**Strategic Risk Management: Deploying Machine Learning to Predict Credit Defaults in German Banking**

**ABSTRACT**

IIn this project, we tackled the challenge of predicting credit risk for customers at a German Bank using a Random Forest Classifier. Our objective was to identify as many high-risk customers as possible, aiming for a high recall to minimize potential financial losses for the bank. We achieved this by employing strategies like stratified splits and k-fold cross-validation to address data imbalance, along with adjusting the classification threshold to improve the precision-recall trade-off. By applying stratified splits and k-fold cross-validation to handle imbalanced data, and adjusting the classification threshold to balance precision-recall trade-offs, we achieved a recall score of 0.8 without significantly compromising precision, which shifted from 0.51 to 0.49. The model's effective identification of high-risk customers supports the bank's risk management and decision-making processes.

Keywords: German Dataset, Random Forest Classifier, k-fold cross validation, stratified split.

# **INTRODUCTION**

In the financial sector, predictive analytics plays a crucial role in safeguarding against risks and enhancing decision-making processes. Accurate prediction of credit risk is particularly essential in credit lending, as it helps avoid substantial financial losses. With the advent of more sophisticated analytical tools, banks are now equipped to analyze customer data more effectively. This project focuses on utilizing a Random Forest Classifier to predict the likelihood of customers defaulting on loans, a critical aspect of risk management in banking.

The primary objective of this study is to develop a predictive model that improves the identification of 'bad' risk customers at a German Bank. By focusing on maximizing recall, the model aims to minimize the number of high-risk loans that go undetected. Enhancing the bank's understanding of credit risk patterns among its clientele and developing a robust method for dealing with imbalanced datasets are also key goals. These objectives stem from the need to refine how financial institutions predict and manage potential defaults, ensuring that lending practices are both secure and profitable.

The significance of this study extends beyond merely improving predictive accuracy. Achieving a high recall rate in identifying bad risks helps the bank reduce the incidence of loan defaults, thus safeguarding its financial health and reputation. Moreover, the methods and insights derived from this study can be applied to similar risk assessment tasks across the industry, potentially leading to broader improvements in credit risk management practices. This research was conducted using historical loan application data from the bank, ensuring that the findings are relevant and applicable to real-world scenarios. Through this approach, the project contributes to more informed decision-making and supports sustainable lending practices within the banking industry.

# **THEORETICAL BACKGROUND**

## Methodology

### 1.1. Stratified K-Fold Cross-Validation

In our credit risk prediction project, we implemented stratified k-fold cross-validation to ensure the accuracy and generalizability of our model across different subsets of data. Stratified k-fold cross-validation is a method that divides the entire dataset into 'k' smaller, equal-sized subsets, or folds. Each fold is constructed to contain a proportional representation of the target classes, mirroring the overall class distribution in the dataset. This is especially crucial in datasets with imbalanced classes, as it prevents any fold from being biased towards the majority class. The model is then trained on 'k-1' of these folds, while the remaining fold serves as the test set. This process is repeated 'k' times with each fold being used as the test set exactly once. This technique helps in validating the model’s performance reliably, ensuring that our findings are not just a result of a particular random split of the data.

Using stratified k-fold cross-validation allows us to assess the model's stability and effectiveness comprehensively. By repeatedly training and testing the model across different combinations of data, we reduce the likelihood of anomalies in any single fold affecting the overall results. This approach not only improves the confidence in the model’s predictions but also ensures that the model performs well in real-world applications, where it needs to accurately classify new and unseen data. It’s an essential method in our toolkit, providing a robust measure of model performance and helping to avoid overfitting, thus ensuring the model's utility in practical scenarios.

### 1.2. Handling Imbalanced Data

To effectively manage the imbalance in our dataset, we utilized class weighting within our machine learning algorithms. This approach involves adjusting the weights of each class according to their representation in the dataset. Specifically, the minority class—representing 'bad' credit risk—is given a higher weight. This adjustment directly impacts the model’s sensitivity to these crucial but infrequent cases by imposing steeper penalties for misclassifying them. As a result, the model pays more attention to the minority class during training, improving its ability to accurately identify high-risk scenarios.

Contrastingly, the Synthetic Minority Over-sampling Technique (SMOTE) generates synthetic samples to balance the class distribution. While SMOTE can be useful, it creates artificial data points through interpolation between existing minority class samples, which might not always capture the true underlying distributions and can lead to model overfitting. This is why we preferred class weights; they enhance model sensitivity without altering the actual data structure, thus maintaining the integrity of the real-world data and avoiding the potential pitfalls of synthetic sample generation. This method ensures our model is robust and reliable, particularly in its capacity to generalize well to new, unseen data.

### 1.3. K-means Clustering

K-means clustering operates by partitioning the dataset into distinct groups, a process that begins with the selection of the number of clusters. This is typically done using the Elbow Method, which involves plotting the total within-cluster sum of squares against the number of clusters and looking for the 'elbow point.' This point, generally where the rate of decrease in the sum of squares slows down significantly, suggests the optimal number of clusters.

Once the number of clusters is determined, K-means clustering initializes with a set of randomly selected centroids, which are the mean positions of the clusters being formed. Each data point in the dataset is then assigned to the nearest centroid, and the mean of the points in each cluster becomes the new centroid. This process is iterative, with data points reassigned to clusters and centroids recalculated until the positions of the centroids stabilize and no longer change significantly. This iterative process ensures that the clusters formed are as compact and distinct from each other as possible, based on the selected features.

The features chosen for clustering are crucial as they directly influence the formation of clusters. In the context of credit risk analysis, selecting features like age, income level, and credit history can help in forming meaningful customer segments that reflect similar financial behaviors or risk profiles. Each cluster can then be analyzed to understand common traits or patterns, which can inform targeted financial strategies or interventions. This method not only categorizes the customers into manageable groups but also enhances the understanding of underlying patterns in customer behavior within a financial dataset.

## Models

### 2.1. Logistic Regression

Logistic Regression is utilized primarily for binary classification problems. It works by using a logistic function to estimate the probability that a given input belongs to the default class, which in the context of credit risk, is typically the likelihood of a customer defaulting on a loan. The output is a probability value between 0 and 1. By applying a threshold (commonly 0.5), the model determines the class of the input, either as high risk (closer to 1) or low risk (closer to 0). This method is straightforward and efficient in scenarios where the relationship between the independent variables and the log-odds of the dependent variable is linear, making it highly interpretable and a good choice for understanding the effect of various predictors.

