CREDIT RISK MODELLING

GROUP 1:

Rahul Upadhyay - 202218003
Shreya Arora - 202218032
Muskan Khare - 202218037
Dhruv Solanki - 202218053
Jatan Sahu - 202218061

PROBLEM STATEMENT

- Develop a risk-based strategy to identify creditworthy customers for fresh loans using machine learning on the German credit dataset.

- Set a risk threshold to balance loan approval and risk mitigation, ensuring controlled false positives and minimizing false negatives.

DATA DESCRIPTION

- The German credit dataset consists of 1,000 instances with 20 attributes, including information about credit applicants such as checking account status, credit history, purpose of the loan, credit amount, savings account/bonds, employment status, and more.

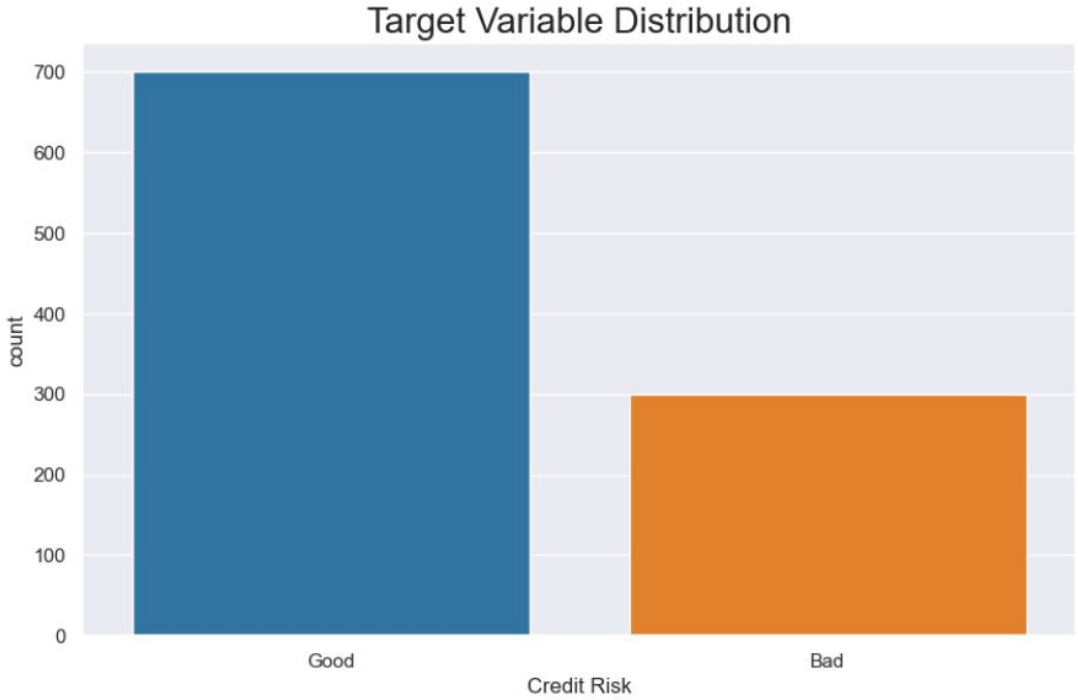
- The dataset includes qualitative and numerical attributes that cover various aspects relevant to creditworthiness, such as the duration of the loan, personal status, property ownership, number of existing credits, age, and other factors that impact credit risk assessment.

- The target variable represents the credit rating of the applicant, categorized as either "good accounts" (1) or "bad accounts" (2), allowing for the development of predictive models to assess credit risk and make informed decisions on loan approvals.

Target variable vs Credit Risk

- There are 700 instances where an applicant was classified as good and 300 instances where am applicant was classified as bad.

- So, overall there's 3/10 chance for an applicant for defaulter.

