# **Loan Default Prediction Using Machine Learning**

# Introduction

Think of data mining as putting together a big puzzle. It's like when you're trying to solve a mystery using all the clues you have. This helps us find out new and interesting things from lots of information, just like a detective uses tools to find hidden clues.

Choosing the right way to solve the puzzle is really important. It's like picking the right tool for a job at home. We need to choose the best way to understand the data and figure out why things are happening the way they are. It's a bit like finding out the story behind events.

Imagine you have many tools, like different tools in a toolbox. You have to pick the tool that works best for the kind of information you have and the things you want to know.

**Bank Loan Decisions: Being Careful with Money**

Now, picture a bank getting requests from different people who want to borrow money. The bank has to decide if it's safe to give them a loan or not. This is a big decision because it affects how much money the bank has.

There are two main situations the bank worries about:

1. **Good Credit Risk**: Some people are really likely to pay back the money they borrow. If the bank says "no" to giving them a loan, it might lose a chance to make money.
2. **Bad Credit Risk**: Other people might have trouble paying back the money they borrow. If the bank gives them a loan and they can't pay it back, the bank could lose money.

So, the bank has to be super careful. Saying "yes" to the wrong people might make the bank lose money. But saying "no" to the right people means the bank might miss out on making money. It's like trying to balance things just right for the bank.

# Objective of Analysis:

**Being Smart with Bank Money:** Our main goal is to help the bank make really smart choices. We want to stop the bank from losing money and help it make as much money as it can.

**Figuring Out Who's Safe:** To avoid losing money, we need a plan. We want to decide who should get a loan and who shouldn't. We look at information about people, like where they live and how much money they have, to help us decide.

**Making a Smart Tool:** We're working on creating a special tool that can learn from this information. This tool will help the bank manager make smart choices. For example, if someone asks for a loan, the tool will check their information and say if it's safe to give them a loan or not.

Our goal is to build this clever tool that tells the bank manager whether to say "yes" or "no" to people who want loans. It looks at their information and helps the bank avoid losing money while making better choices.

# Methods and Materials:

This report centers around a well-known dataset in the world of Machine Learning: the German Credit Risk dataset.

Given the sheer number of loan applications that need quick processing daily, having a predictive model can greatly aid decision-making for executives. Such a model can provide insights to guide the approval or rejection of new loan applications efficiently.

In this, I will outline a step-by-step process to create a Machine Learning predictive model for scenarios like this. This approach can be used as a template to tackle any supervised ML classification problem.

The case study follows this sequence:

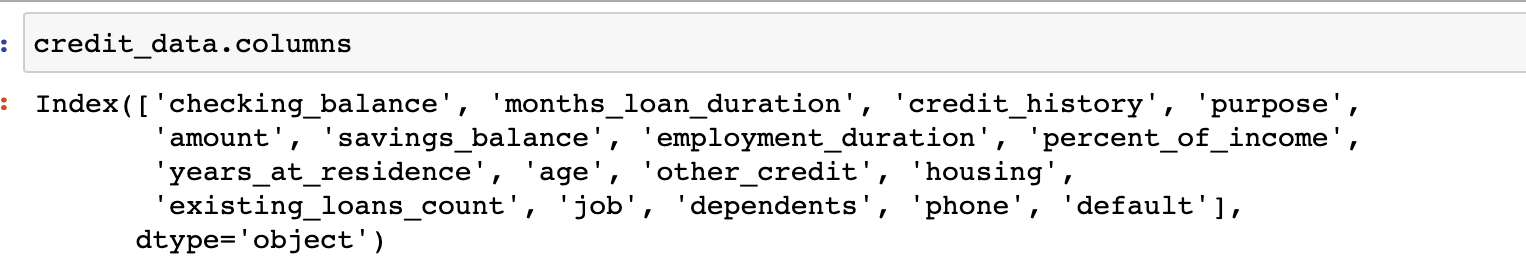
* Understand The data
* Identifying the Target variable and Examining the distribution of the Target variable to understand its nature
* Basic exploration of the data to grasp its characteristics and visual exploratory techniques like Histograms and Bar charts to understand data distribution and Selecting features (data attributes) based on how they're distributed
* Handling outliers (unusually extreme data points) and Managing missing values in the dataset
* Conducting visual analysis to understand feature correlations and Applying statistical methods to identify features most correlated with the Target variable
* Data Pre-processing for Machine Learning
* Trying out various classification algorithms to see which one works best for the given problem
* Choosing the most effective Model based on performance
* Deploying the selected model for real-world use

This case study's flow provides a structured path to create a Machine Learning predictive model for classification problems. You can adapt this process to tackle similar challenges and make informed decisions using data-driven insights.

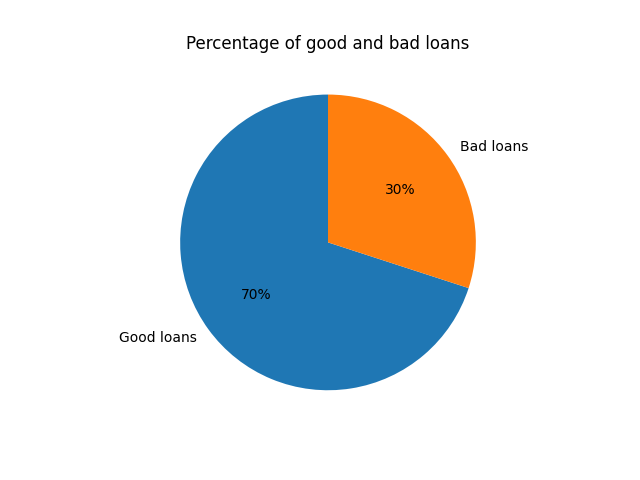
**Understand The data**

This is one of the most important steps in machine learning! You must understand the data and the domain well before trying to apply any machine learning algorithm.

The bank has historical information on relevant features for each customer such as employment duration, existing loans count, saving balance, percentage of income, age, default status.

The data set has 17 columns and 1000 rows. Columns are described below and each row is a customer. ****

1. checking\_balance - Amount of money available in account of customers
2. months\_loan\_duration - Duration since loan taken
3. credit\_history - credit history of each customers
4. purpose - Purpose why loan has been taken
5. amount - Amount of loan taken
6. savings\_balance - Balance in account
7. employment\_duration - Duration of employment
8. percent\_of\_income - Percentage of monthly income
9. years\_at\_residence - Duration of current residence
10. age - Age of customer
11. other\_credit - Any other credits taken
12. housing- Type of housing, rent or own
13. existing\_loans\_count - Existing count of loans
14. job - Job type
15. dependents - Any dependents on customer
16. phone - Having phone or not
17. default - Default status (Target column)

**Identifying the Target variable and Examining the distribution of the Target variable to understand its nature**

* Target Variable: default
* Predictors: duration, history, purpose, amount, savings etc.
* default=1 means the loan was a good decision.
* default=0 means the loan was a bad decision.

Target variable's distribution is vital. Extreme imbalance can hinder predictive modelling, while slight skew is acceptable. Balanced class distribution is crucial for classification tasks. Unbalanced data may impede Machine Learning algorithms' learning.

Luckily, this data's target variable distribution is favourable. Adequate rows per category allow confident progress. This balance enables accurate models for predictions in each class.

**Basic exploration of the data to grasp its characteristics and visual exploratory techniques like Histograms and Bar charts to understand data distribution and Selecting features (data attributes) based on how they're distributed**

This step is performed to understand the overall data. The volume of data, the types of columns present in the data. Initial assessment of the data should be done to identify which columns are Quantitative, Categorical or Qualitative.

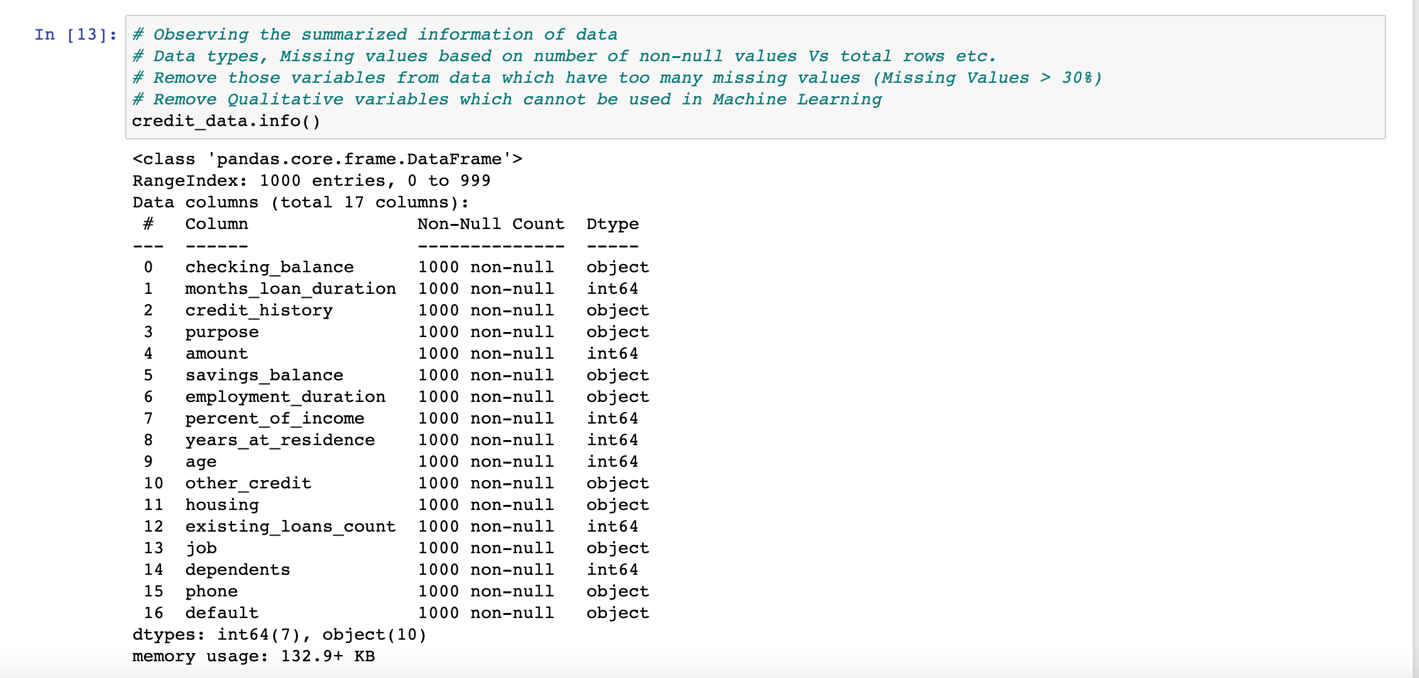
This step helps to start the column rejection process. You must look at each column carefully and ask, does this column affect the values of the Target variable? For example in this case study, you will ask, Does this column affect the approval or rejection of loan? If the answer is a clear "No" the remove the column immediately from the data otherwise keep the column for further analysis.