### 2.2. Random Forest Classifier

The Random Forest Classifier builds multiple decision trees and merges them together to get a more accurate and stable prediction. The 'forest' it builds is an ensemble of Decision Trees, usually trained with the 'bagging' method. The basic idea is that each tree in the ensemble votes for a class, and the most popular class is chosen as the final output. This process improves accuracy by reducing the model’s variance without substantially increasing bias. Because each tree in the forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set, and by using random subsets of features, Random Forest can model complex interactions within the data that are often missed by other models.

### 2.3. XGBoost

XGBoost stands for eXtreme Gradient Boosting and it is an implementation of gradient boosting machines. The model is built in a stage-wise fashion like other boosting methods but it incorporates several important enhancements. XGBoost improves upon the base Gradient Boosting framework by using more accurate approximations to find the best tree model. While Gradient Boosting builds one tree at a time, XGBoost also parallelizes the process, which speeds up the training considerably. It also includes a regularization term in the loss function to prevent overfitting, which is often not taken into account in traditional Gradient Boosting. This results in performance improvement both in terms of speed and predictive accuracy, particularly on datasets with a large number of noisy features.

### 2.4. Decision Trees

A Decision Tree is a non-linear predictive model that recursively splits the data into subsets based on the value of input features. Each node of the tree represents a feature, which is selected based on how well it separates the classes in the target variable. Branches are created under each node, corresponding to the possible values of that feature, and this process continues until the algorithm reaches a stopping condition defined by the user, such as a maximum depth of the tree. The end nodes, or leaves, represent the outcome or decision reached after considering all input features. Decision Trees are particularly useful for their simplicity and ease of visualization, which aids in understanding how decisions are made. However, they are susceptible to overfitting, especially when the tree is very deep. This can be mitigated by techniques such as pruning, limiting the maximum depth of trees, or requiring a minimum number of samples to split a node.

## Model Evaluation and Selection

### 3.1. Confusion Matrix

The Confusion Matrix is integral for assessing the classification accuracy of our credit risk prediction model. It categorizes the outcomes of model predictions into four distinct groups, providing a clear picture of the model's performance:

* True Positives (TP): Instances where the model accurately predicts high-risk customers.
* True Negatives (TN): Instances where the model accurately predicts low-risk customers.
* False Positives (FP): Instances where the model erroneously identifies low-risk customers as high-risk.
* False Negatives (FN): Instances where the model fails to identify actual high-risk customers.

From these values, we derive several important performance metrics:

* Accuracy: Represents the overall effectiveness of the model in correctly classifying both high-risk and low-risk customers. It is calculated with the formula:
* Precision: Indicates the reliability of the model’s positive (high-risk) predictions. A higher precision means fewer low-risk customers are incorrectly labeled as high-risk:
* Recall (Sensitivity): Measures the model's capability to identify all actual high-risk customers. A higher recall value indicates fewer high-risk customers are overlooked:
* F1 Score: Balances precision and recall in a single metric, useful when equal importance is assigned to both precision and recall:

Understanding these metrics helps us navigate the trade-offs between maximizing the detection of high-risk customers and minimizing incorrect high-risk classifications. The implications of these trade-offs are significant: while False Positives can lead to unnecessary denial of services and potential loss of revenue, False Negatives pose a risk of unexpected financial losses due to undetected default risks.

A thorough evaluation using the Confusion Matrix enables us to fine-tune the model, enhancing its predictive power while aligning it with the bank’s risk management strategies. This methodical approach ensures that the model not only performs with high accuracy but also adheres to the operational and strategic objectives of the bank.

### 3.2. Area Under the Curve (AUC)

The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a crucial indicator of the effectiveness of our credit risk prediction model. The ROC curve depicts the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds, showcasing the model's capacity to differentiate between customers who are likely to default and those who are not. A higher AUC value indicates superior model performance, with 1.0 representing perfect classification and 0.5 equivalent to random chance.

The AUC metric is highly valuable in credit risk assessment because it aggregates the model's performance across all possible classification thresholds into a single measure. This aspect is particularly important for ensuring that the bank effectively identifies likely defaulters while minimizing the rejection of creditworthy customers. The advantage of the AUC is that it is unaffected by any specific decision threshold, which can change based on business needs or market conditions.

Utilizing the AUC allows for an objective comparison between different predictive models, helping to identify which model best balances identifying high-risk customers and avoiding false positives. This metric assists in optimizing the models to suit the bank’s specific risk tolerance and operational goals, facilitating smarter, data-driven decisions in credit management.

### 3.3. Hyperparameter Tuning

In optimizing our Random Forest model, we chose Bayesian Optimization due to its sophisticated approach in handling complex parameter spaces effectively. This method is particularly useful when dealing with models that have interdependent parameters, as it uses a probabilistic model, often a Gaussian process, to estimate the performance of various hyperparameter configurations. The strength of Bayesian Optimization lies in its iterative process, where each step builds upon the previous ones by updating the probabilistic model with the latest evaluation results. This allows the method to refine its predictions continuously and focus the search on the most promising hyperparameter regions, thereby maximizing efficiency.

The use of Bayesian Optimization in our project was crucial for balancing the exploration of new parameter spaces against exploiting known configurations that yield good results. This balance is managed through an acquisition function, which directs the optimization process to areas with the highest expected improvement based on the current model predictions. This strategic exploration is vital in avoiding the computational waste typical of grid searches, where every conceivable combination is evaluated regardless of its likelihood of success.

Moreover, integrating Bayesian Optimization with our specific needs, such as adjusting for class imbalance through class weights, further enhanced our model's performance. By placing greater emphasis on the minority class, high-risk customers, we ensure that our model remains sensitive to this critical group, thereby improving its predictive accuracy and reducing potential financial risks. Additionally, employing stratified K-Fold Cross-Validation within this framework ensured that each fold of the dataset was representative of the overall class distribution, thus maintaining the integrity of the evaluation process. This methodical approach to hyperparameter tuning not only fine-tuned our model to achieve high precision and recall but also ensured that the predictions it generated were robust and reliable, supporting informed decision-making in risk management effectively.

# **DATA**

## Data Description

The dataset utilized in this study is sourced from the UCI Machine Learning Repository and is publicly available on Kaggle. This dataset provides detailed credit data for individuals, aimed at facilitating the analysis and prediction of credit risk based on several demographic and financial attributes.

The dataset comprises data collected on individuals who have taken credit from a bank, with each record detailing various aspects of the individual's financial status and personal demographics. It consists of 1,000 rows, each representing a unique customer, and 10 columns, which correspond to different attributes related to the customer's financial and personal background.Out of these variables, four are numerical: 'Age', 'Credit amount', 'Duration', and 'Checking account balance'. The remaining six variables are categorical, including 'Sex', 'Job', 'Housing', 'Saving accounts', 'Purpose', and 'Risk'.