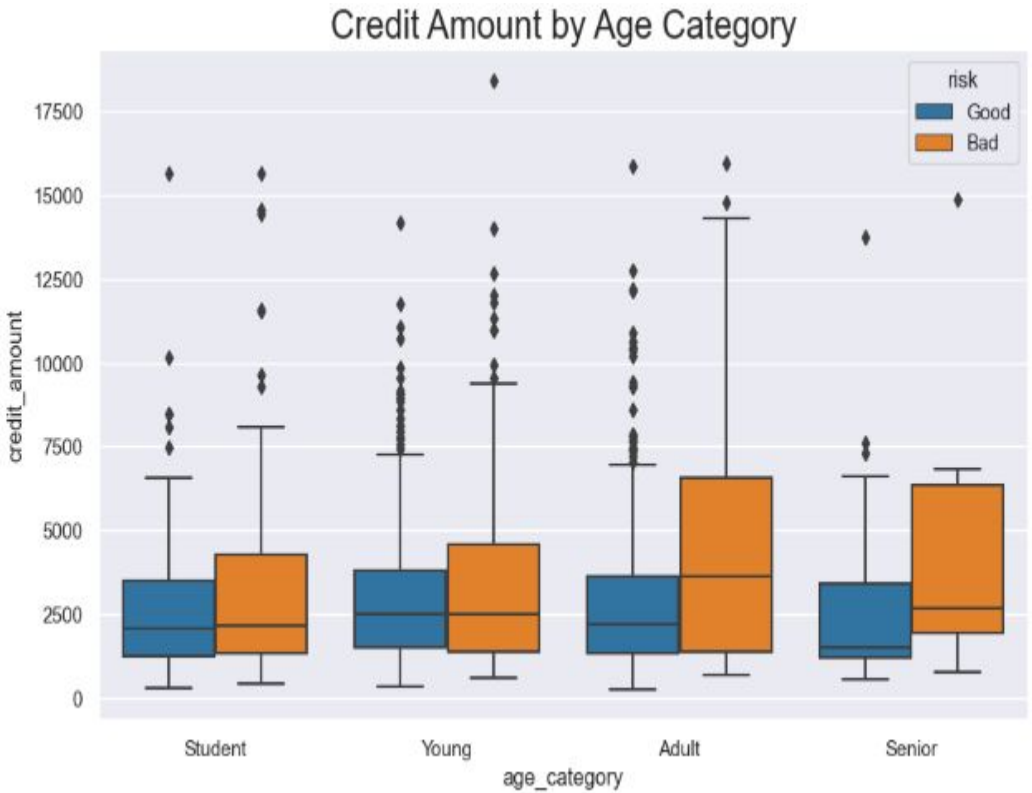


Distribution of Numerical Values

- All graphs have a positive skew indicating that the mean is greater than the median.
- Applicants between the ages of 20 to 30 are more likely to apply for a loan.
- Applicants are less likely to apply for a high credit loan.
- More loans have been paid off around 20 months after being issued.
- The bank is more likely to receive applicants between the ages of 20 and 30 and request loans between 250 and 2500 DM.