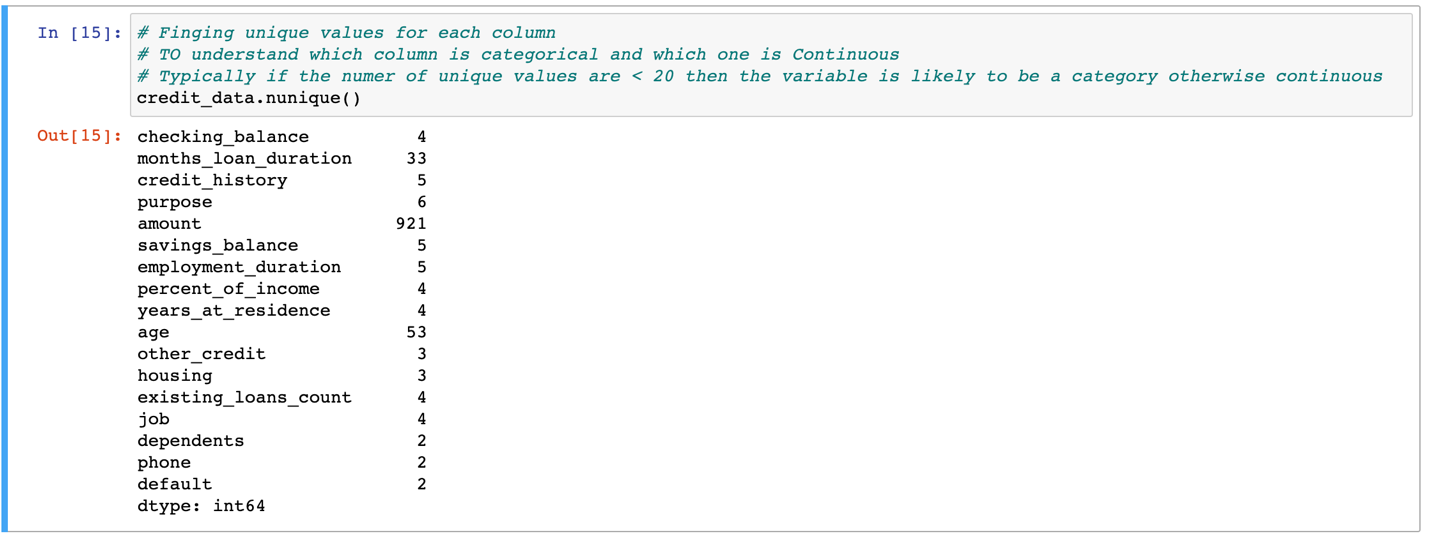
**Basic exploration:**

Below is an initial report based on our basic exploration of the data. These observations serve as a roadmap for our upcoming analysis.

Categorical Predictors: checking\_balance, credit\_history, purpose, savings\_balance, employment\_duration, percent\_of\_income, years\_at\_residence, other\_credit, housing, existing\_loans\_count, job, dependents, phone

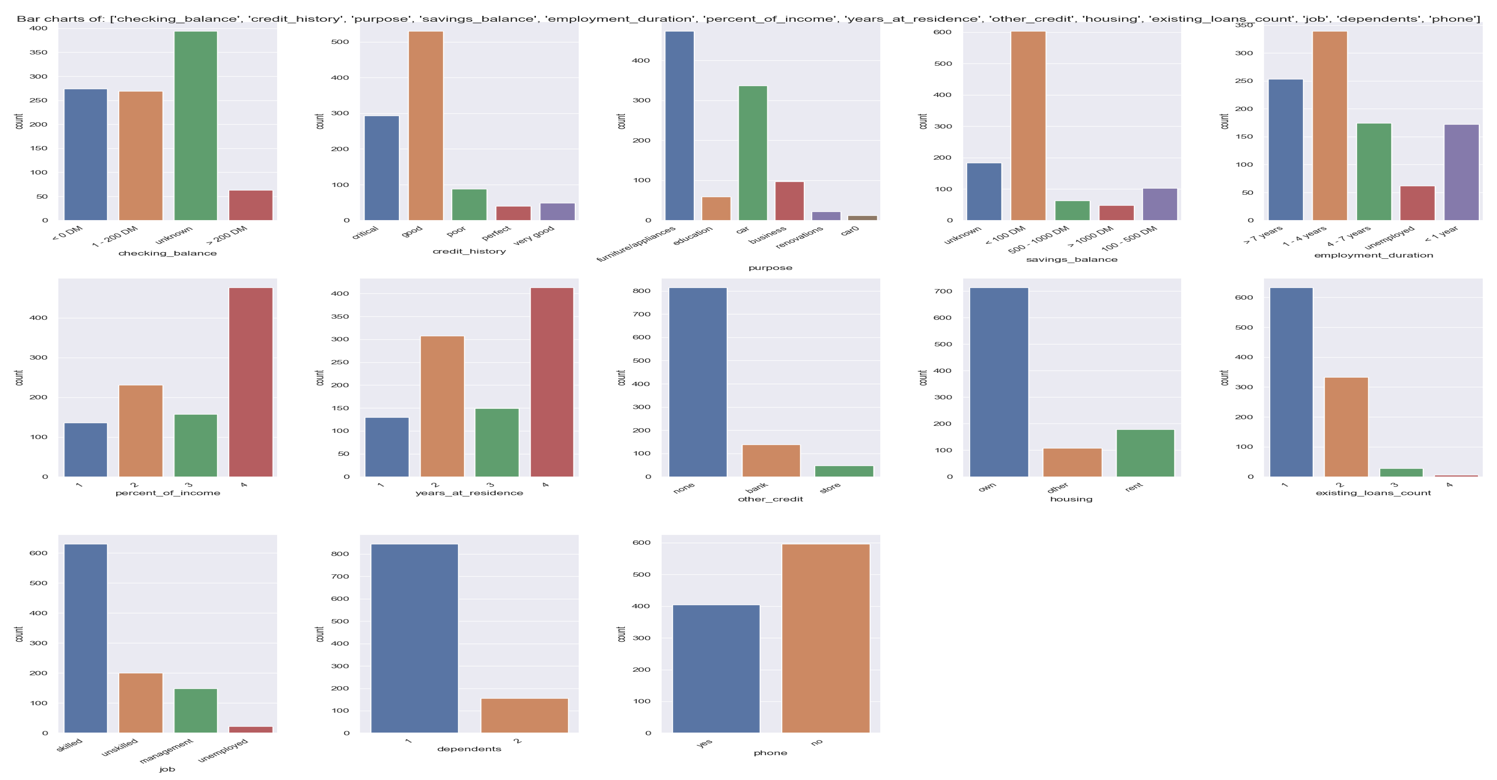
Continues Predictors: months\_loan\_duration, amount and age.





**Visual Exploratory of Categorical variables:** Based on the Basic Data Exploration above, we have spotted seventeen categorical predictors in the data

Categorical Predictors: checking\_balance, credit\_history, purpose, savings\_balance, employment\_duration, percent\_of\_income, years\_at\_residence, other\_credit, housing, existing\_loans\_count, job, dependents, phone

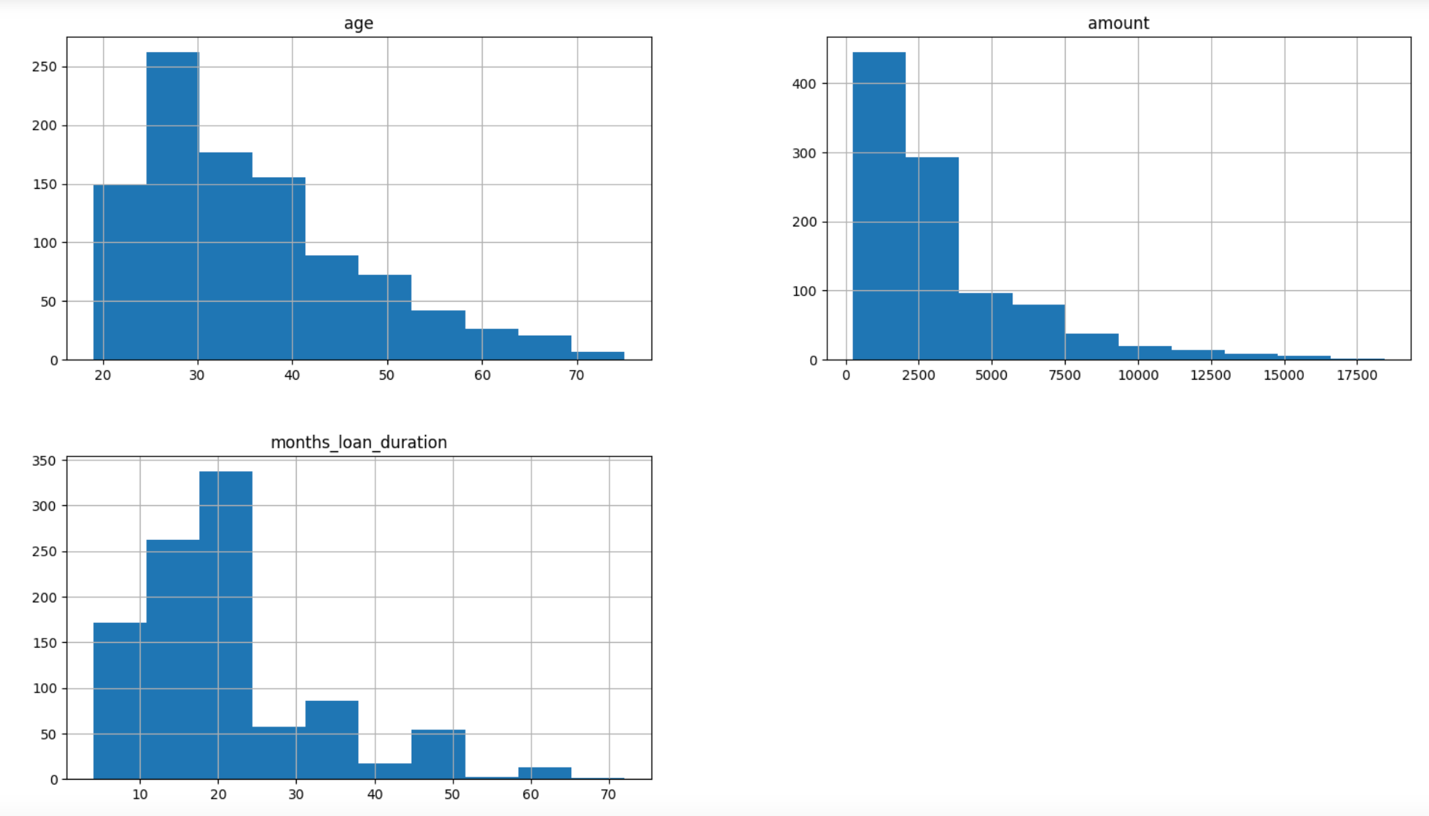
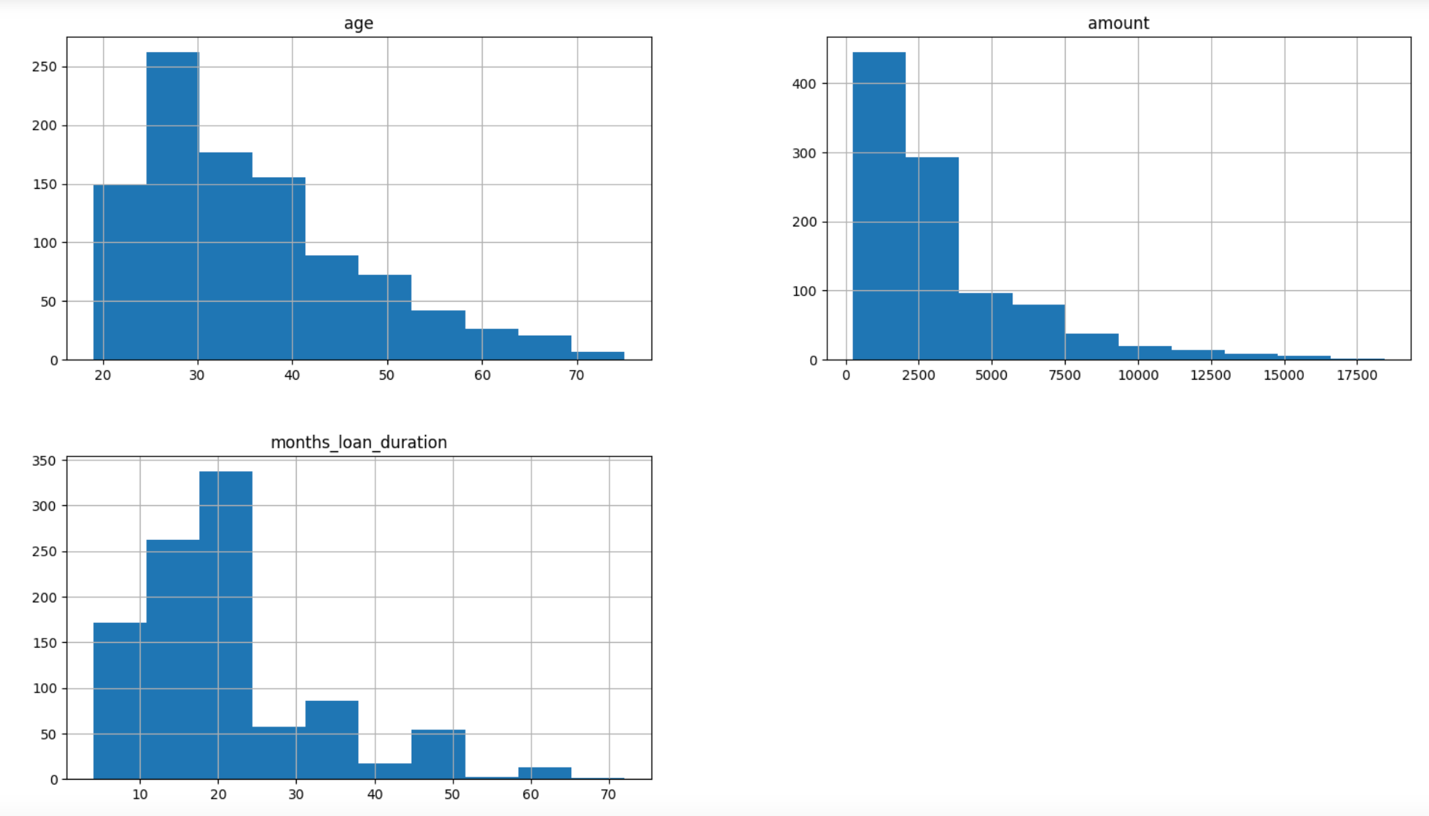
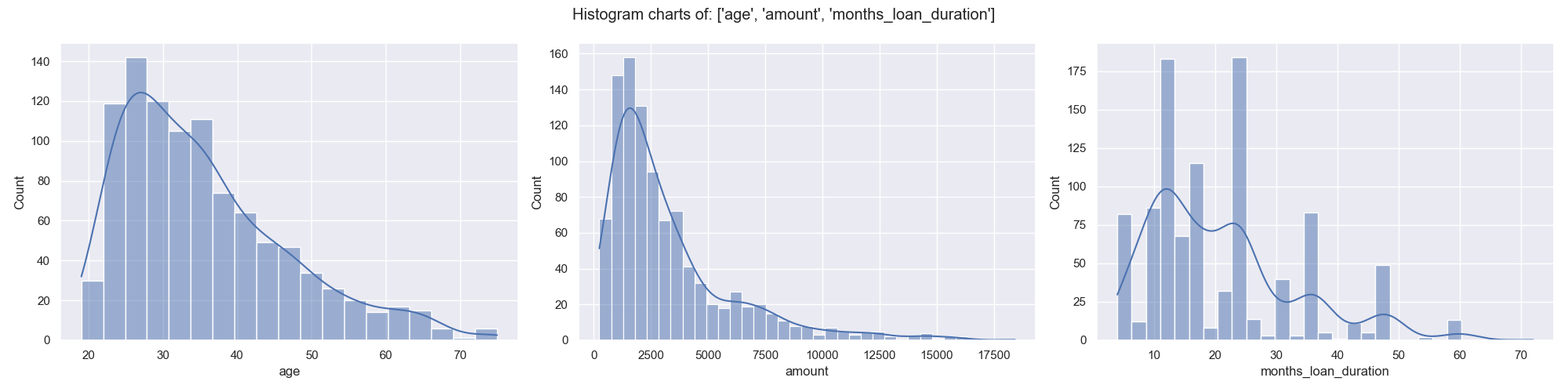
Bar charts show how often each category appears on the Y-axis, while the category names are on the X-axis. A perfect example is the " years\_at\_residence" column chart, where the categories have similar frequencies. This means there's a good amount of data for each category, helping the Machine Learning algorithm learn effectively. However, some columns might have a skewed distribution. This could make those columns less useful for Machine ****Learning .This means one category dominates, and others have very few instances.

