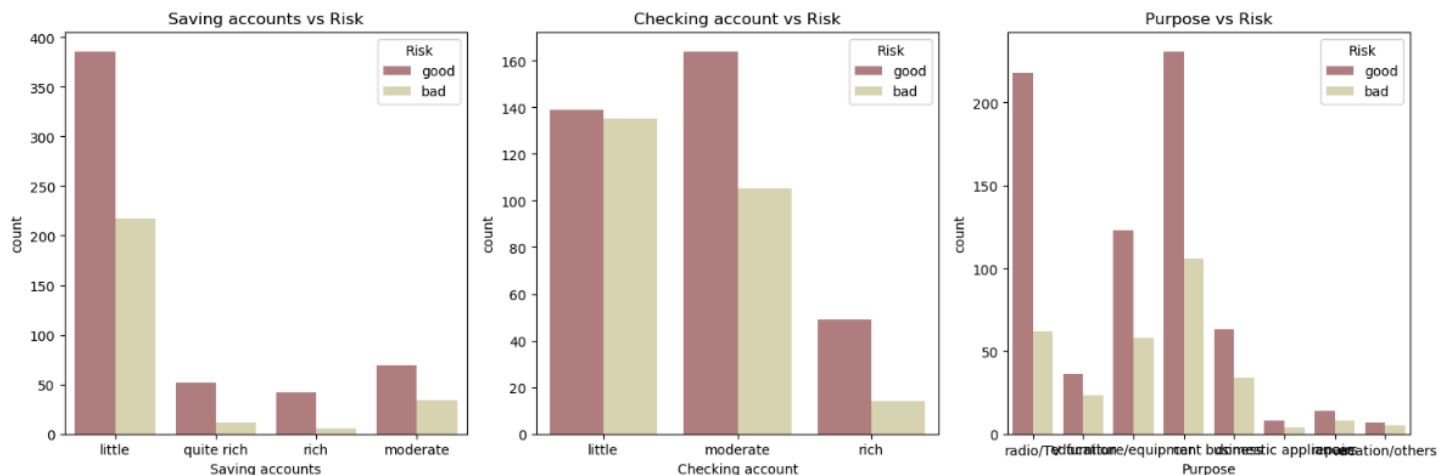
Figure 3: Risk vs Other categorical variables

Analyzing the distribution of age across credit risk categories reveals potential differences between these groups. The boxplot suggests that applicants with good credit risk might be older on average compared to those with bad credit risk. This is evident from the higher median age in the "good" risk category. Additionally, the wider spread of the "good risk" box indicates a larger variation in ages for the good credit risk group. While outliers are present on both sides for both categories, further analysis is needed to confirm the statistical significance of these observations.