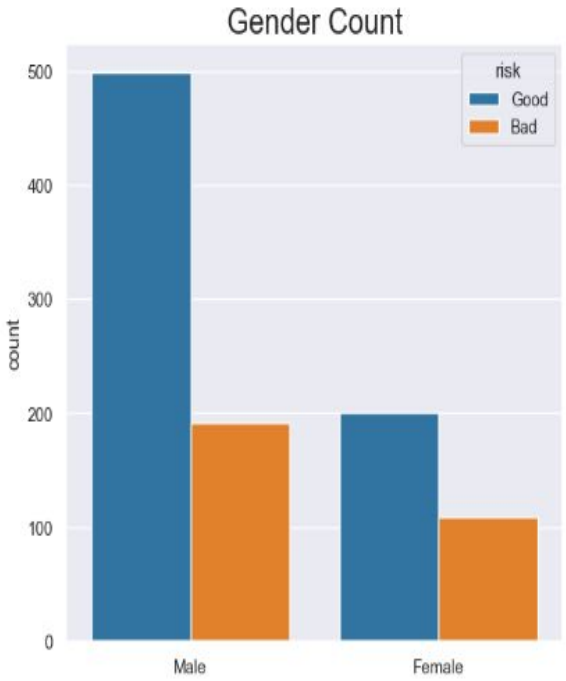
Distribution of Credit by Age

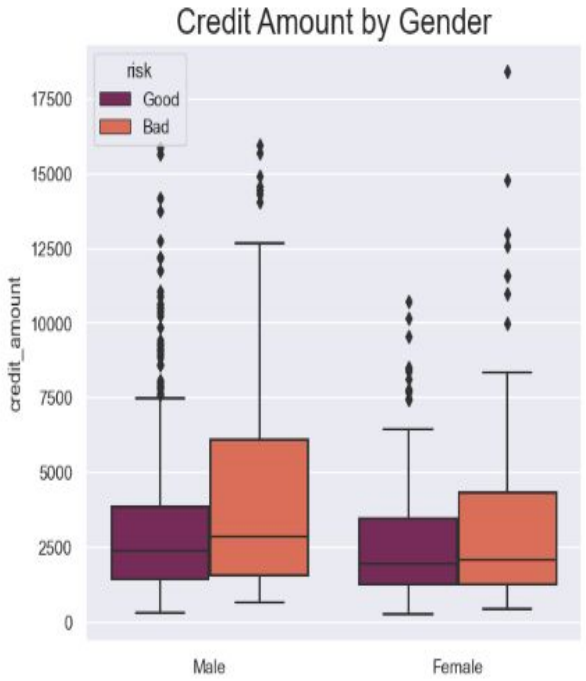
- More than 50% of the applicants with credit amounts below 5,000 DM are classified as good.
- Adults with loan credit greater than 5,000 DM are more likely to be classified as bad.
- Students and Young applicants are most likely to apply for loans with a credit amount of less than 5,000 credit amount.



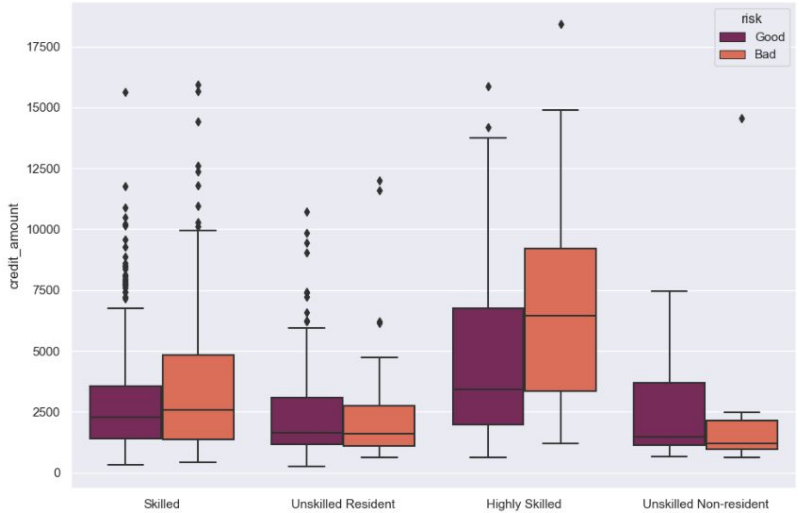
Distribution by Gender

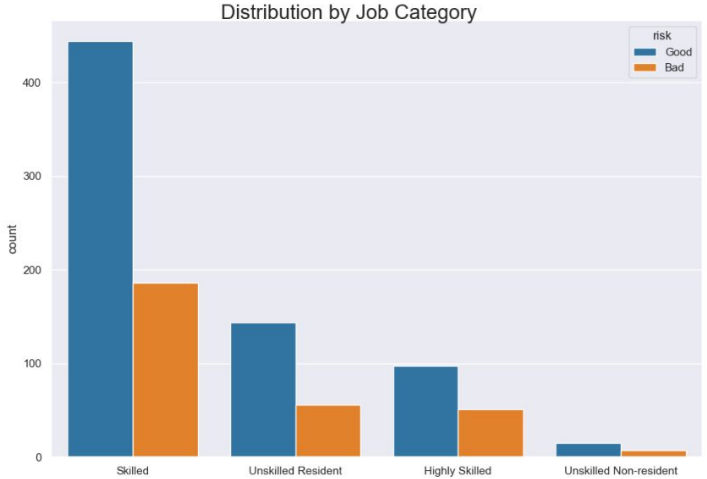
- There are 2x more male applicants than females in the data
- About 2/5 of male applicants and 1/2 of female applicants are classified as bad