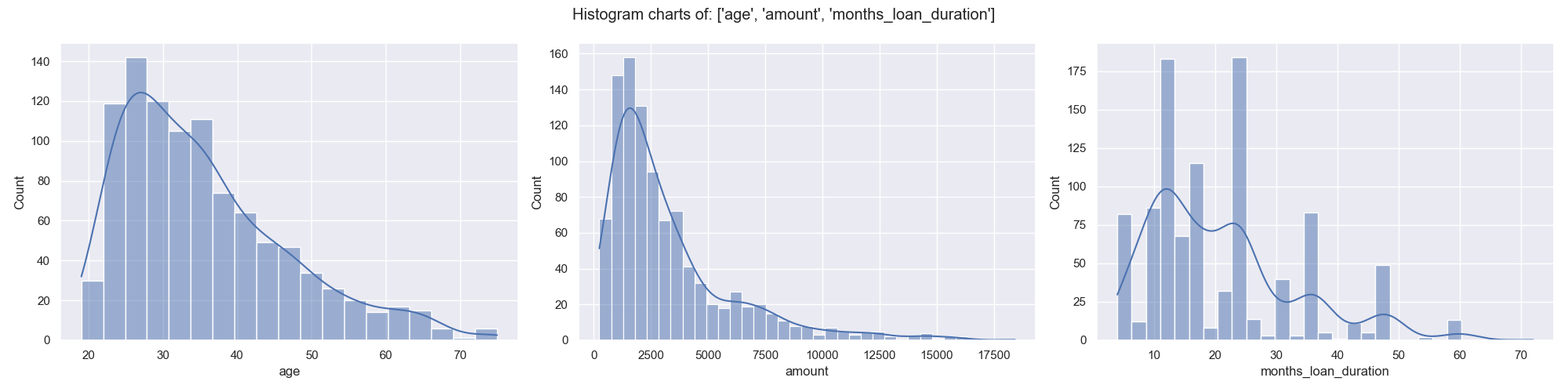
We'll confirm this in the correlation analysis and then decide if we should keep or remove those columns

In this dataset, all the categorical columns except "other\_credit" have balanced distributions that are great for Machine Learning.

We're including all the categorical variables in our analysis except "other\_credit." We're a bit uncertain about " other\_credit " and will make a final decision after the correlation analysis. This way, we ensure our data is ready for effective Machine Learning.**Top of Form**

**Visual Exploratory of Continuous variables:** Based on the Basic Data Exploration, There are Three continuous predictor variables months\_loan\_duration, amount and age.

****

****

histograms visually depict the distribution of a single continuous variable. The X-axis denotes the value range, while the Y-axis displays the count of values falling within each range. For instance, in the "age" histogram, approximately 260 data rows exhibit ages between 25 to 30.

An ideal histogram outcome is a bell curve or a slightly skewed version. If skewness is extreme, addressing outliers is essential. If that doesn't resolve the issue, column reconsideration is necessary.

For the chosen variables:

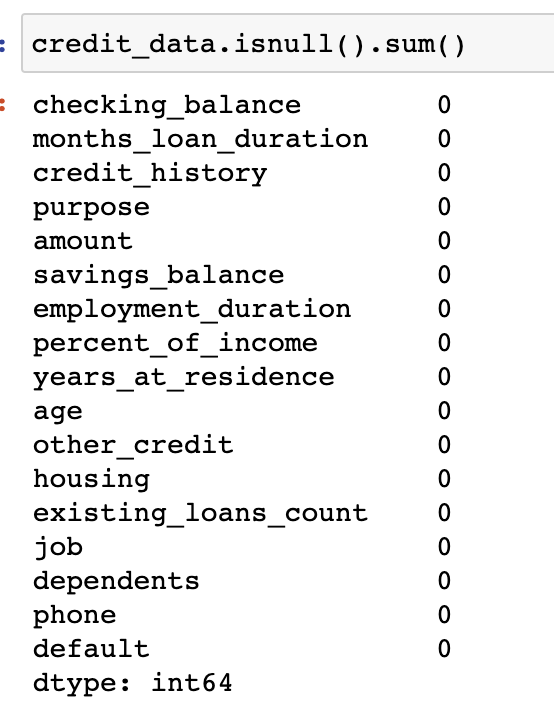
* "age": Slightly skewed but acceptable.
* "amount": Slightly skewed but acceptable.
* "months\_loan\_duration": Slightly skewed but acceptable.

In summary, these variables exhibit acceptable distributions, ensuring their suitability for analysis.

Top of Form

Bottom of Form

**Handling outliers (unusually extreme data points) and Managing missing values in the dataset**

outliers are unusual data points that deviate from the common pattern and are usually seen as tails in histograms. Each column's outliers should be treated differently since their impact varies. Outliers can mislead machine learning models by pushing them towards extreme values, away from the majority of the data. There are options to either remove outlier records (if a few rows are lost) or replace outlier values with logical alternatives based on business context. Fortunately, in this dataset, all continuous variables exhibit slightly skewed distributions, which is acceptable, making outlier treatment unnecessary.

missing values are managed for individual columns. If a column contains over 30% missing data, it's best to reject it due to significant information loss.

Possible treatments for missing values include:

* Deleting rows with missing values, applicable when only a few records are affected.
* Imputing missing values with the median for continuous variables.
* Imputing missing values with the mode for categorical variables.
* Interpolating values based on nearby data.
* Interpolating values using contextual business logic.

In the context of this dataset, fortunately, there are no missing values present.

**Conducting visual analysis to understand feature correlations and Applying statistical methods to identify features most correlated with the Target variable**

Now its time to finally choose the best columns(Features) which are correlated to the Target variable. This can be done directly by measuring the correlation values or ANOVA/Chi-Square tests. However, it is always helpful to visualize the relation between the Target variable and each of the predictors to get a better sense of data.

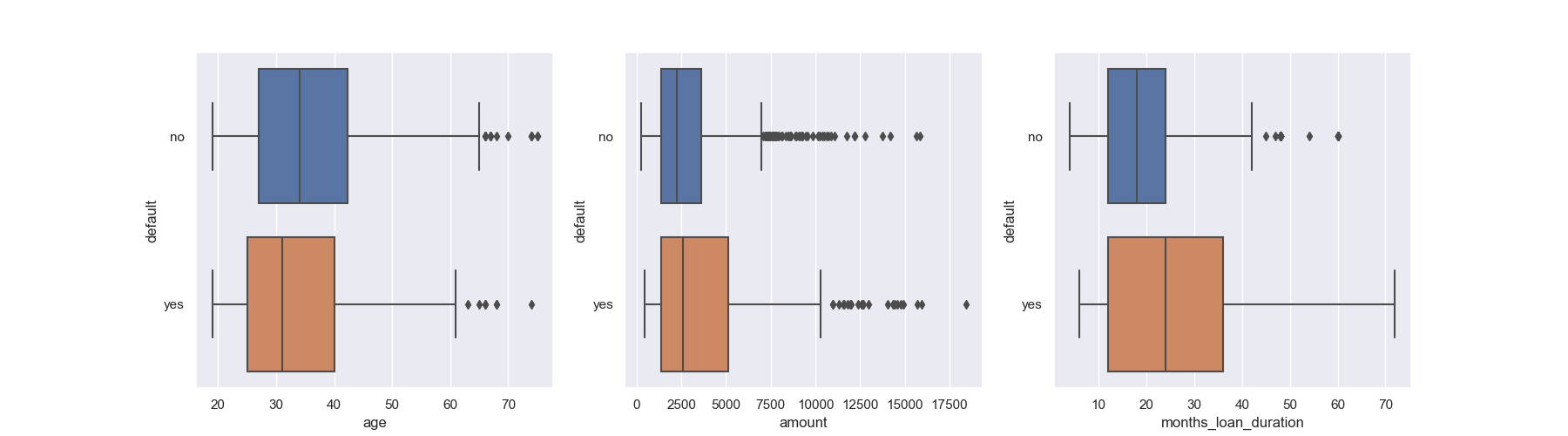
**Categorical Vs Continuous:** When the target variable is Categorical and the predictor variable is Continuous we analyze the relation using bar plots/Boxplots and measure the strength of relation using Anova test

these plots illustrate the data distribution of continuous predictors on the Y-axis for each category on the X-axis.