The target variable within this dataset is 'Risk', which classifies each customer into one of two categories: 'good' or 'bad'. This classification serves as the basis for our predictive modeling, aiming to identify patterns that indicate the likelihood of a customer defaulting on a loan.

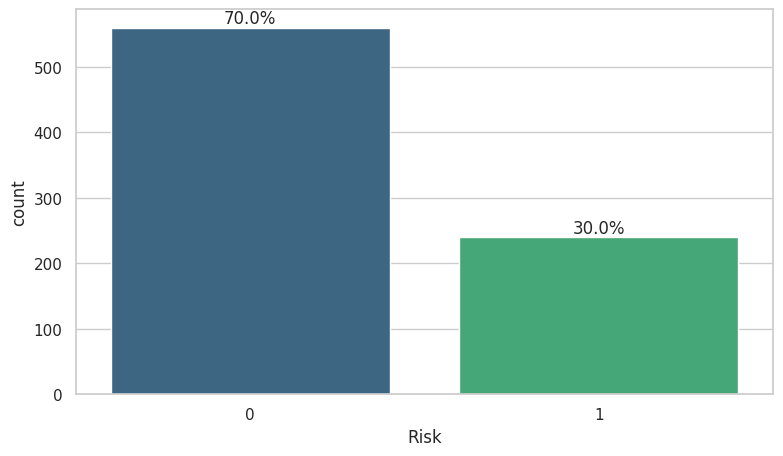
***Table 1. Columns Description***

| **Variables** | **Description** |
| --- | --- |
| Age | Age of the customer in years |
| Sex | Gender of the customer (male, female) |
| Job | Job classification of the customer( 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled) |
| Housing | The housing status of the customer (own, rent, or free) |
| Saving accounts | Indicates the amount of money in the customer’s saving accounts (little, moderate, quite rich, rich)) |
| Checking account | Shows the balance of the individual's checking account in Deutsche Marks (DM), the former German currency |
| Credit amount | Account balance of the customer, measured in Deutsch Marks (DM) |
| Duration | The tenure of the credit amount, in months |
| Purpose | The purpose for which the credit is taken ( car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)) |
| Risk | The risk classification of the customer, based on the likelihood of failing to repay the credit ( good and bad) |

## Exploratory Data Analysis

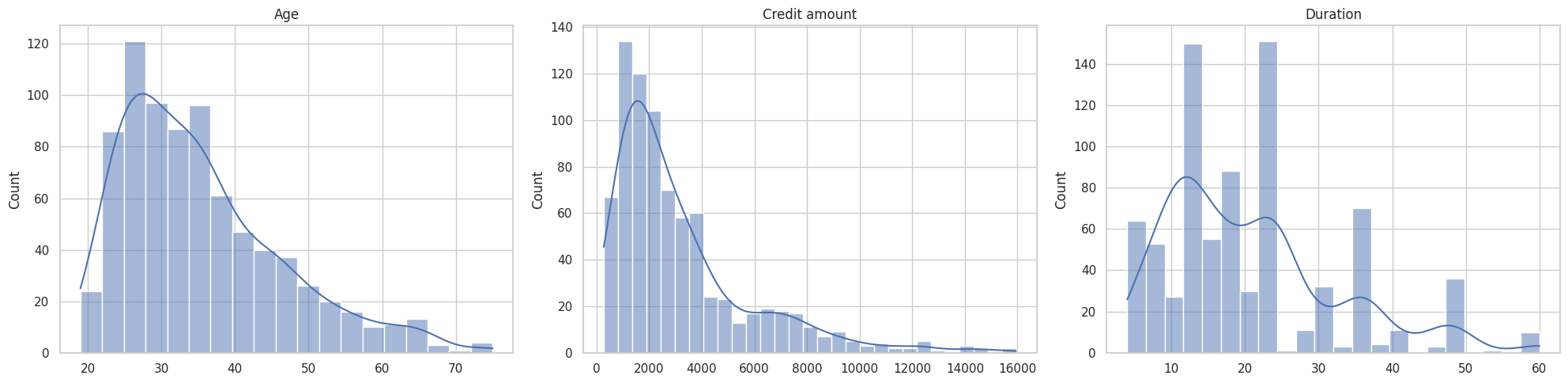
In our analysis, the target variable 'Risk' shows a slight imbalance with 30% of the dataset labeled as 'bad risk'. To address this and ensure accurate model performance, we will implement stratified k-fold cross-validation and adjust class weights to enhance the model's sensitivity to 'bad risk' predictions.