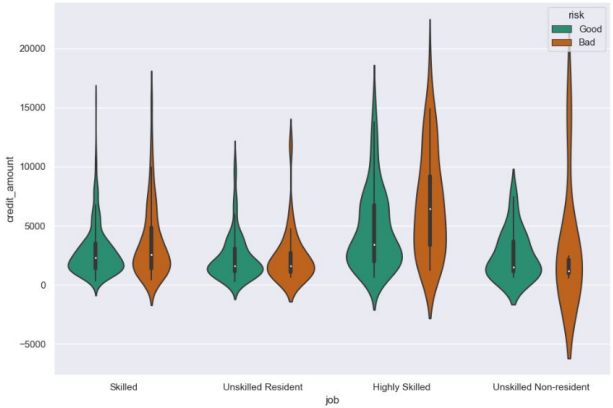




Distribution by Job Category







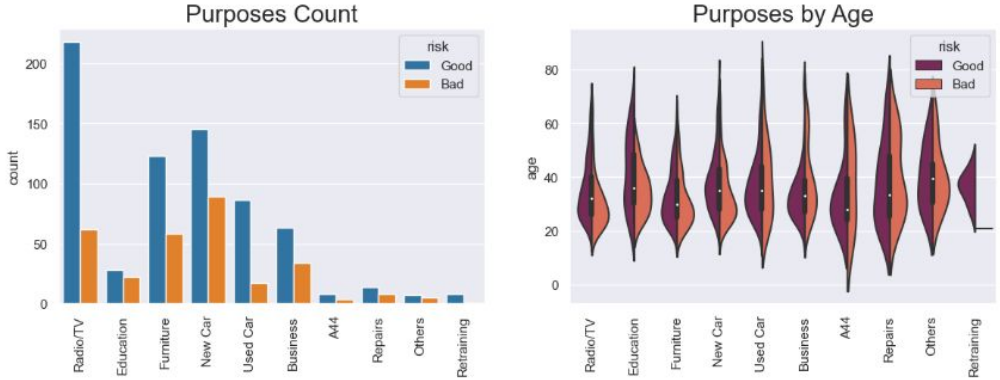
- There are 2x more applicants with skilled jobs that are classified as good
- More than 50% of applicants are under the skilled and unskilled and resident job categories
- Applicants that are highly skilled are more likely to take out larger loans

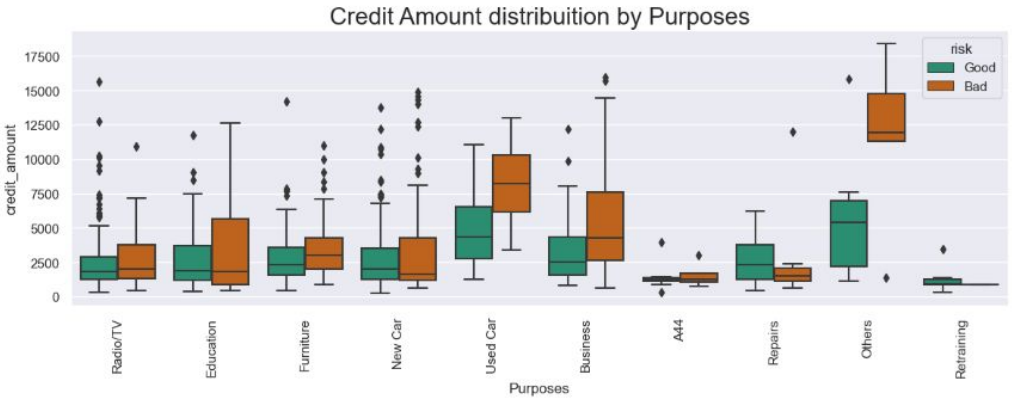
Distribution by Savings Account

- Applicants with little or no saving accounts are more likely to apply for loans
- The majority of the applicants are in the little category
- 50% of the applicants in the little category are between the age range of 25 and 45
- Applicants with moderate, quite rich, and rich savings accounts are more likely to be classified as good
- Applicants with little and no savings accounts with a credit amount loan that exceeds 5,000 DM are more likely to be classified as bad

Distribution by Purpose

- A large portion of applicants requested loans for buying cars, radios/tv's.
- More than half of the applicants applied for loans less than 5,000 DM.
- Applicants with high credit loans are more likely to be classified as bad.