* If the distributions are similar for each category (boxes align), it implies the continuous variable has little effect on the target. This means the variables are not correlated.

For instance, consider the first chart, "age" vs. "default." The boxes align, suggesting that age doesn't distinctly differentiate loan approval or rejection. Hence, this column is not correlated with "default."

Conversely, the other two charts show opposite characteristics, indicating that "amount" and "month\_loan\_duration" are correlated with the target variable.

****These observations highlight how specific continuous variables relate to the target, aiding in understanding their correlation.

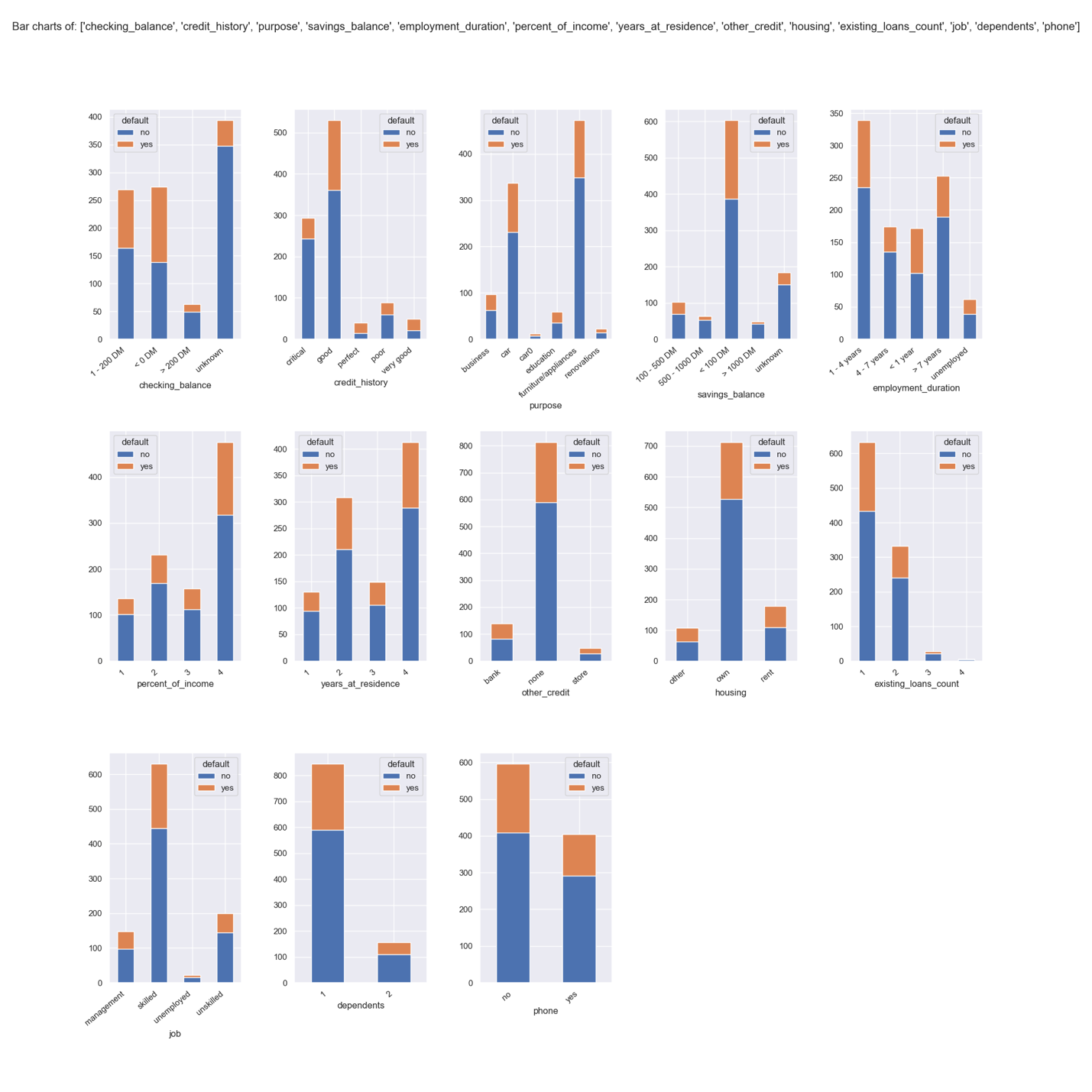
the ANOVA results validate our earlier box plot observations.

* The "age" P-value hovers near the threshold, aligning with our earlier doubts from the box plots.
* ****On the other hand, the P-values for the other two variables are definitively zero, strongly affirming their correlation.

Conclusively, all three columns are indeed correlated with "default." The ANOVA findings reinforce our initial visual analysis.

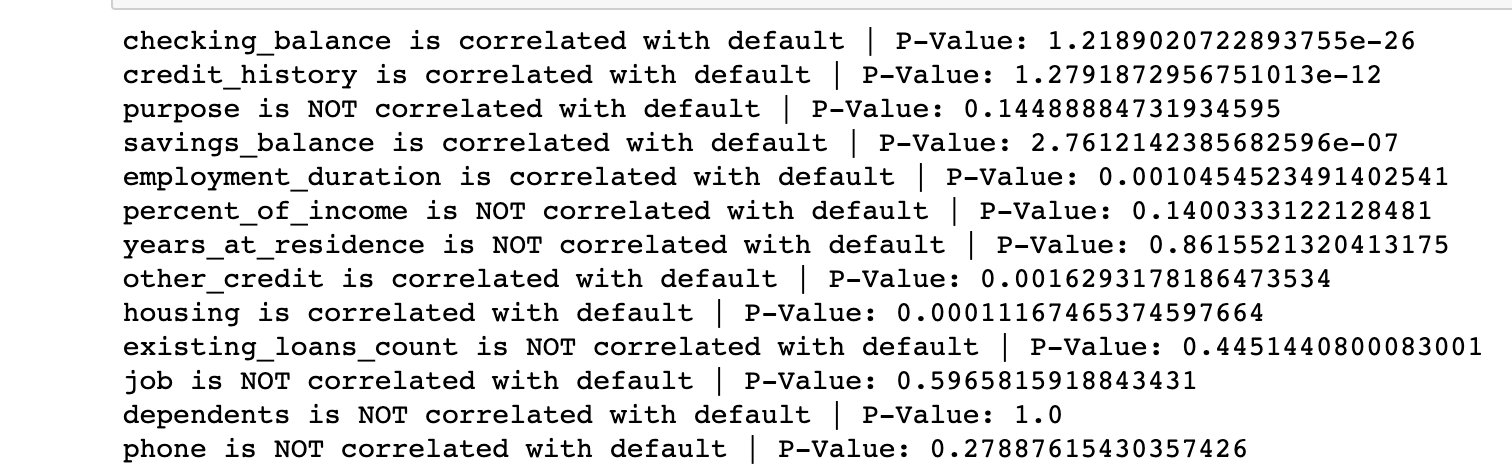
**Categorical Vs Categorical:** When the target variable is Categorical and the predictor is also Categorical then we explore the correlation between them visually using barplots and statistically using Chi-square test

These grouped bar charts show the frequency in the Y-Axis and the category in the X-Axis. If the ratio of bars is similar across all categories, then the two columns are not correlated. For example, look at the "phone" Vs "default" plot. The no vs yes ratio similar, it means tele does not affect the Good/Bad Credit!. Hence, these two variables are not correlated. On the other hand, look at the "credit\_history" vs "default" plot. The number of Bad Credits are very high if credit\_history if good or critical. It means history affects the Good/Bad Credit! Hence, two columns are correlated with each other. We confirm this analysis in below section by using Chi-Square Tests.



Chi-Square test is conducted to check the correlation between two categorical variables

* Assumption(H0): The two columns are NOT related to each other
* Result of Chi-Sq Test: The Probability of H0 being True



Based on the results of Chi-Square test, checking\_balance, credit\_history, savings\_balance, employment\_duration, other\_credit, 'housing' categorical columns are selected as predictors for Machine Learning

Based on the above tests, selecting the final columns for machine learning

**checking\_balance, credit\_history, savings\_balance, employment\_duration, other\_credit, housing, age, amount, months\_loan\_duration**

**Data Pre-processing for Machine Learning**

List of steps performed on predictor variables before data can be used for machine learning

1. Converting each Ordinal Categorical columns to numeric
2. Converting Binary nominal Categorical columns to numeric using 1/0 mapping
3. Converting all other nominal categorical columns to numeric using pd.get\_dummies()
4. Data Transformation (Optional): Standardization/Normalization/log/sqrt. Important if you are using distance based algorithms like KNN, or Neural Networks

**Trying out various classification algorithms to see which one works best for the given problem**

We dont use the full data for creating the model. Some data is randomly selected and kept aside for checking how good the model is. This is known as Testing Data and the remaining data is called Training data on which the model is built. Typically 70% of data is used as Training data and the rest 30% is used as Testing data.

In our study, we have explored a range of classifiers to better understand their capabilities and performance. Here's a brief overview of each classifier:

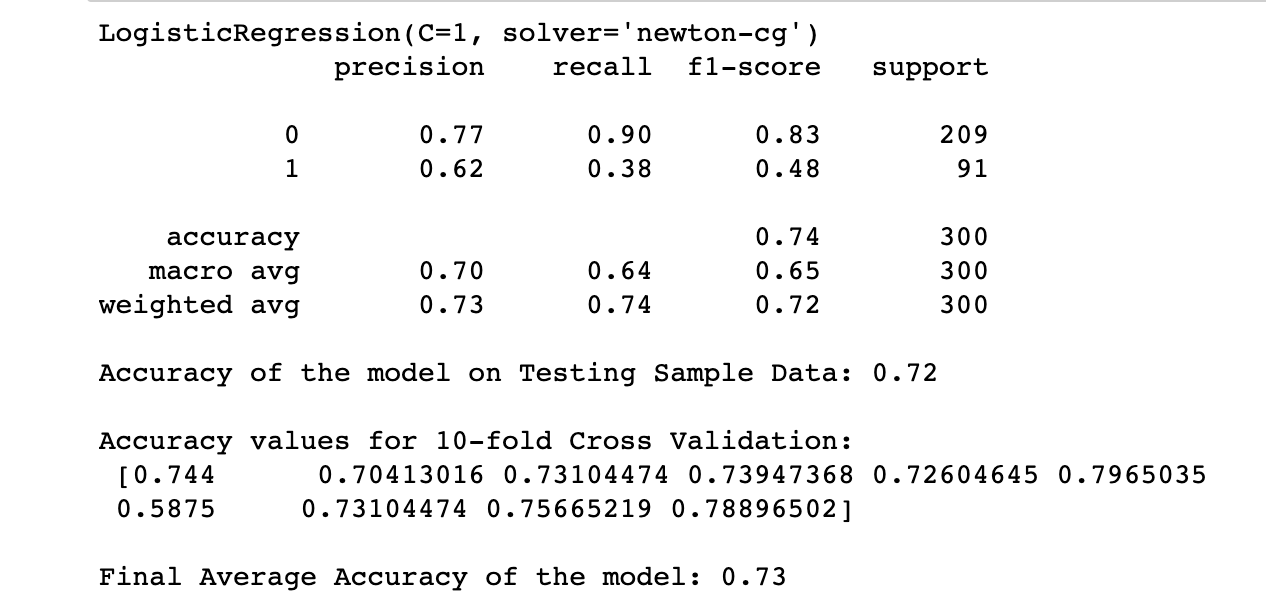
1. **Logistic Regression:**
   * A fundamental linear classifier suitable for binary classification tasks.
   * Well-suited for scenarios where the relationship between features and outcomes is approximately linear.
   * Provides probabilities indicating the likelihood of class membership.
2. **Decision Trees:**
   * Non-linear classifiers that make decisions through hierarchical splitting of data.
   * Visual and intuitive, they capture intricate relationships in the data.
   * Pruning is important to prevent overfitting and improve generalization.
3. **Random Forest:**
   * An ensemble of decision trees that mitigates overfitting by combining predictions.
   * Robust and versatile, making them suitable for various tasks.
   * Effective at handling high-dimensional data and capturing complex patterns.
4. **AdaBoost:**
   * Adaptive boosting technique that leverages multiple weak learners to build a strong classifier.
   * Focuses on correcting errors from previous iterations, enhancing accuracy.
   * Offers improved generalization and resistance to overfitting.
5. **XGBoost:**
   * An advanced gradient boosting algorithm known for its competitive performance.
   * Optimizes boosting process, resulting in higher accuracy and reduced overfitting.
   * Widely recognized in machine learning competitions due to its efficacy.
6. **K-Nearest Neighbors (KNN):**
   * Instance-based learning that classifies based on the majority class of its nearest neighbors.
   * Simple and intuitive, but computationally demanding for larger datasets.
   * Effectiveness depends on the choice of distance metric and the value of 'k'.
7. **Support Vector Machines (SVM):**
   * A powerful classifier that identifies the best hyperplane to separate classes while maximizing margin.
   * Effective for both linear and non-linear classification using kernel functions.
   * Particularly useful in scenarios with separable classes or clear decision boundaries.
8. **Naive Bayes:**
   * A probabilistic classifier that applies Bayes' theorem under the naive assumption of feature independence.
   * Efficient and commonly used for text classification, sentiment analysis, and spam detection.
   * Simple yet effective for certain types of datasets.