**Figure 1. Distribution of Risk**

****

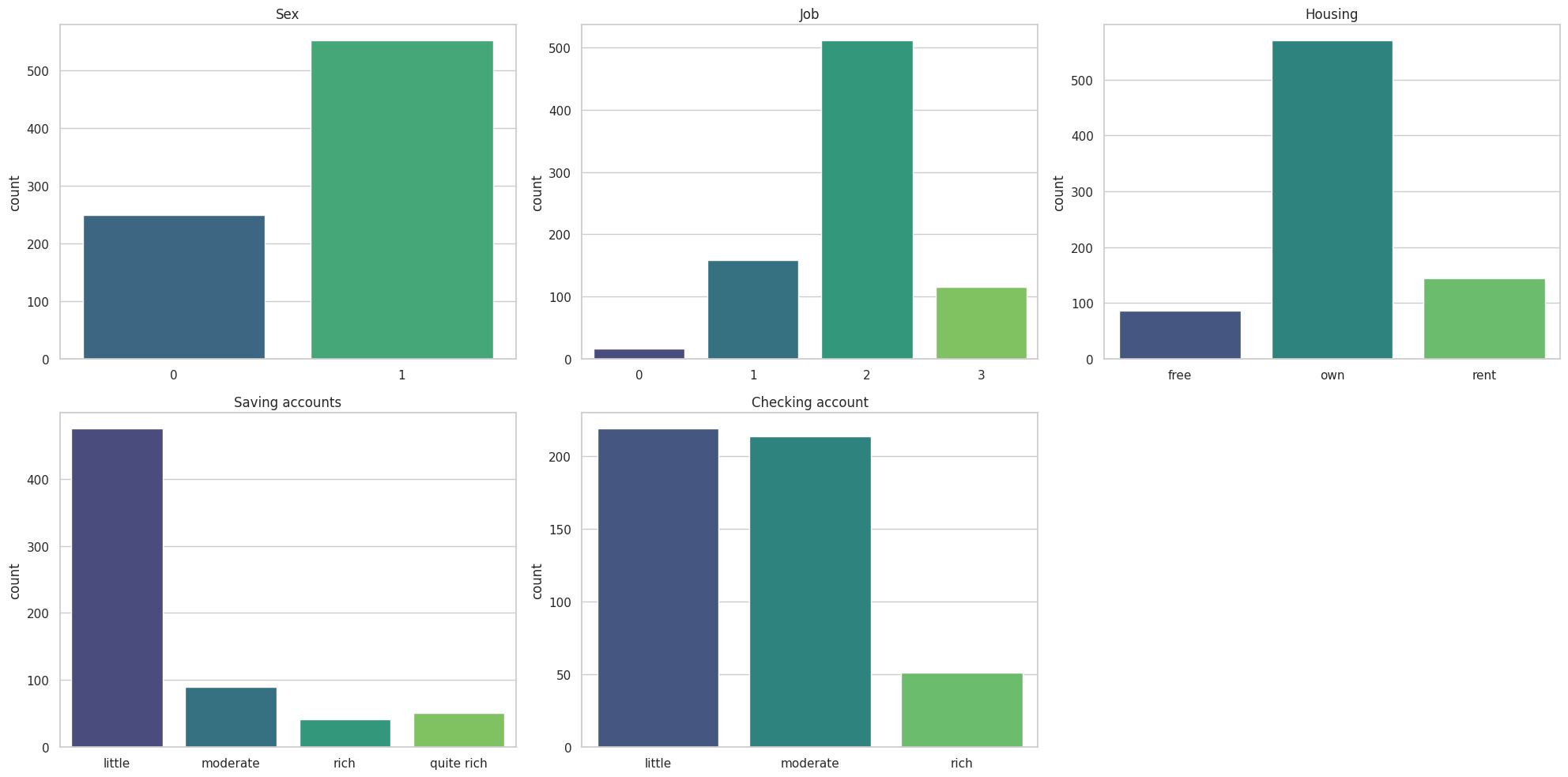
Our analysis highlights that most of the bank's clients are young, primarily aged between 20 and 40 years. The distribution of credit amounts is right-skewed, showing that although most clients opt for smaller loans, there are outliers with significantly higher balances. Additionally, loan duration data indicates a preference for 12 or 24-month terms among many customers, pointing to standardized borrowing behaviors that may influence financial product offerings.

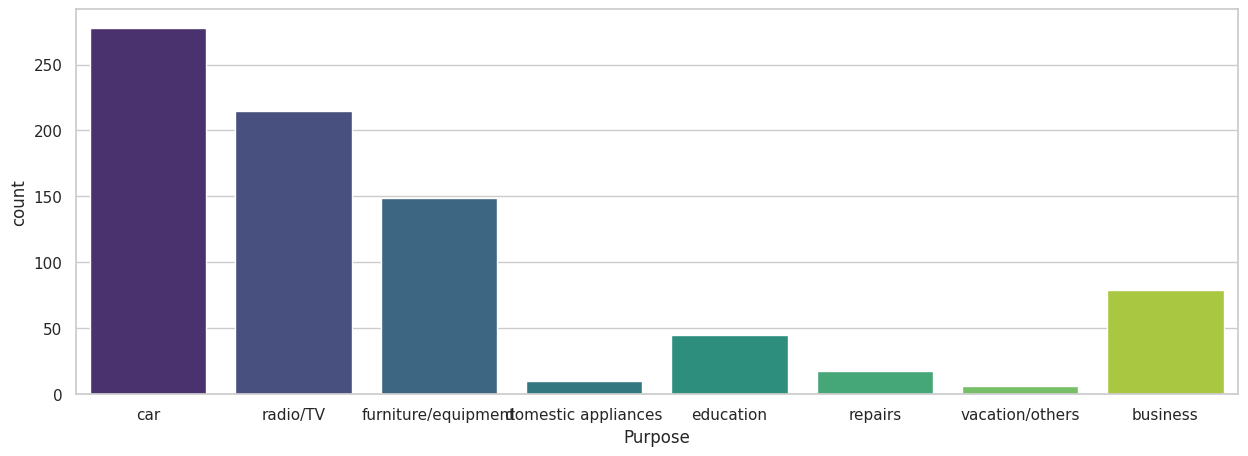
**Figure 2. Distribution of Numerical Features**



We have male customers outnumber female customers by more than two to one, indicating a gender disparity in banking services usage. In terms of employment, the majority of customers are classified as skilled, which may reflect their financial stability and borrowing capacity. Additionally, a significant portion of customers own their homes, providing them with a potential collateral advantage in credit transactions. Financial habits show that most customers have little in their saving accounts, and their checking accounts typically hold little or moderate amounts, suggesting a cautious or limited ability to save. Regarding the purposes for which credits are taken, there is a clear preference for financing cars, radio/TV, and furniture/equipment, highlighting common consumer needs and spending patterns.

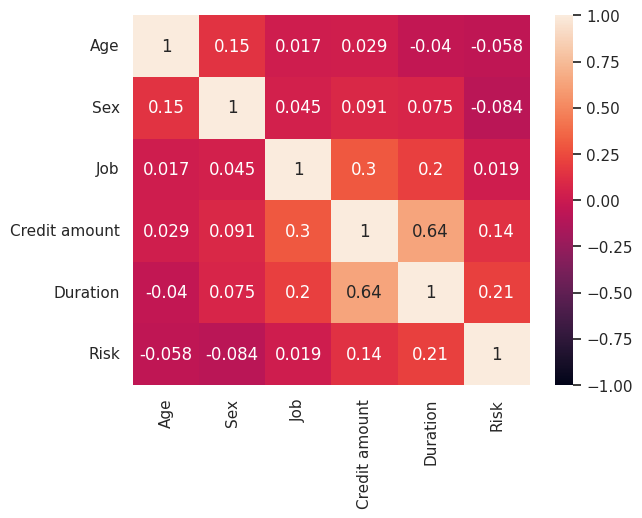
**Figure 3. Distribution of Categorical Features**

****

****

In addition, we also use the correlation matrix to determine the linear relationship between features in the data, especially with the target variable.