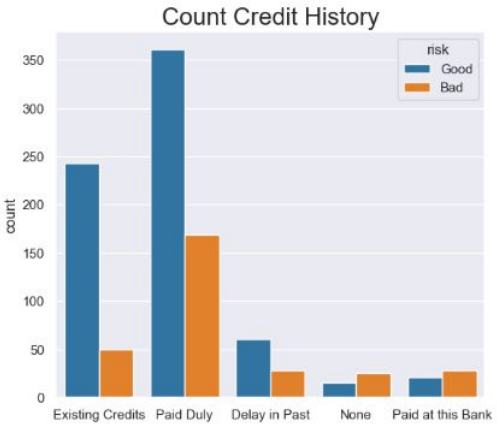


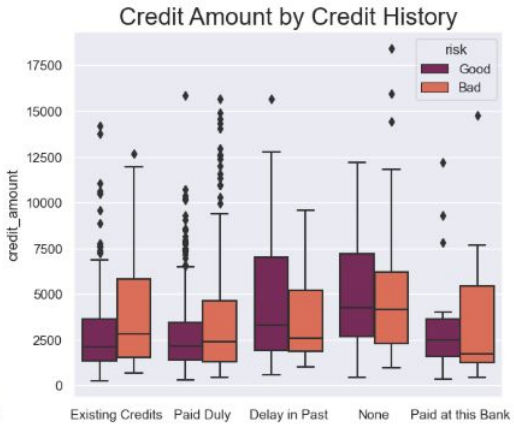


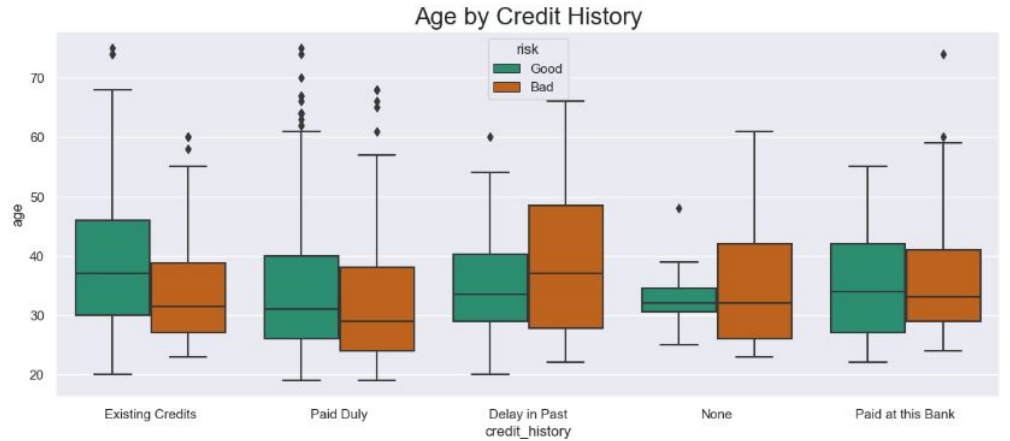
Credit History Distribution

- People with existing credits or whom who have duly paid their credits are more likely to apply for a loan.

- Such people are more likely to be classified as defaulters if the credit amount is more than 5000 DM.