I analyzed credit amount by risk category in the German Credit Dataset and it reveals a potential connection between creditworthiness and loan amount. The mean and median credit amounts are both higher for applicants classified as bad credit risks compared to good credit risks. This suggests that on average, and for the typical applicant within each category, those with bad credit risk tend to apply for larger loans. Furthermore, the wider standard deviation for bad credit risk indicates a greater spread in credit amounts requested by this group. The presence of outliers on both ends suggests there might be applicants in both good and bad credit risk groups who deviate significantly from the typical credit amount requested within their category.
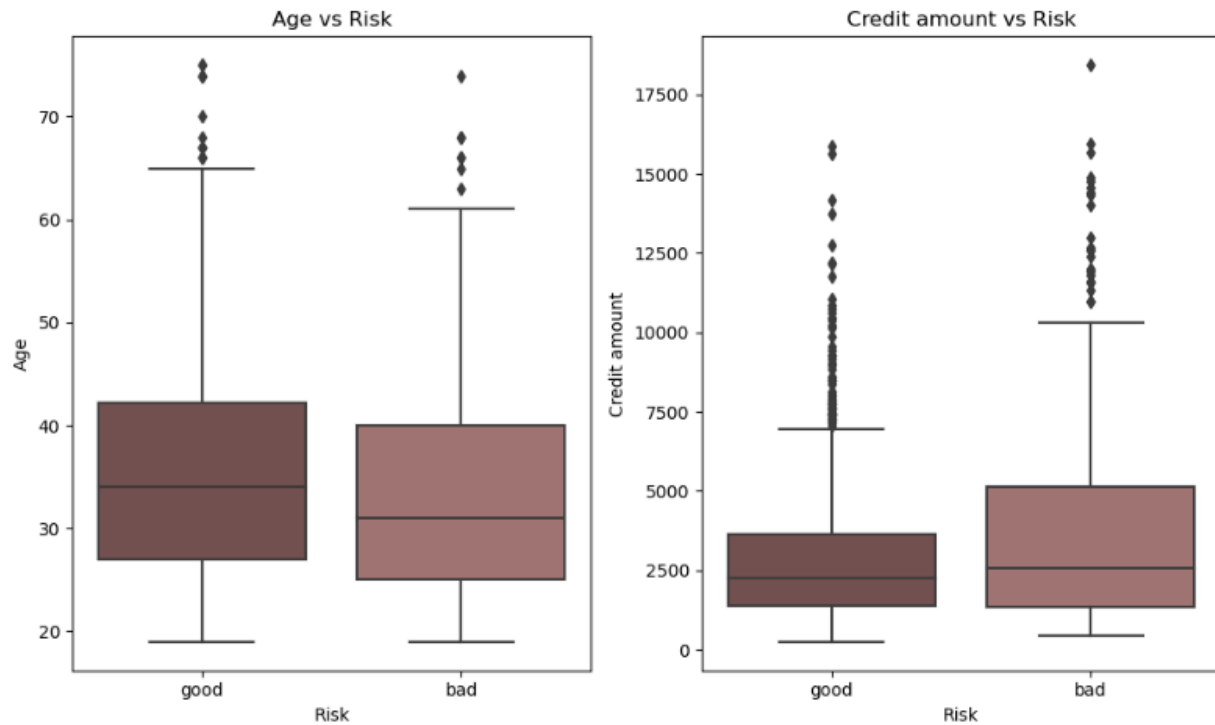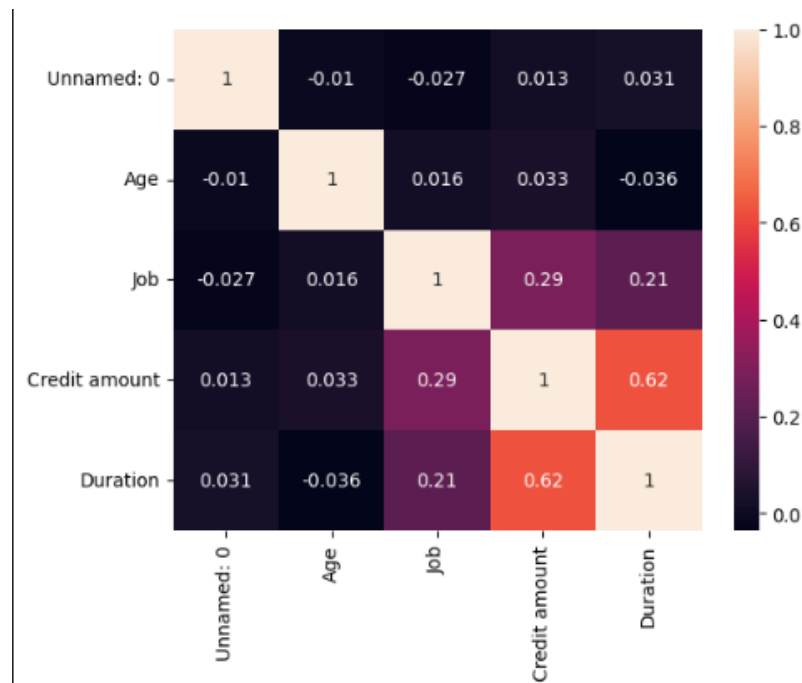
Figure 4:

Calculating correlation between numerical variables is a statistical technique used to assess the strength and direction of the linear relationship between two numerical variables in a dataset. The near-zero correlation between Age and Credit amount (0.0327) suggests that age has little influence on the loan amount applicants request. Similarly, the weak negative correlation between Age and Duration (-0.0361) indicates a negligible tendency for younger applicants to have slightly longer loan terms. In contrast, a strong positive correlation between Credit amount and Duration (0.6250) highlights a clear association. Applicants with larger credit amounts tend to have longer loan durations, possibly due to lenders offering extended repayment periods for bigger loans. It's important to remember that correlation doesn't imply causation, and further exploration might be needed to understand the underlying factors influencing these relationships.

By analyzing the correlation between duration and credit amount further, I gained valuable insights into loan term decisions. The heatmap reveals an interesting initial finding: the only correlation coefficient above 0.50 (indicating a moderately

strong positive correlation) is between duration (loan term) and credit amount. A correlation above 0.5 suggests that borrowers with higher credit amounts tend to be offered (or request) longer loan durations. This might be because lenders see larger loans as riskier and require a longer repayment period to make them more manageable for borrowers.