**Figure 4. Correlation Matrix**

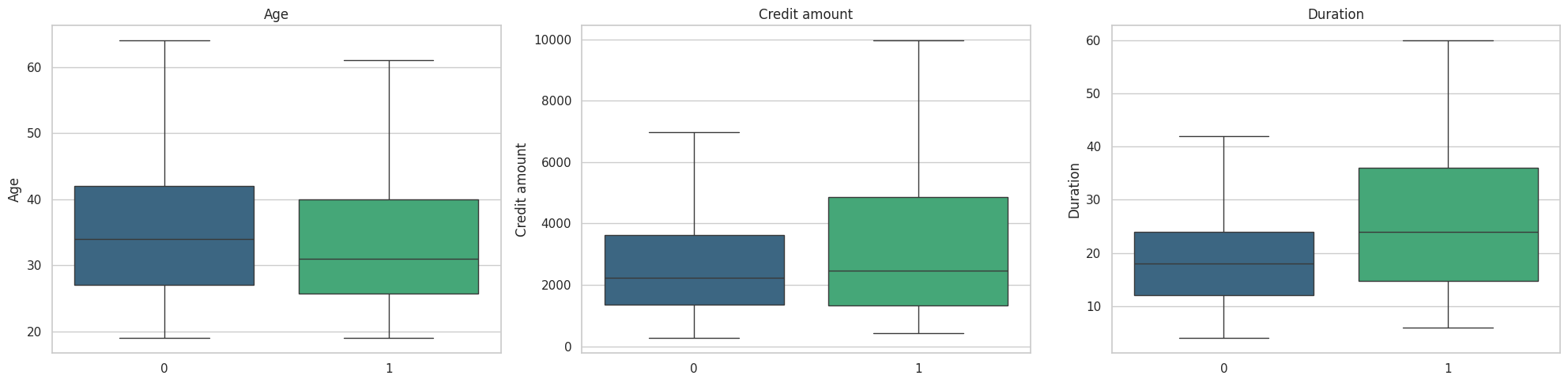
****

Credit amount and duration show a slight positive correlation with the target. This relationship suggests that higher credit amounts and longer loan durations typically imply greater risk, while smaller amounts and shorter terms indicate lower risk.

Duration and credit amount exhibit a strong positive correlation. Generally, longer loan durations are associated with higher loan amounts, while shorter durations correspond to lower amounts. This is intuitive as larger loans typically require more time to repay, whereas smaller loans can be paid off more quickly.

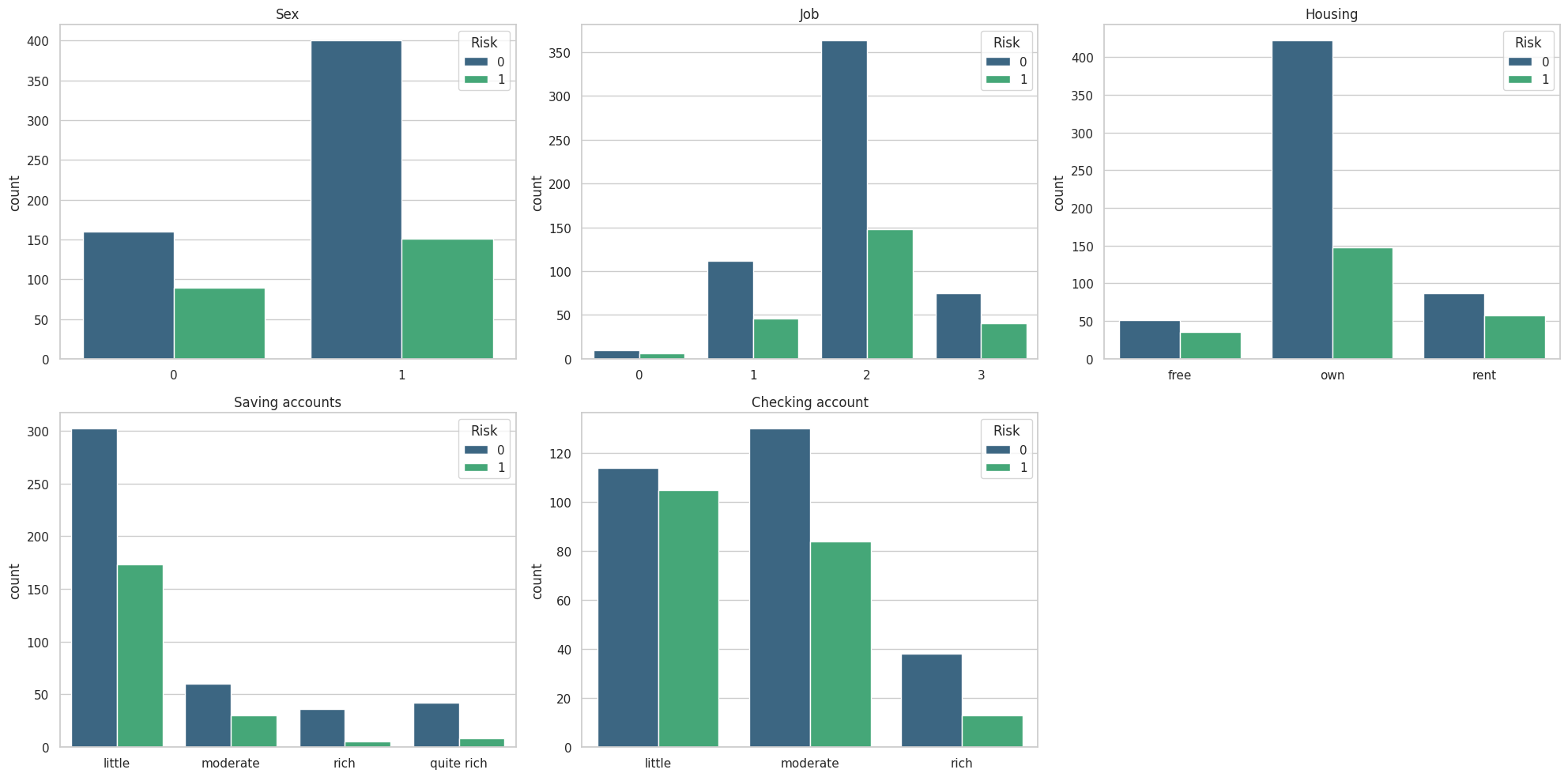
Credit amount and job show a moderate positive correlation, indicating that customers in more skilled professions often qualify for larger credit amounts, while those in less skilled roles usually receive smaller loans.