By exploring these classifiers, we aim to identify the most suitable model for our our classification task, considering factors such as interpretability, accuracy, and generalization capability. The choice of classifier ultimately depends on the characteristics of our dataset and the desired outcome.

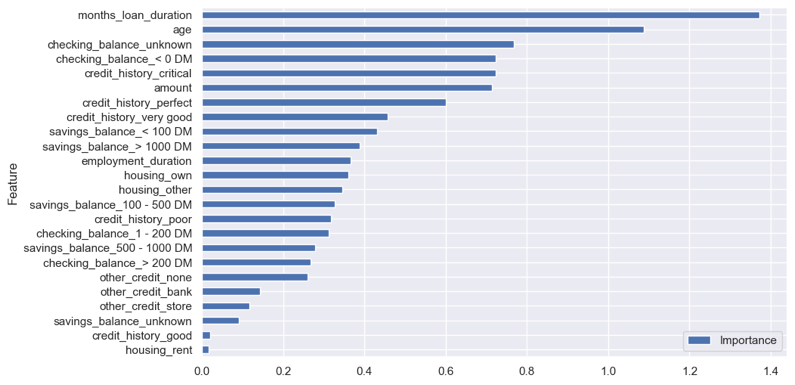
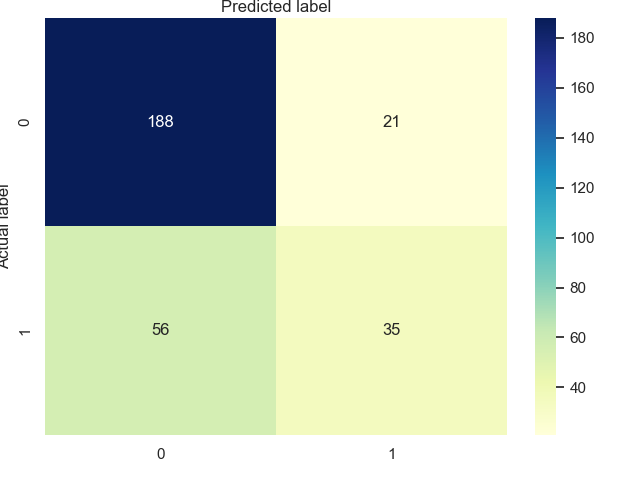
# Results:

**Logistic Regression:**

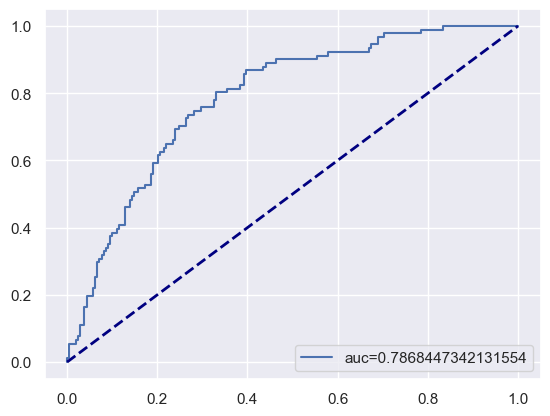
* **accuracy and F1 score**

****

* **Confusion matrix and feature importance**

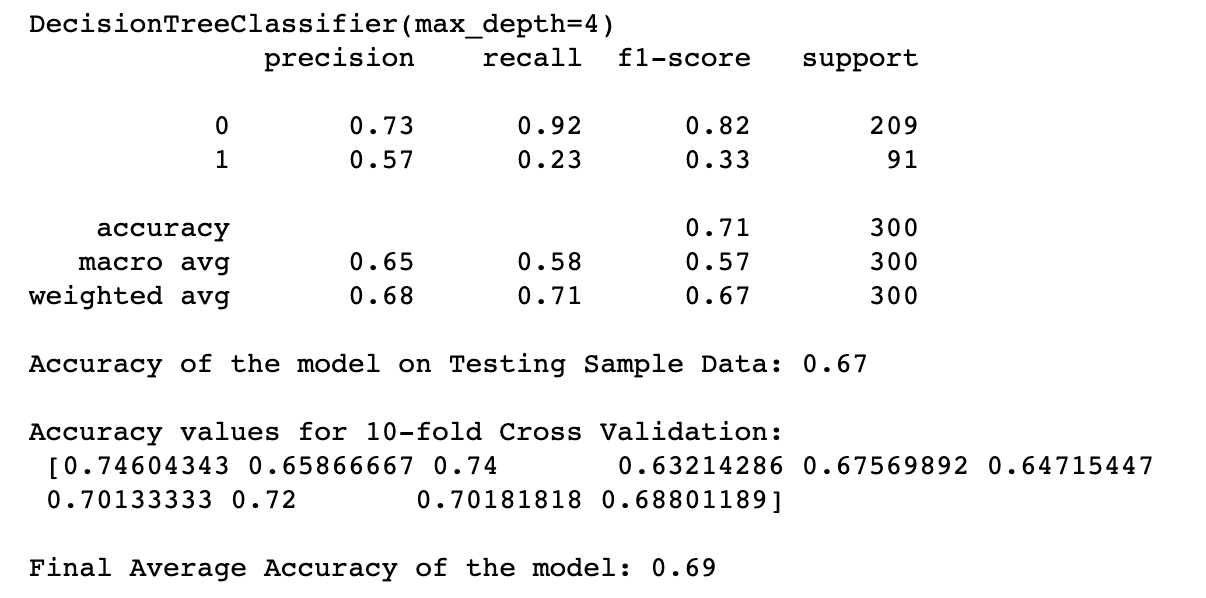
****

* **ROC curve**

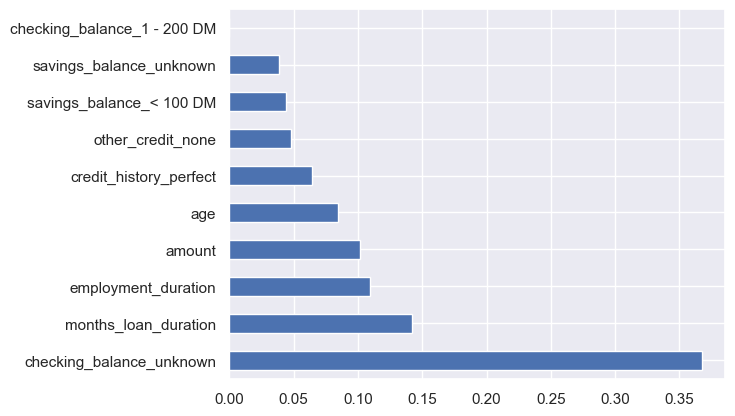
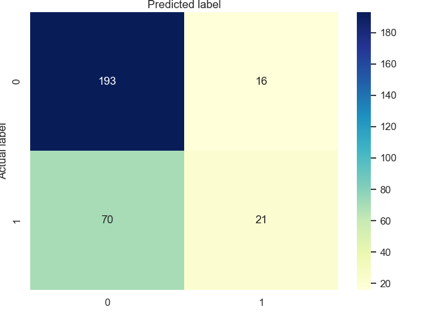
****