Figure 5: Heatmap

3.3. Data Processing

3.3.1. Missing values handling

After EDA, we noted that, "Saving accounts" and "Checking account" variables have 20% and 40% missing data in their columns, respectively. For this problem, we will use 3 approaches (mode imputation, constant imputation, or k-NN imputation with cross-validation) is a good starting point for addressing missing values in "Saving accounts" and "Checking account" variables, but the most suitable method depends on

the data type. Trial of all 3 imputation options and their accuracy reports are found below.

Imputer	Accuracy
A constant value or a new category	0.725
Mode	0.7
k-NN Model	0.7

Table 2: Accuracy of 3 imputation options

For 20% Missing Data (Saving Accounts), both mode and constant imputation are potential options here. However, constant imputation might be less ideal due to potential bias. For 40% Missing Data (Checking Accounts), mode imputation becomes less effective. Since the constant imputer yielded the highest accuracy score, both missing data points have been imputed with "Unknown".

3.3.2. Variables encoding

Variable encoding is a crucial step in data preprocessing for machine learning models. One-hot encoding is a popular technique for transforming categorical variables into numerical representations. It works by creating a new binary feature for each category within the original categorical variable. This method is suitable since there are small numbers of possible values for all categorical variables. List of categorical columns to be one-hot encoded includes 'Sex', 'Job', 'Housing', 'Saving accounts', 'Checking account', 'Purpose' columns.

This approach ensures that models can leverage the categorical information present in the data without introducing biases due to ordinal assumptions.

However, there are some disadvantages when applying One-hot Encoding. One-hot encoding creates a new binary feature for each unique category. This can significantly increase the dimensionality of your data, especially for features with many

categories. Therefor, it can lead to increased computational cost and training time for models; potential overfitting, where models become too specific to the training data and perform poorly on unseen data. One-hot encoded features are typically sparse, meaning most entries are zeros. This can affect the performance of certain machine learning algorithms.

3.3.3. Data transformation

One key technique is standardization, achieved using StandardScaler. This process centers and scales the numerical features in the dataset to have a mean of 0 and a standard deviation of 1. By standardizing the training data, we ensure all features contribute equally to the model's learning process. We then use the same scaler to transform the testing data, maintaining consistency across both sets. This data transformation step helps create a well-prepared dataset for machine learning, potentially leading to improved model performance.

3.4. Modeling

Following preprocessing, we divided the dataset into two segments using a 70/30 ratio: a training set comprising 700 rows and a test set with 300 rows. The training set, a substantial dataset, serves as the foundation for training our machine learning model, enabling it to grasp and retain critical features. Conversely, the test set serves as a means to validate the effectiveness of the model's learned insights post-training.

We employed various algorithms—Logistic Regression, Decision Tree, Random Forest, CatBoost, XGBoost, and KNN—to train our model. In Logistic Regression, we additionally incorporated the WOE Binning dataset to gauge performance.

Post-training, we utilized the feature_importances function from the sklearn library to assess and rank attributes within the German Credit Dataset, aiding in predicting model performance. Notably, the analysis revealed that Credit amount, Age, and Duration emerged as the most influential features.
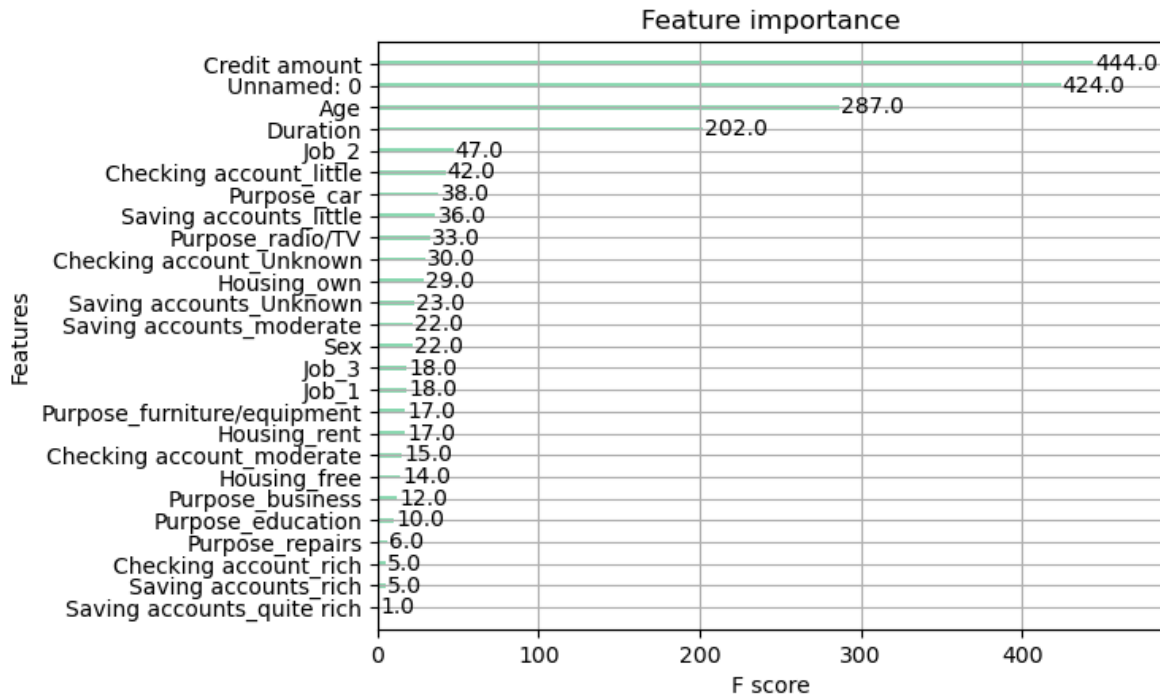
Figure 6: Feature Importance