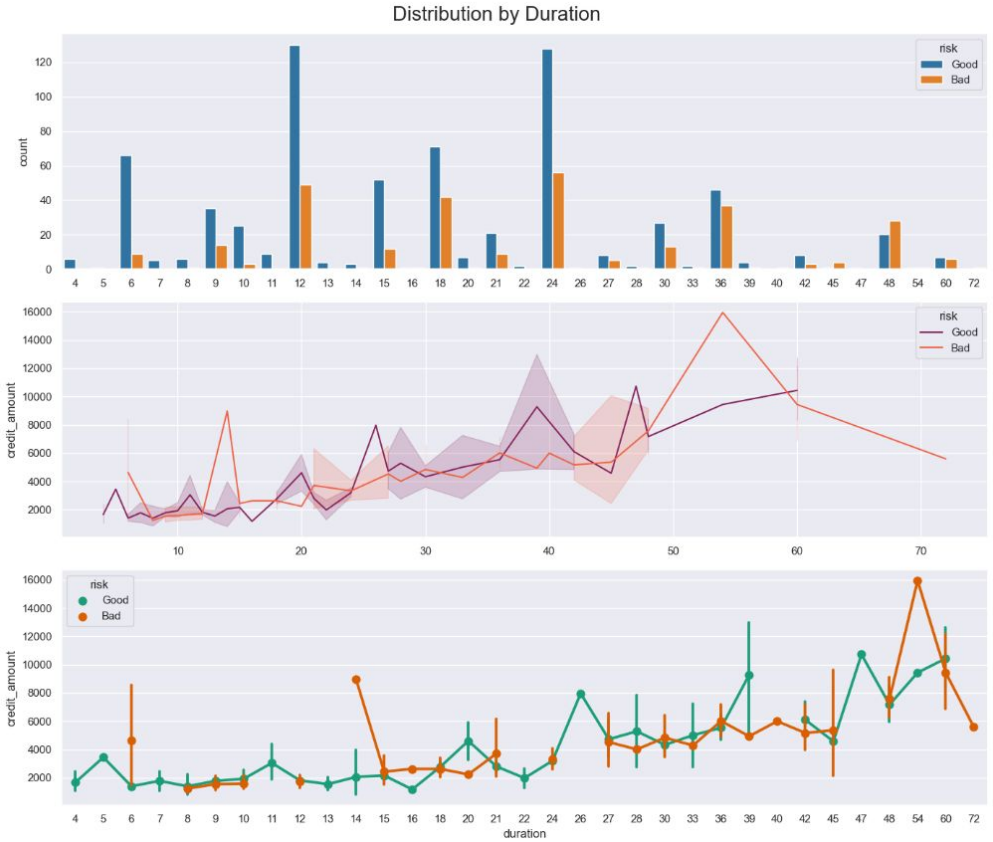






Distribution by Duration

- Most of the loans issued had a duration of 12 and 24 months.
- Most applicants that repaid their loans within 24 months are classified as good.
- Most applicants with a loan duration that exceeds 24 months are classified as bad.