**Decision Trees:**

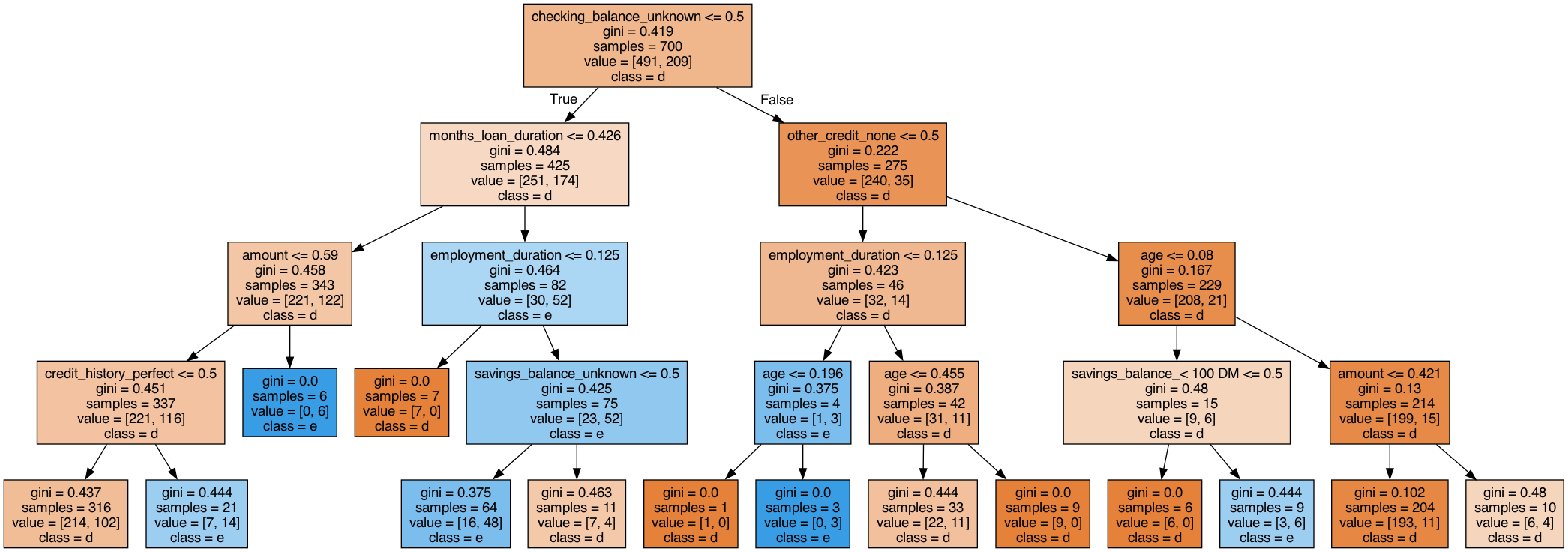
* **accuracy and F1 score**

****

* **Confusion matrix and Feature importance**

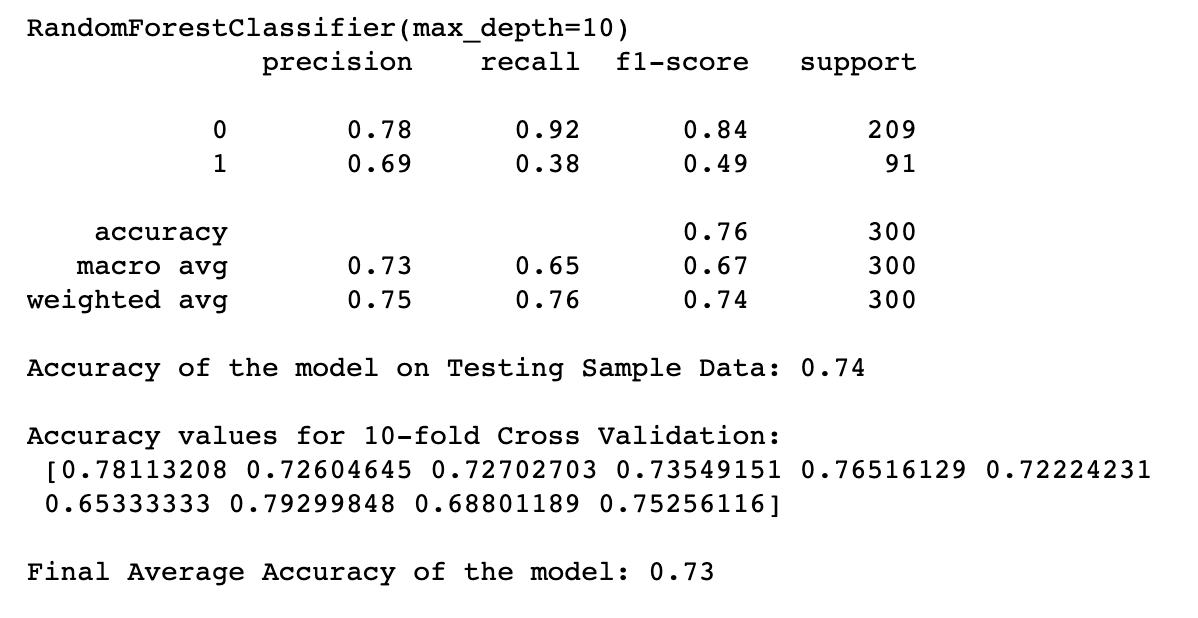
****

* **Decision Tree**

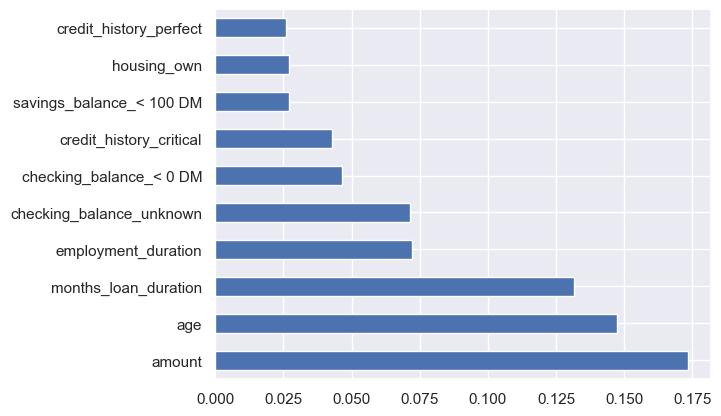
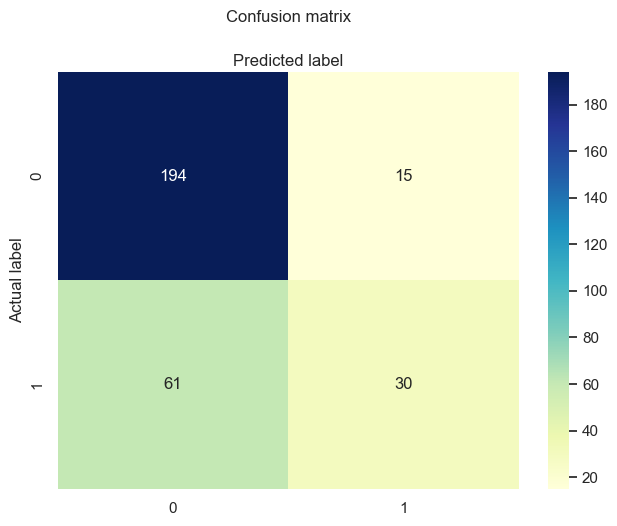
****

**Random Forest:**

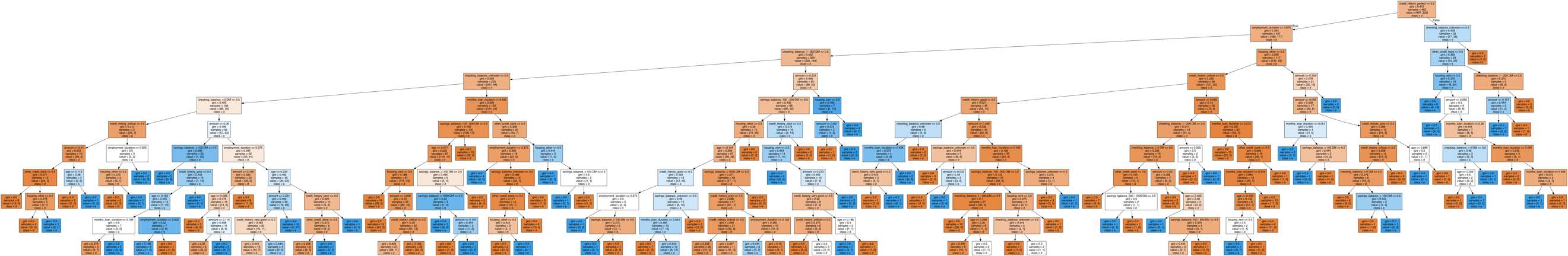
* **accuracy and F1 score**

****

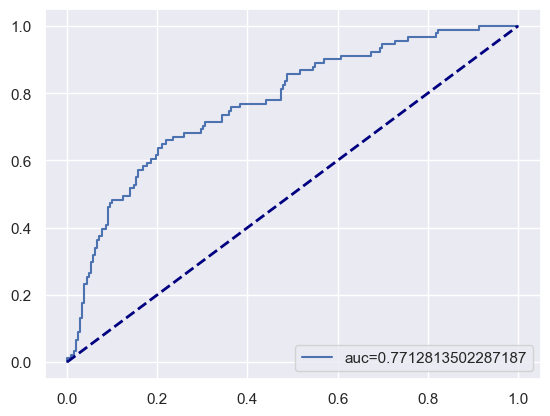
* **Confusion matrix and Feature importance**

****

* **Decision Tree**

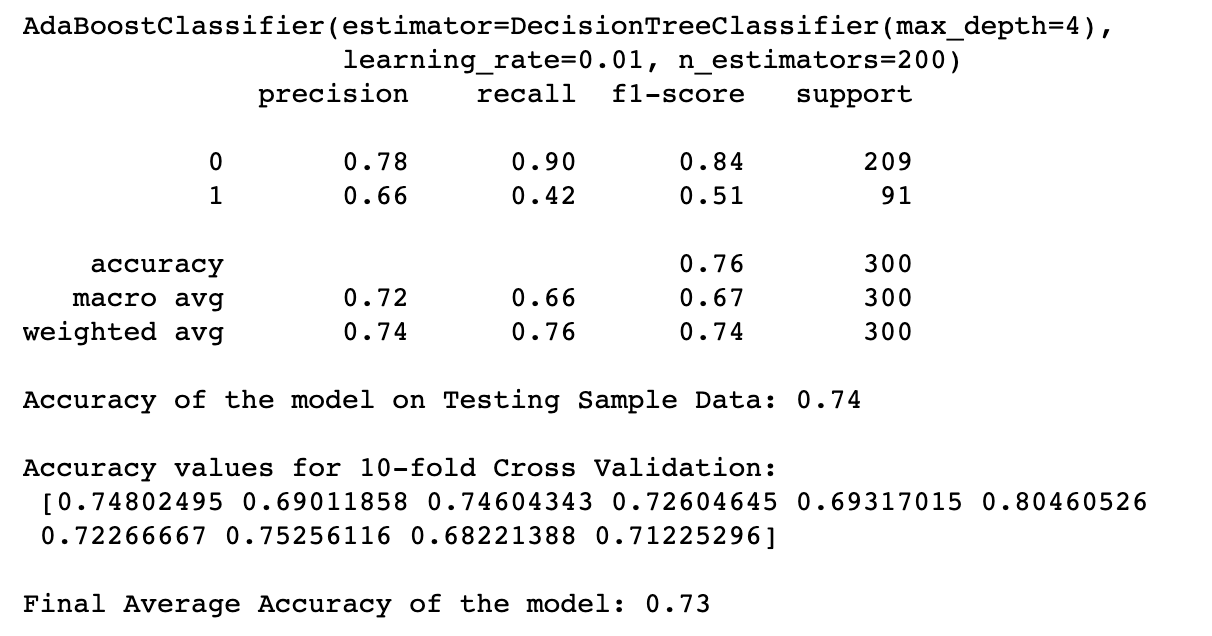
****

* **ROC curve**

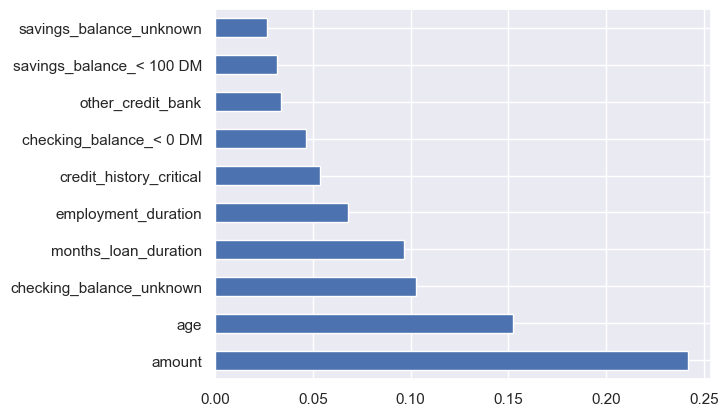
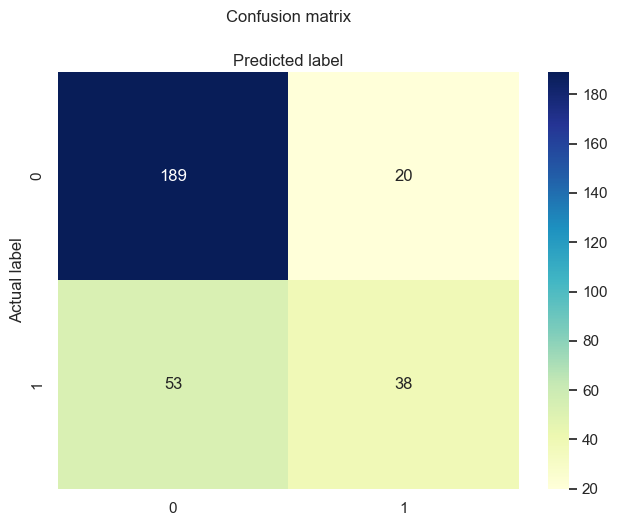
****