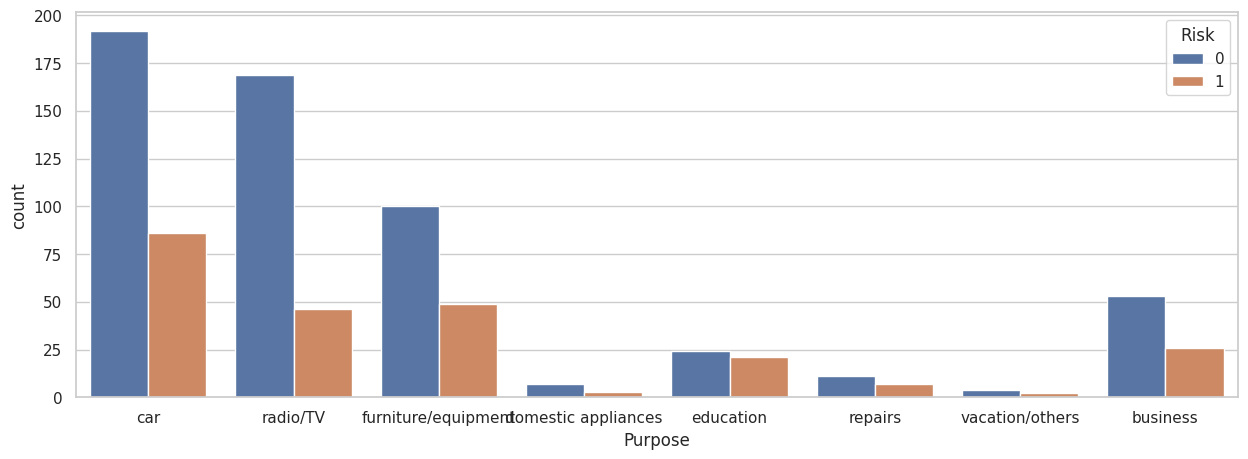
**Figure 5. Distribution of Numerical Features and Risk**



Bad risk customers are typically younger than those classified as good risks, which aligns with the notion that younger individuals often lack financial stability. Additionally, bad risk customers are prone to taking out larger credit amounts than their good risk counterparts, which is understandable given that higher loan amounts are usually associated with increased repayment challenges. Moreover, bad risk customers generally have longer loan durations compared to good risk customers. This trend correlates with the fact that longer loan periods often involve larger sums, thereby elevating the risk, and extended repayment periods increase the likelihood of default.

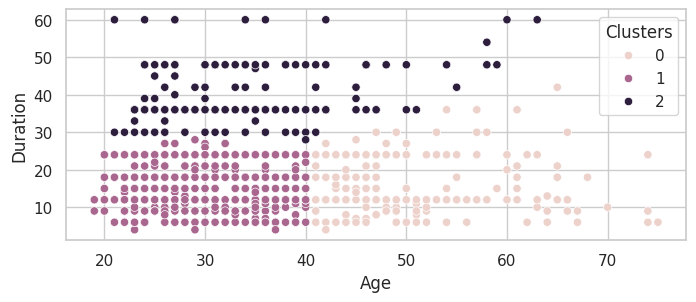
**Figure 6. Distribution of Categorical Features and Risk**





Female customers and those living in rented or free housing exhibit higher bad risk proportions at 36% and 40%, respectively, compared to their counterparts. Financial stability also plays a critical role; customers with minimal savings or checking account balances show significantly higher risk, with nearly 50% of those with low checking balances falling into the bad risk category. Additionally, about half of the customers taking loans for education are labeled as bad risk, reflecting the financial instability often associated with younger demographics.

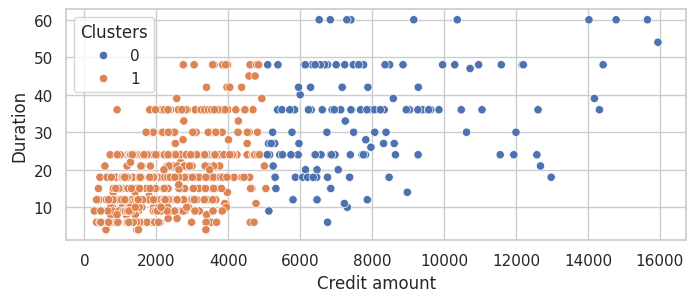
**Figure 6. Age and Duration clustering**



Using the Elbow Method, we determined that three clusters optimally represent the customer data. The clustering yielded the following groups:

* Group 0: Comprises older customers, ranging from 40 to 70 years old, who typically opt for credit services with shorter durations, spanning 10 to 30 months. This group represents mature individuals likely seeking credit for established financial needs.
* Group 1: Consists of younger customers, aged 20 to 40 years, who also prefer shorter-duration credits of 10 to 30 months. This demographic often includes individuals who may lack financial stability, possibly taking smaller loans for immediate needs such as education.
* Group 2: Includes customers of various ages who choose longer credit durations, which could indicate higher loan amounts for substantial expenses such as vacations or other significant expenditures. This group potentially represents a higher risk due to the extended commitment to longer loan terms.

**Figure 7. Credit amount and Duration clustering**



Using the Elbow Method, we identified that two clusters optimally categorize the relationship between credit amount and duration. Here's a concise overview of the groups formed:

* Group 0: Comprising customers who secure higher credit amounts, between 5,000 and 16,000, this group also displays a broader range of durations, from 0 to 60 months. Notably, the longer duration credits within this cluster suggest more substantial financial commitments typical of this group.
* Group 1:This cluster consists of customers who generally take lower credit amounts, ranging from 0 to 5,000, over durations varying from 0 to 50 months. This group likely represents individuals seeking short-term, lower-value financial assistance.

## Data Preprocessing

### 3.1. Handle missing values

In preparing our dataset for analysis, we identified missing values in the 'Saving accounts' and 'Checking account' columns, with 144 and 316 missing entries, respectively. Given the dataset's modest size of 1,000 entries, removing records with missing data was deemed impractical. Such an action would significantly reduce our sample size and could potentially introduce bias into our analyses.

To address this issue, we decided to impute the missing values using the mode, which is the most frequently occurring value in each respective column. This method of imputation helps maintain the integrity and the representational accuracy of our dataset. It preserves the original distribution of values, ensuring that the dataset remains robust and comprehensive for effective analysis and modeling.

### 3.2. Handle categorical features

To prepare our data for machine learning, we converted categorical features into numeric formats suitable for algorithmic processing. For features like 'Saving accounts' and 'Checking account' with an ordinal relationship, Ordinal Encoding was used to preserve their inherent order. This method involves assigning a unique integer to each category according to its relative ranking, which is crucial for maintaining the natural structure of the data and ensuring that the model interprets these features accurately.

For other categorical variables that do not exhibit an inherent order, we employed Target Encoding. This technique replaces categorical values with a number derived from the average target value for each category. It effectively reduces dimensionality and prevents data sparsity issues that might arise from using techniques like OneHot Encoding, especially beneficial in datasets of limited size. Target Encoding is particularly effective as it helps prevent overfitting by incorporating the influence of the target variable directly within the feature, thus maintaining significant information while reducing the number of dummy variables created.

After encoding, we applied StandardScaler to all features to ensure they contribute equally to the model, addressing potential issues where larger scale features could dominate the learning process. This normalization step, by removing the mean and scaling to unit variance, is essential for maintaining balance in the model's sensitivity to feature variations.