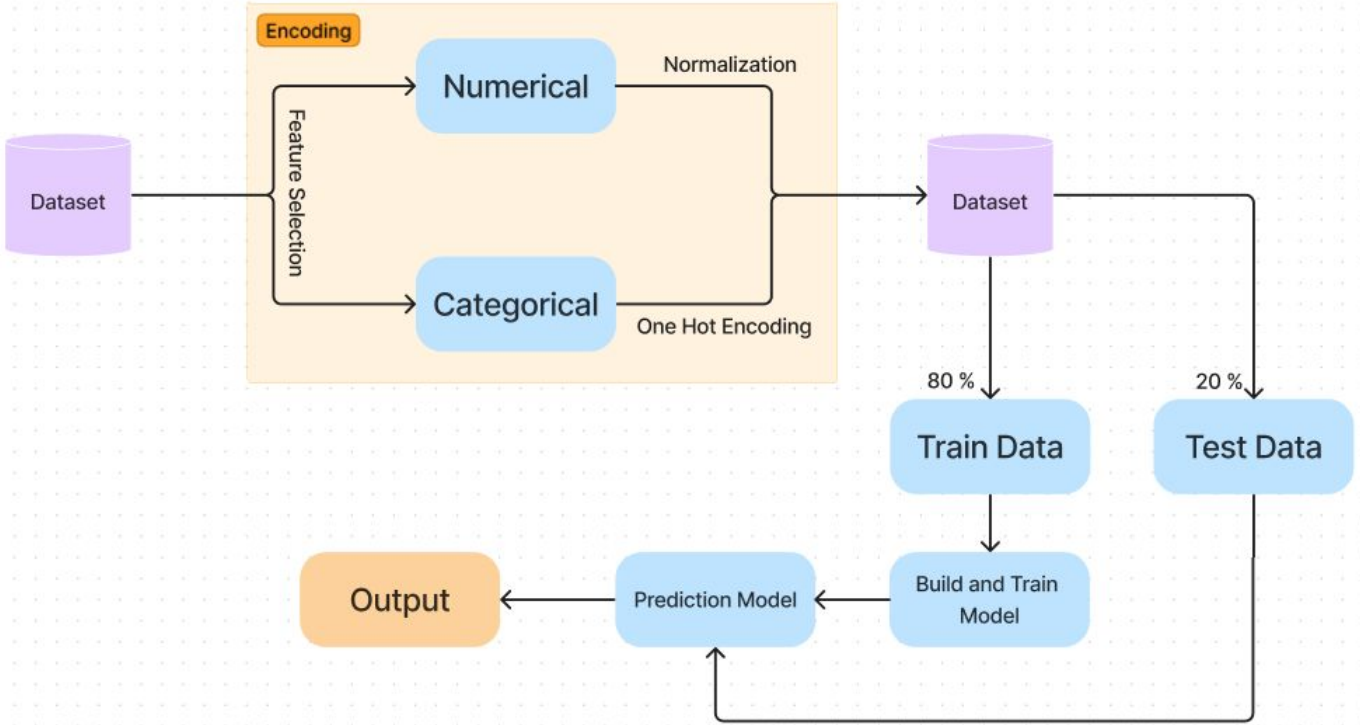
Inference from other Variables

- More than 50% of the people who applied for a loan own a house. Around 25% of them are likely to be classified as Bad accounts. The percentage of bad accounts for people who live on rent or for free is much higher.
- People with no savings account are mostly classified as Good accounts compared to people with little or moderate savings.
- People who are already co-applicants or guarantor for other debts do not apply for loans.
- People who don't own property are much more likely to be defaulters compared to people who own at least some property.

Conclusion from Analysis

- Loans with durations terms less than 24 months are more likely to be repaid.
- It’s safer to issue loans with a credit amount less than 5,000 DM and a duration term of fewer than 24 months.
- Applicants who own property show that they are financially independent and are better candidates for a loan.
- Applicants with skilled and highly skilled jobs are safer candidates to issue loans.
- Car loans are the most common loans with a high gain to loss ratio issued by the bank (most profitable loan).
- It’s more profitable to issue loans that are less than 2,500 DM than higher credit loans that are less likely to be repaid.