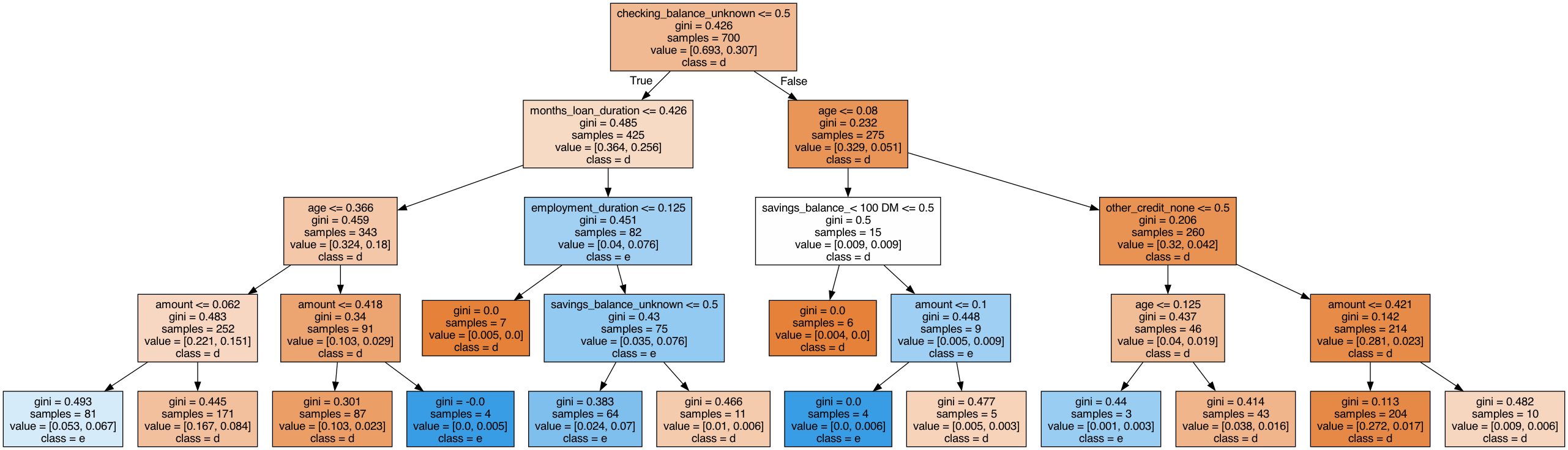
**AdaBoost:**

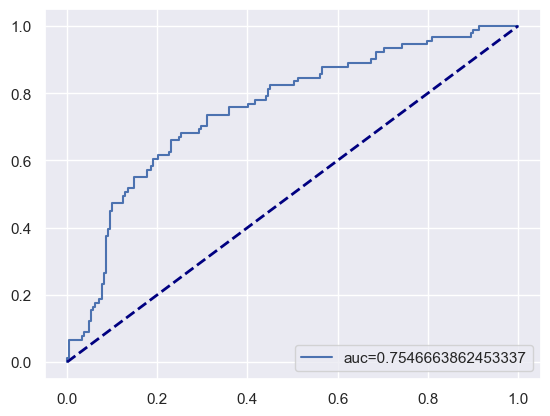
* **accuracy and F1 score**

****

* **Confusion matrix and Feature importance**

****

* **Decision Tree**
* **ROC curve**

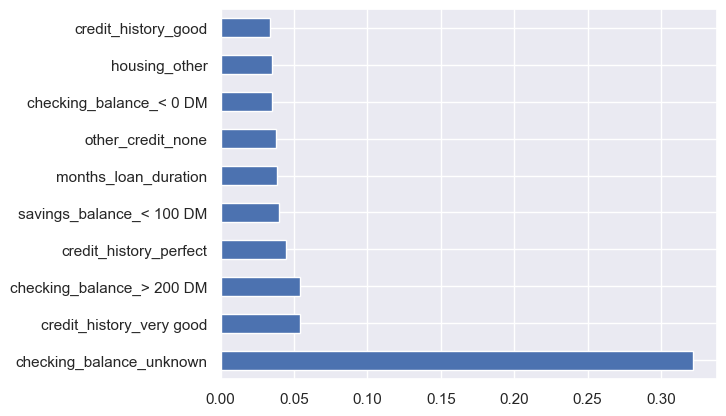
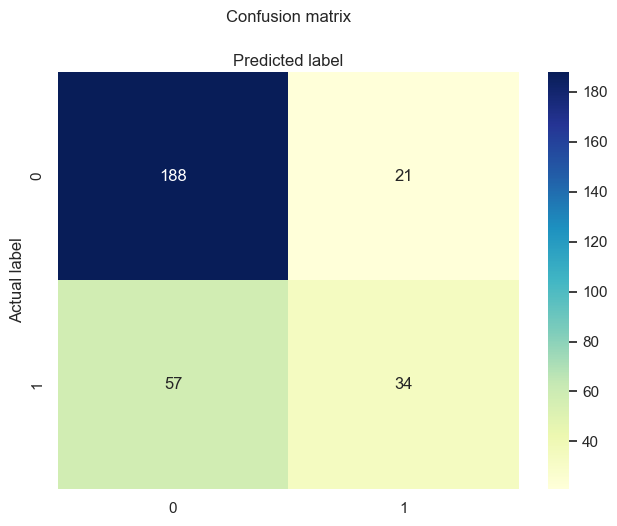
****

**XGBoost:**

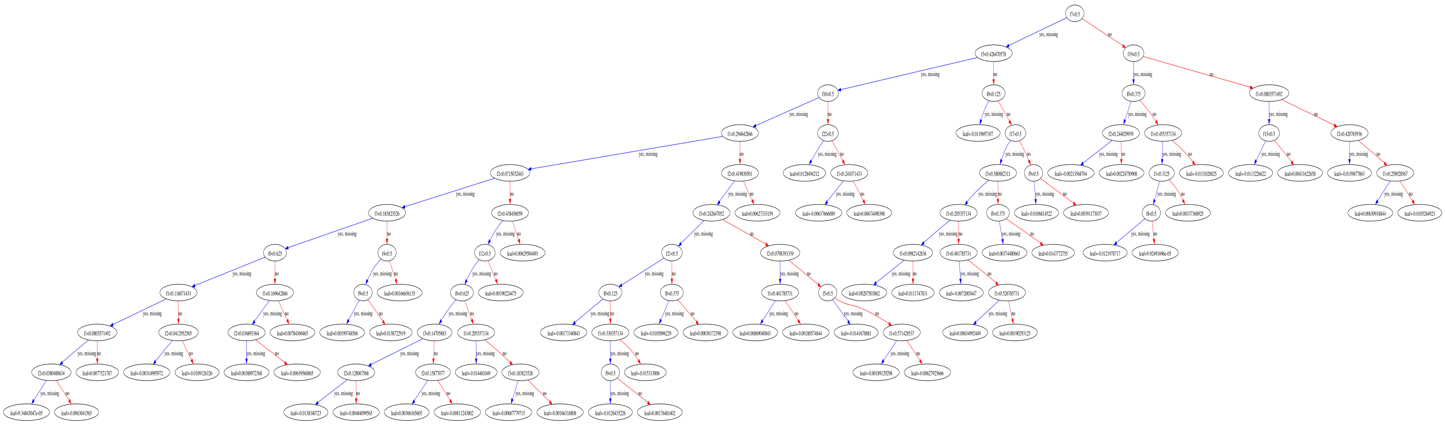
* **accuracy and F1 score**

****

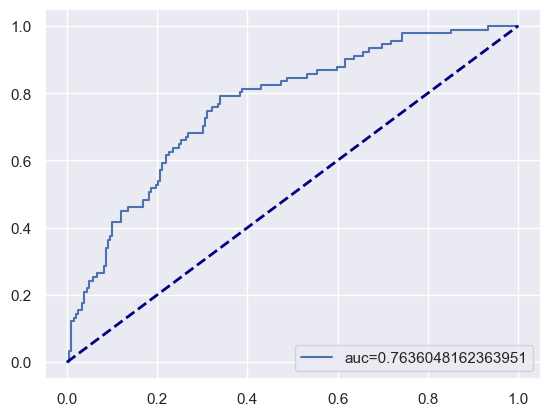
* **Confusion matrix and Feature importance**