### 3.3. Handle numerical features

For numerical attributes in our dataset, scaling is a crucial preprocessing step due to the nature of certain machine learning algorithms. Algorithms that involve distance calculations or employ optimization techniques like gradient descent are particularly sensitive to the scale of input features. Features on larger scales can disproportionately influence the model's learning process, potentially leading to skewed or suboptimal results.

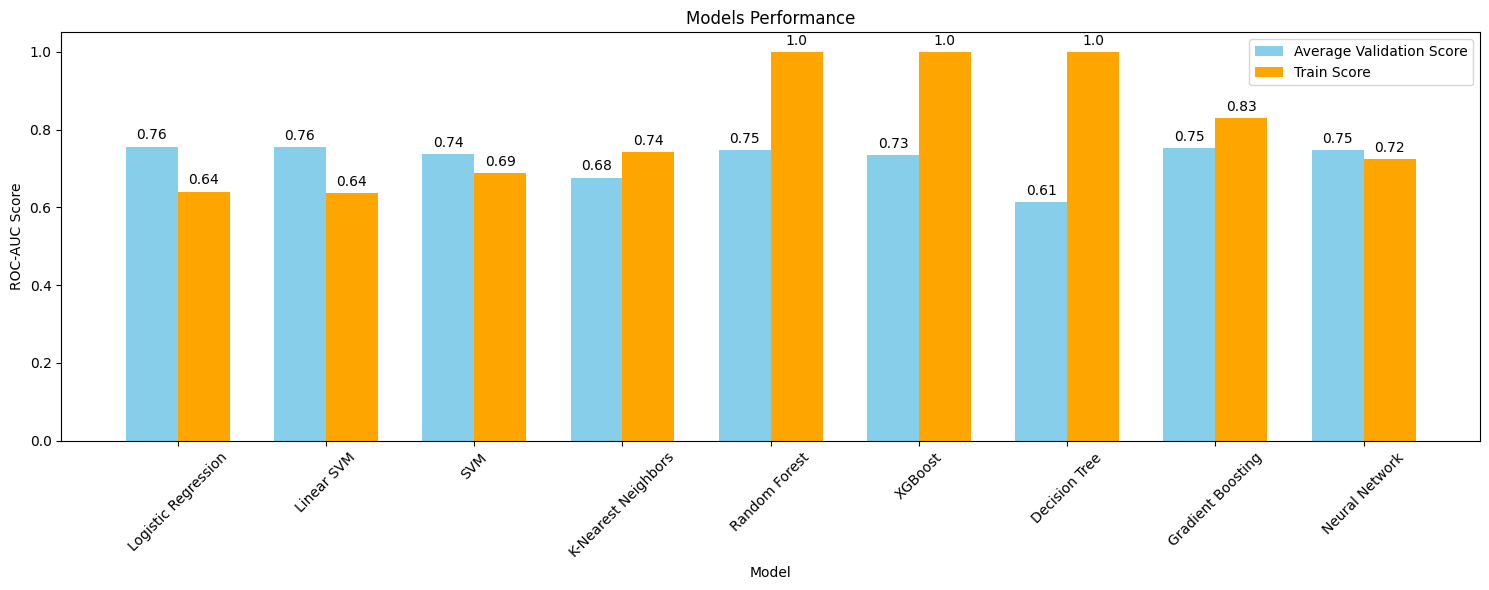
To address this, we applied StandardScaler to all numerical features. This transformation standardizes the data by removing the mean and scaling each feature to unit variance. Essentially, StandardScaler adjusts the data so that the distribution of each feature centers around zero with a standard deviation of one. This normalization ensures that no single feature will dominate others due to differences in scale, leading to a more stable and fair learning process across all attributes. By standardizing numerical data, we facilitate faster convergence during training and enhance algorithm performance by providing a level playing field for all input features. This step is particularly beneficial in environments where algorithms are highly sensitive to the magnitude of input variables, ensuring consistent and reliable modeling outcomes.

# **IV. RESULTS**

In our analysis, we initially divided the data into training and testing sets, allocating 80% for training and the remaining 20% for evaluation to ensure a balanced approach between model training and validation.

Among the various models evaluated, Logistic Regression performed best in terms of the average validation ROC-AUC score, demonstrating strong generalizability to unseen data. However, the Random Forest model, while showing perfect training performance, significantly overfitted, as evidenced by the large gap between training and validation scores. Despite this, the proximity of its validation score to that of the Logistic Regression and its potential for substantial improvement led us to select it for further hyperparameter tuning and final evaluation.

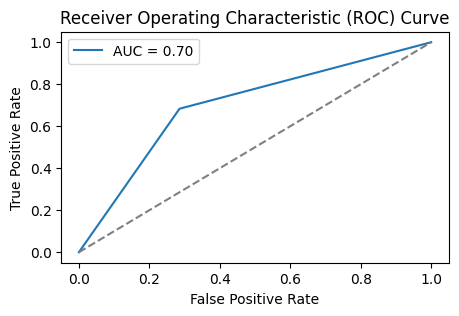
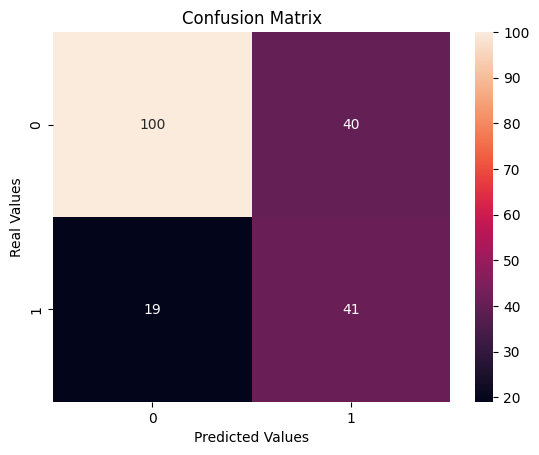
**Figure 8. Models Performance**



We opted for Bayesian Optimization over Grid Search for tuning the Random Forest model due to its adaptability and efficiency. This method allowed us to intelligently navigate the hyperparameter space, optimizing key parameters like the number of estimators, max depth, and class weight, which was set to 'balanced' to address class imbalances.