4. Results

4.1. Metrics

Assessing the performance of a machine learning model is crucial for gauging its effectiveness and pinpointing areas for enhancement. Various standard metrics are employed to evaluate model performance, particularly in classification tasks such as those involving the German Credit dataset. These metrics include Precision rate, Recall rate, F1 value, AUC value, and Accuracy. For this study, the primary evaluation metrics encompass AUC value, F1 value, Precision rate, Accuracy, and Recall rate. Initially, four key components are established: True Positive (TP), indicating cases where the model accurately predicts a positive class; False Negative (FN), signifying instances where the model erroneously predicts a negative class (Type II error); False Positive (FP), representing instances where the model inaccurately predicts a positive class (Type I

error); and True Negative (TN), indicating instances where the model correctly predicts a negative class.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Table 3: Confusion Matrix

From the Confusion Matrix, several metrics can be computed to assess the classifier's predictive performance. It measures the overall correctness of the classifier's predictions across all classes. Accuracy is the ratio of correctly classified instances (true positives and true negatives) to the total number of instances in the dataset.

The formula for accuracy is:

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

In Credit Risk Analysis, identifying Risk is one of the most important to banks. In this regard, recall and precision are two evaluation indicators that hold significant value. recision measures the accuracy of positive predictions made by the classifier. It's the ratio of correctly predicted positive samples (true positives) to the total number of positive predictions made by the classifier (true positives + false positives). Recall, also known as sensitivity or true positive rate, measures the classifier's ability to identify all positive instances. It's the ratio of correctly predicted positive samples (true positives) to the total number of actual positive samples (true positives + false negatives).

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

Picking the best way to judge a classifier can be tricky. While accuracy is popular, it can be misleading, especially for datasets with uneven class sizes. Recall is another option, but it might miss important details. To get a more well-rounded picture, data scientists use the F1-score. This metric considers both a classifier's ability to find the right positives (precision) and its ability to catch all positive cases (recall). By combining these two aspects, the F1-score offers a clearer view of how well a classifier actually performs.

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

When evaluating how well a model separates positive and negative cases, AUC (Area Under the Curve) is a valuable tool. This metric, specific to binary classification, summarizes the model's performance across all possible decision thresholds. Imagine a graph plotting true positive rate (catching the good guys) against false positive rate (mistakenly labeling bad guys as good). A perfect model would have an AUC of 1 (flawless separation), while a random guesser would score 0.5. By providing a single score that considers all thresholds, AUC allows us to easily compare different models and assess their overall effectiveness in distinguishing between positive and negative cases.

4.2. Results

This table summarizes the performance metrics of various models. However, it excludes the logistic regression model trained on the WOE binning dataset. It's important to note that this model achieved an impressive AUC score of 0.9809 on the test set. The values in the table are rounded for easier readability.