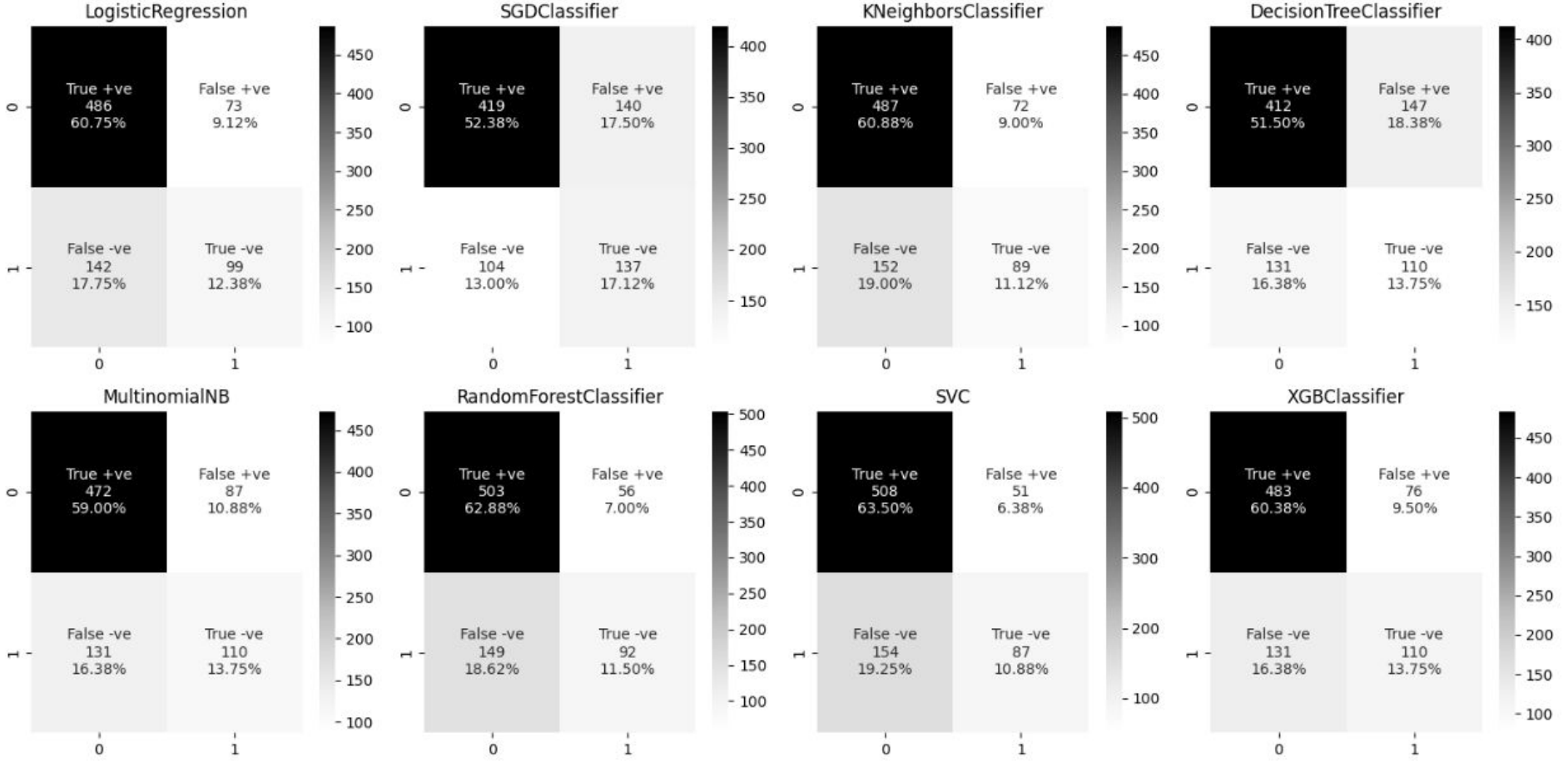
PIPELINE



Models Used

- Logistic Regression
- SGD Classifier
- KNN Classifier
- Decision Tree
- Naive Bayes
- Random Forest
- SVM Classifier
- XGB Classifier

Confusion Matrices for Different Models



Classification Report

- Considering Accuracy and precision, SVM is the best model.
- Random Forest model is the second best model, with a better recall value.

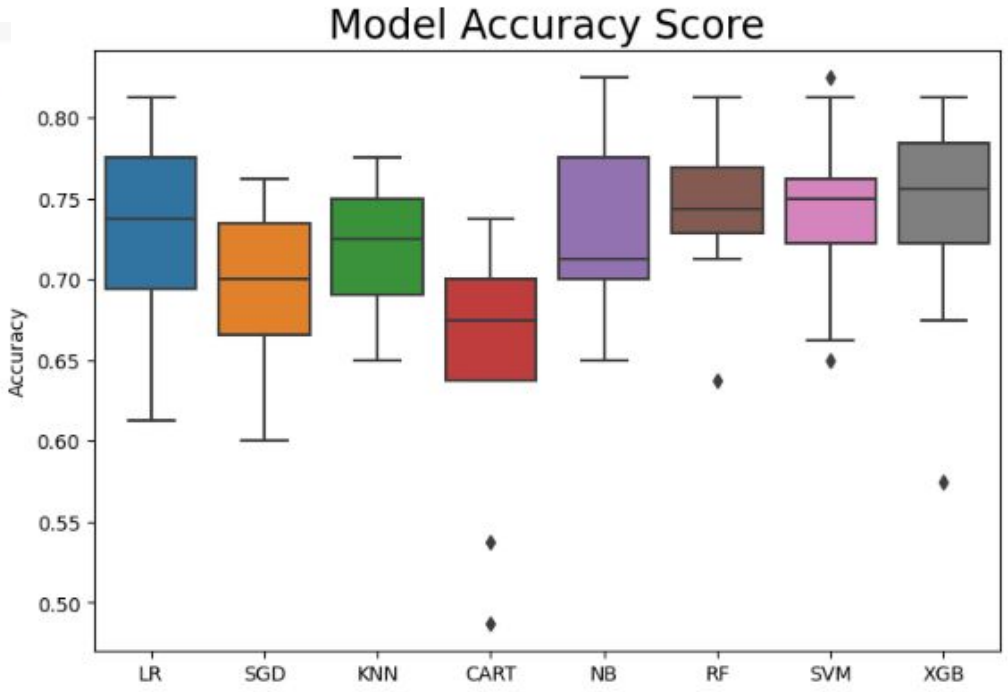
  However, the recall is comparatively low for both these models.

- SGD gives the best recall value but compromises the precision.
- XGB can be considered the best model as it gives accuracy very close to SVM and also gives a better F1 score.

Model	Accuracy	Precision	Recall	F1
LR	0.73125	0.579468	0.418066	0.480578
SGD	0.69500	0.517079	0.568728	0.510933
KNN	0.72000	0.562600	0.367792	0.436600
CART	0.65250	0.432721	0.460667	0.442371
NB	0.72750	0.565938	0.458155	0.503404
RF	0.74375	0.642824	0.382729	0.468993
SVM	0.74375	0.656374	0.368077	0.462303
XGB	0.74125	0.606312	0.458178	0.514269

CONCLUSION

Model	Accuracy	Precision	Recall	F1
SVM	0.775	0.659091	0.491525	0.563107



THANK YOU!