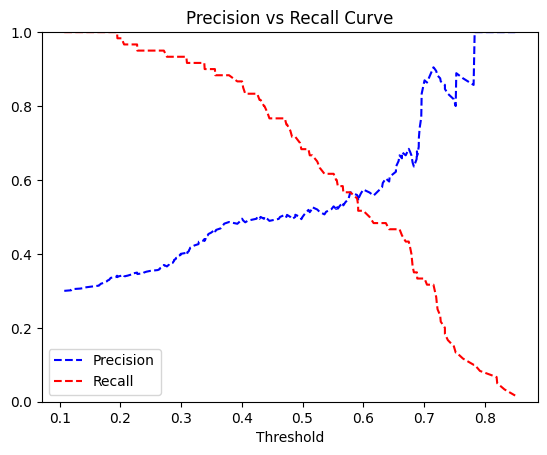
Post-tuning, the Random Forest model demonstrated a recall score of 0.68, meaning it correctly identified 68% of the bad risk customers. Despite a precision score of 0.51, indicating that about half of the predicted bad-risk customers were accurately identified, the model excelled at detecting bad-risk customers, aligning with our primary goal. Furthermore, it achieved an AUC score of 0.70, effectively distinguishing between good and bad risk customers.

**Figure 9. Performance after tuning the final Random Forest model**



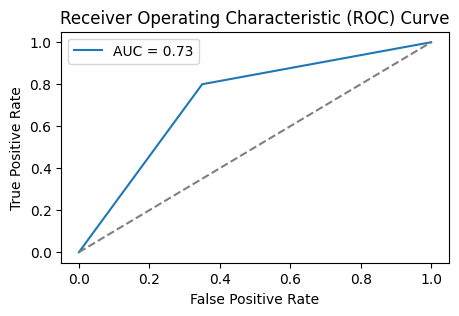
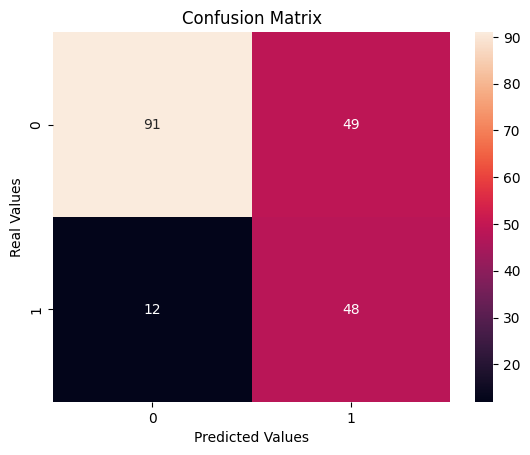
The precision-recall curve analysis highlighted potential improvements in recall without significantly compromising precision. By adjusting the threshold, we enhanced the recall from 0.68 to 0.8, with a slight decrease in precision from 0.51 to 0.49. This adjustment significantly increased the model's ability to correctly identify bad risk customers, predicting 48 out of 60 correctly.

**Figure 10. Precision and Recall Curve**



This refinement in the model's performance not only improved recall and precision but also raised the ROC-AUC score from 0.70 to 0.73, indicating better overall model discrimination between customer risk categories.

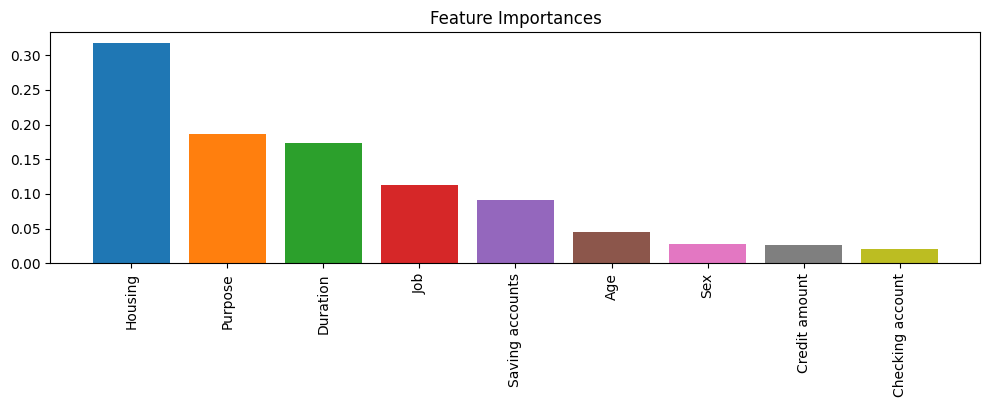
**Figure 11. Performance after balancing the precision-recall trade-off**



In conclusion, by effectively managing the precision-recall trade-off, we have fine-tuned the model to meet our business needs accurately. The model now successfully identifies 80% of bad risk customers, allowing the bank to make well-informed decisions that balance risk management with profit generation, thus ensuring sustainable growth and operational stability in its credit services.

To determine the predictive strength of features, we utilized feature importance analysis within the Random Forest model. This method aggregates the uneven decrease in impurity (Gini or entropy) across all decision trees in the ensemble. Features causing a more significant reduction in impurity when used for splitting are deemed more important.

**Figure 12. Feature importances of Random Forest**



The Random Forest model's feature importance analysis reveals that 'Housing', 'Purpose', 'Duration', 'Job', and 'Saving accounts' are the most significant predictors of credit risk. These features reflect the model's reliance on concrete financial indicators and customer characteristics to make its predictions.

# **V. CONCLUSIONS**

Our project aimed to enhance the way a German bank assesses credit risk, and the journey has been both challenging and enlightening. We explored various machine learning techniques and settled on a Random Forest model, which, after careful tuning, showed it could reliably identify customers likely to present a higher credit risk.

In refining our model, we meticulously handled imbalanced data and adjusted for overfitting, which brought our recall score up to 80%, a significant achievement, as it means the model can now detect the majority of high-risk loan applicants. The precision score, while not perfect, is adequate given our focus on recall, because the primary risk to the bank comes from undetected high-risk loans rather than from false positives.

The analysis revealed that factors like housing status, loan purpose, employment type, and savings account balances were highly predictive of credit risk. These insights are not just numbers; they reflect real-world behaviors and circumstances that affect a person's ability to repay a loan. By understanding these relationships, the bank can make more nuanced lending decisions.

In essence, this project has not only yielded a predictive model that can spot potential risks with impressive accuracy, but it has also offered a clearer understanding of what might lead a customer to become a risk. These insights are invaluable, as they can guide the bank in offering services that are both supportive of customers' needs while protecting its own interests. It's a step towards more empathetic and informed banking, and it's been gratifying to see data science pave the way for such advancements.

# **REFERENCES**

[1] Kaur, H. (n.d.). Understanding K-means Clustering in Machine Learning. Towards Data Science. Retrieved April 20, 2024, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[2] GeeksforGeeks. (n.d.). Stratified K-Fold Cross Validation. Retrieved April 20, 2024, from https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics.

[5] Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189-1232.

This is the code of the report.