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.76	0.72	0.67	0.68	
Decision Tree	0.62	0.55	0.54	0.54	0.54
Random Forest	0.72	0.67	0.63	0.64	0.63
KNN	0.75	0.71	0.63	0.64	
CatBoost	0.75	0.72	0.64	0.66	0.64
XGBoost	0.75	0.70	0.75	0.67	0.662

Table 4: Models and Prediction metrics results

After observing the Logistic Regression Model as the best model in terms of its performance by accuracy, precision, recall, and F1-score metrics, some feature engineering methods are performed to further improve the performance of the Logistic Regression Model. To further improve the Logistic Regression model, three logistic regression models were trained after feature extraction and selection. Surprisingly, the third logistic regression model using only the features: 'Checking account_Unknown', 'Checking account_little', 'Credit_per_duration','Saving accounts_little'. Surprisingly, a model using less data achieved the best accuracy (0.79) among all those tested. This goes against the common assumption that more data is always better. This unexpected result highlights the power of careful feature selection in improving model performance. These findings add valuable insights to credit risk assessment research and underscore the importance of both choosing the right model and strategically selecting features for training.

Model	Accuracy	Precision	Recall	F1-score
Initial Logistic Regression	0.76	0.72	0.67	0.68
LR Model #1 (after feature selection)	0.765	0.73	0.66	0.67
LR Model #2 (after feature selection)	0.775	0.76	0.66	0.67
LR Model #3 (after feature selection)	0.78	0.75	0.68	0.70

Table 5: Model Logistic Regression after Features Selections

4.2. Conclusion and discussions

In conclusion, our analysis of the German Credit dataset delved into various machine learning techniques for credit risk assessment. We meticulously prepared the data by employing data transformation methods such as standardization and explored the performance of multiple models. Among these models, the Logistic Regression classifier emerged as the top performer after Feature Engineering, exhibiting an impressive accuracy of 0.78 and an F1-score of 0.70 on the test set. This finding underscores the significance of employing ensemble learning techniques capable of handling non-linear relationships and minimizing overfitting. Furthermore, our study highlighted the critical role of feature selection in enhancing model performance. We were able to significantly improve the accuracy of the Logistic Regression by feature engineering.

One limitation of the German Credit Dataset analysis is the absence of certain critical variables that could significantly impact credit risk assessment. While the dataset provides extensive information on borrower characteristics, credit history, and loan attributes, it may lack variables related to broader economic factors, regulatory changes, or external market conditions. These external factors can play a crucial role in

influencing credit risk but may not be captured within the dataset. Additionally, the dataset's static nature may pose limitations in capturing temporal trends and dynamics, as it may not reflect changes in borrower behavior or creditworthiness over time. Furthermore, the dataset's size and scope may constrain the complexity and sophistication of the predictive models developed, potentially limiting their predictive accuracy and generalizability to broader populations or different lending environments. Finally, the dataset's representativeness of the broader population of borrowers or specific market segments may be a concern, as it may not fully capture the diversity and heterogeneity present in real-world lending scenarios.

In future work, several avenues of exploration exist to further enhance the predictive capability and applicability of credit risk analysis using the German Credit Dataset. This includes delving deeper into feature engineering techniques to extract more nuanced insights from existing variables and potentially integrating external data sources to enrich the dataset. Additionally, fine-tuning machine learning models through rigorous optimization and exploring ensemble methods could lead to improved predictive performance and generalization ability. Temporal analysis to capture evolving trends over time, enhancing model interpretability, and conducting external validation on diverse datasets are also essential considerations. Furthermore, addressing ethical and fairness considerations and facilitating the deployment and integration of developed models into operational systems within lending institutions are vital for practical implementation. Through these future research endeavors, the field of credit risk analysis stands to benefit from enhanced accuracy, transparency, and fairness in assessing borrower creditworthiness and mitigating risk in lending practices.

Our exploration contributes valuable insights to the field of credit risk assessment, demonstrating the effectiveness of advanced machine learning techniques and underscoring the importance of meticulous feature engineering. Looking ahead, future research could expand upon our findings by investigating the generalizability of these techniques to other credit risk datasets and exploring the impact of incorporating alternative features or model architectures.

References

[Statlog (German Credit Data) - UCI Machine Learning Repository](#)

[Credit risk measurement methodologies](#)

[Random Forest: A Complete Guide for Machine Learning | Built In](#)

[Credit Card Analytics: A Review of Fraud Detection and Risk Assessment Techniques](#)

[Analysis of German Credit Data | STAT 508](#)

[German Credit Risk Classification : modeling and metrics | by Adam Hajjej | The Startup | Medium](#)

[Full article: Credit risk prediction with and without weights of evidence using quantitative learning models](#)

[Identify creditworthy borrowers [ML project]](#)

You can find the code [here](#)