****

* **Decision Tree**

****

* **ROC curve**

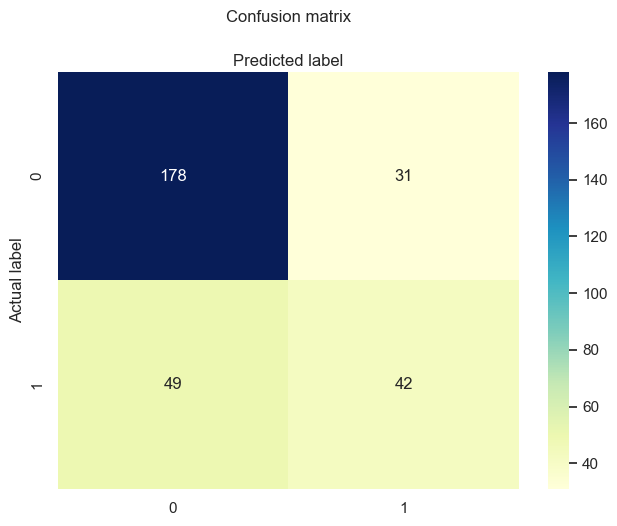
****

**K-Nearest Neighbors (KNN):**

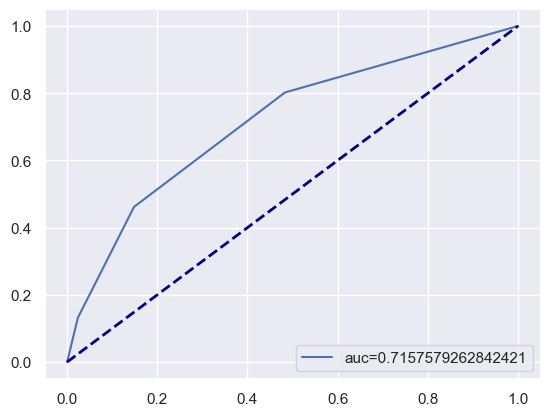
* **accuracy and F1 score**

****

* **Confusion matrix**

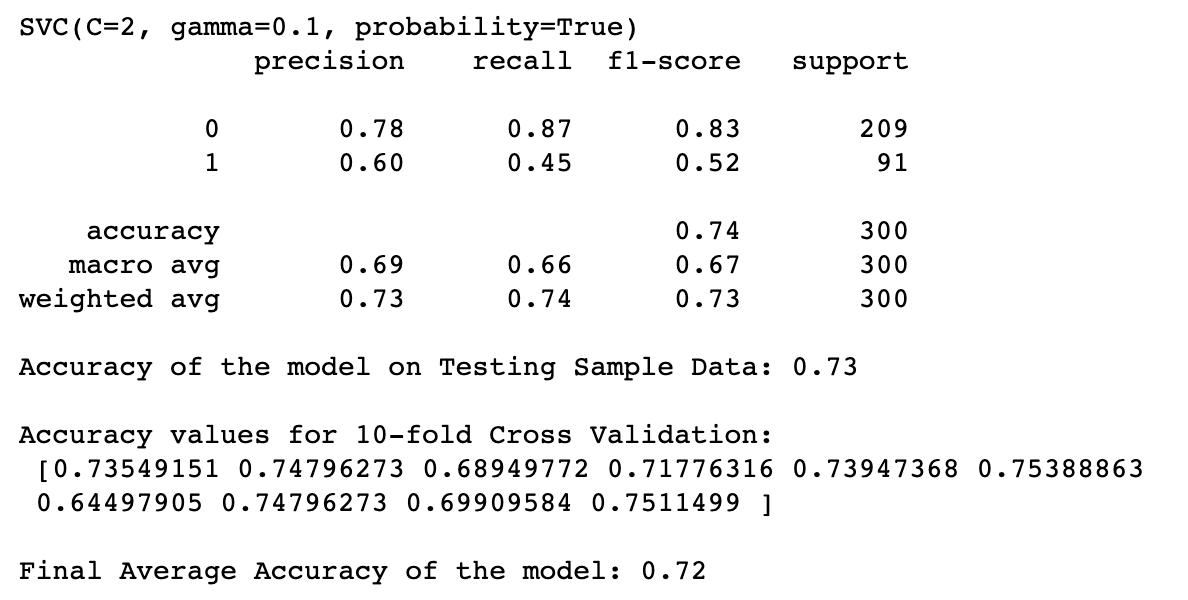
****

* **ROC curve**

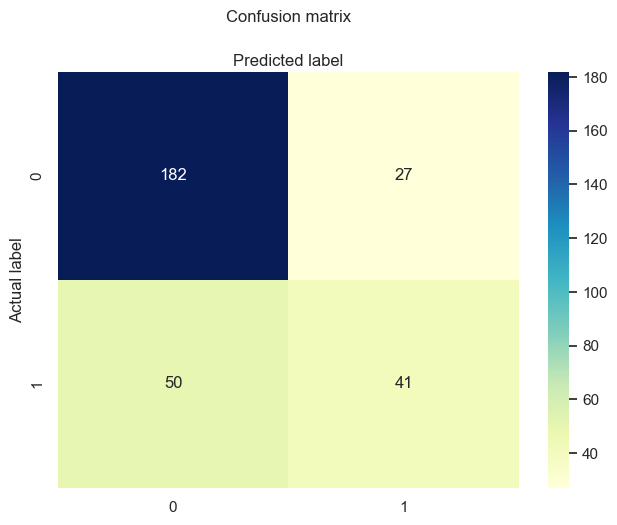
****

**Support Vector Machines (SVM):**

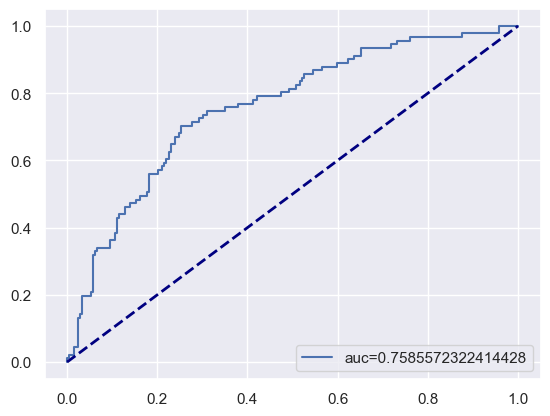
* **accuracy and F1 score**

****

* **Confusion matrix**

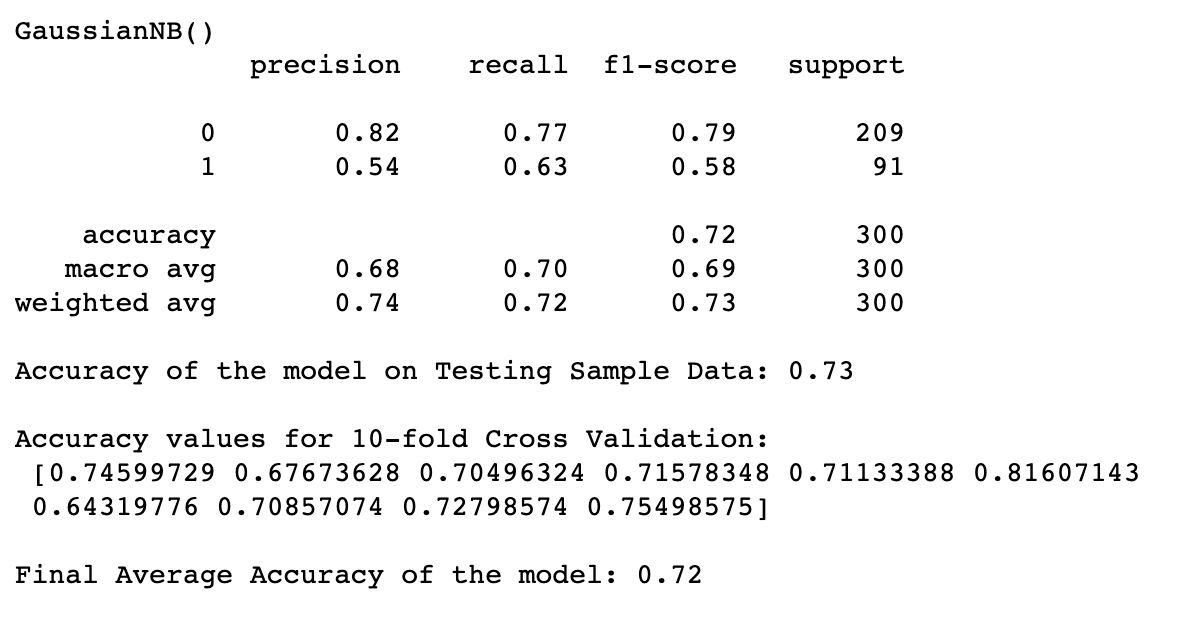
****

* **ROC curve**

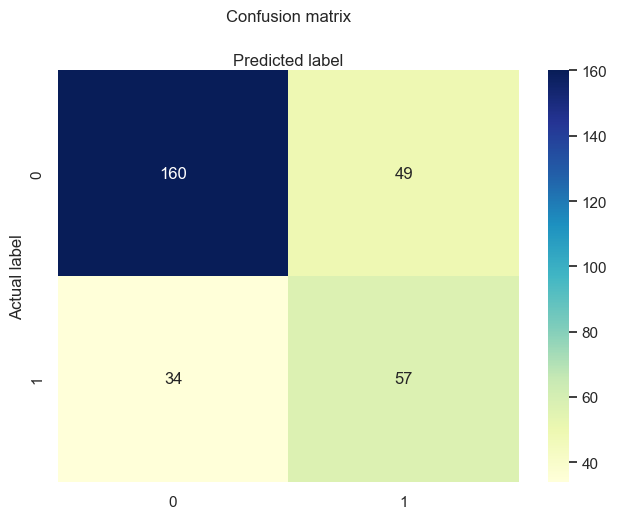
****

**Naive Bayes:**

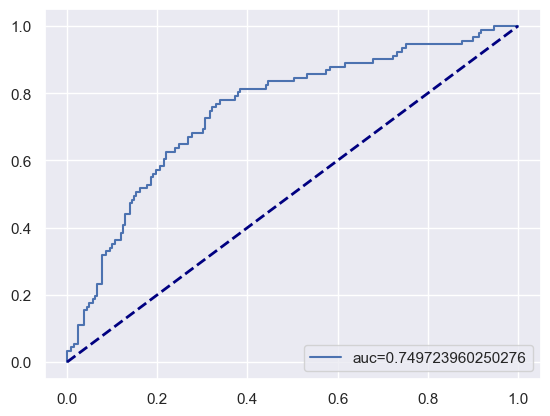
* **accuracy and F1 score**

****

* **Confusion matrix**

****

* **ROC curve**

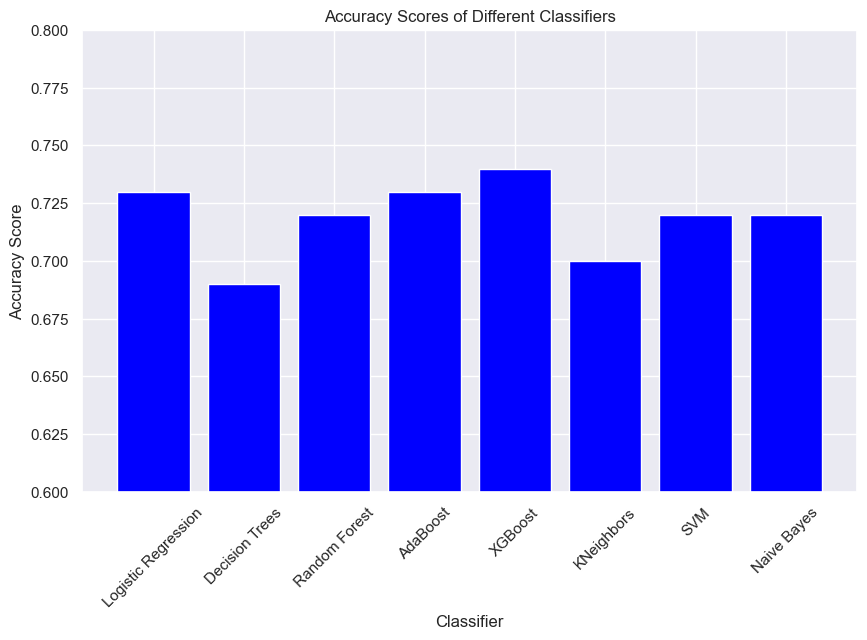
****

By applying this structured evaluation approach to each model, we gained a comprehensive understanding of their respective strengths, weaknesses, and suitability for our classification task.

# Discussion:

**Classifier Performance Evaluation**

In our study, we evaluated the performance of various classification algorithms to determine their effectiveness in predicting loan credit risk. The accuracy scores obtained from each classifier provide valuable insights into their capabilities. Below, we present a detailed analysis of the results along with a visual representation of the accuracy scores.

****

We used the following classification algorithms and computed their accuracy scores:

The bar chart above visually represents the accuracy scores of each classifier. The y-axis represents the accuracy score, while the x-axis corresponds to the different classifiers. The chart provides a quick overview of how each classifier performed in terms of accuracy.

Summary and Insights

**From the accuracy scores obtained, several key observations and insights emerge:**

1. XGBoost Leads the Pack: The XGBoost classifier exhibited the highest accuracy score of 0.74. This signifies its strong predictive power and suggests that it can effectively distinguish between different credit risk categories.
2. Consistent Performance: Both Logistic Regression and AdaBoost achieved accuracy scores of 0.73, demonstrating their robustness in capturing relevant patterns within the data. These classifiers offer competitive accuracy and can be considered suitable choices.
3. Competitive Mid-Range Performers: Random Forest, SVM, and Naive Bayes consistently delivered accuracy scores of 0.72. This indicates that these models maintain a commendable level of predictive performance and could serve as viable options.
4. Balanced Accuracy: KNeighbors achieved an accuracy score of 0.7, positioning itself as a balanced performer. While not the highest, its accuracy is still within a reasonable range, suggesting its suitability for the task.
5. Decision Trees Performance: Decision Trees, while achieving an accuracy score of 0.69, demonstrated slightly lower performance compared to some other models. However, it still provides a foundation for understanding feature importance and decision pathways.
6. After thorough analysis, the most influential predictor variables for credit risk prediction have been identified: 'age','checking\_balance', 'credit\_history', 'savings\_balance', and 'months\_loan\_duration'. These variables consistently demonstrate high importance across multiple algorithms, making them the final set of predictors for enhancing the accuracy of our credit risk prediction model.

**Considerations for Model Selection:** The selection of the most appropriate classifier should encompass factors beyond accuracy alone. Model complexity, interpretability, computational efficiency, and the specific requirements of the task play crucial roles. The decision should be driven by a comprehensive assessment that balances accuracy with other relevant considerations.

In conclusion, our evaluation of different classifiers provides valuable insights into their respective strengths and weaknesses. The XGBoost classifier emerged as the top performer in terms of accuracy, closely followed by Logistic Regression and AdaBoost. The choice of classifier should align with the specific goals and characteristics of the classification task.

For the final model selection, we have opted for Support Vector Machines (SVM). The decision to choose SVM is influenced by several factors, including its commendable accuracy performance, its compatibility with high-dimensional data, and its efficiency in training and prediction. SVM's swiftness in processing this dataset's dimensionality makes it a promising choice for our predictive task.

**Ref:**  *Classification Evaluation Metrics: Accuracy, Precision, Recall, and F1 Visually Explained*. (2022, June 7). Context by Cohere. https://txt.cohere.com/classification-eval-metrics/

Gandhi, R. (2018, July 5). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Medium. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

**Limitations of the research:** Machine learning algorithms are limited to the dataset used to train and test the model. Data is governed by data protection laws making it challenging to access primary data for the purpose of research. This limits the generalization of the model as it is specific towards the dataset used to train and test the model. It would be beneficial to look comprehensively at the main features that are relevant to the characteristic that drive default and can be applied. Further limitation is in reference to the variables provided in the dataset, although the dataset is open due to data protection laws some factors may not be available to the public and these may have had an impact on the predictions of the probability of default. Lastly, the research focused on the probability of default in a default state however default loans may still be recovered during the recollection process.

**Recommendations and Future Work:** This research explores using machine learning algorithms to improve the accuracy of predicting loan default. This model will be instrumental to mobile lending institutions in evaluating their customer credit risk. The best performing model in the research which is SVM achieves an accuracy of about 72%. This is a fair performance and can further be improved through different methods of parameter tuning and feature selection which may possibly yield improvements in the model performance. It may also be beneficial to do a cross validation with other sources of open dataset as they become more accessible to compare the performance of the model. Since the research is also limited to the probability of default in a default state , further exploration may be made in determining the expected return of the loan based on borrower’s characteristics , loan characteristics the recollection of loans processes.

# Conclusions:

This report delved into the significance and applicability of utilizing machine learning for loan prediction. We explored existing methodologies and datasets employed in loan eligibility prediction, and examined the potential of AI in streamlining this process. Furthermore, we conducted a comparative analysis of several machine learning models, utilizing publicly available datasets, to predict loan outcomes. Throughout this project, we gained insights into the impact of various features on model predictions, and how specific attributes wield more influence on decisions compared to others.

The classifiers used in this analysis included Logistic Regression, Decision Trees, Random Forest, AdaBoost, XGBoost, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes.