Loan Approval Prediction

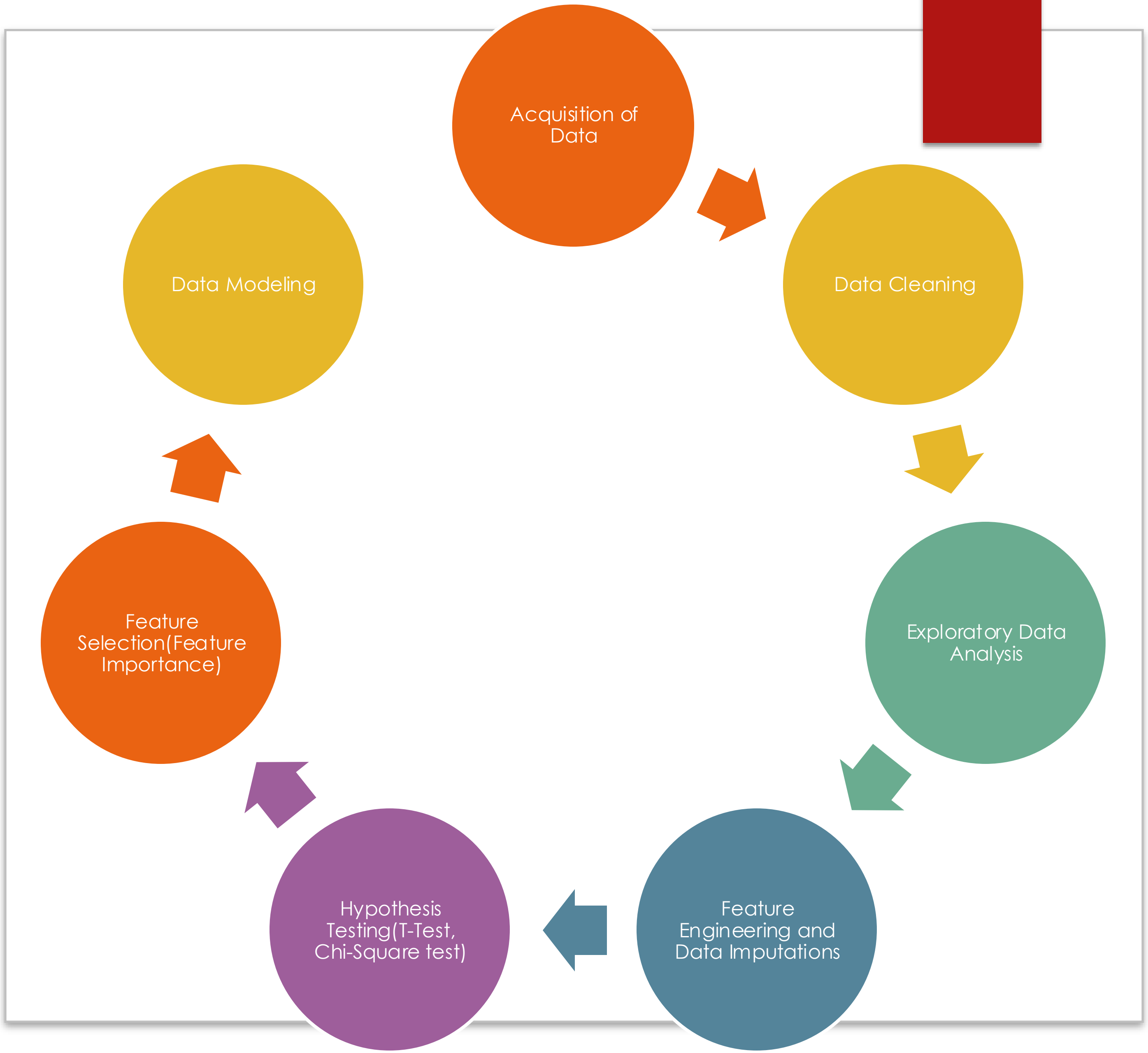
SAMAYAK MALHOTRA

Agenda

To create a Loan Approval prediction model that accurately predicts if the loan for a customer would be approved.

Generate a UI using Tkinter/Flask.

Data Flow/ Pipeline



Acquisition of Data

Data Cleaning

Exploratory Data Analysis

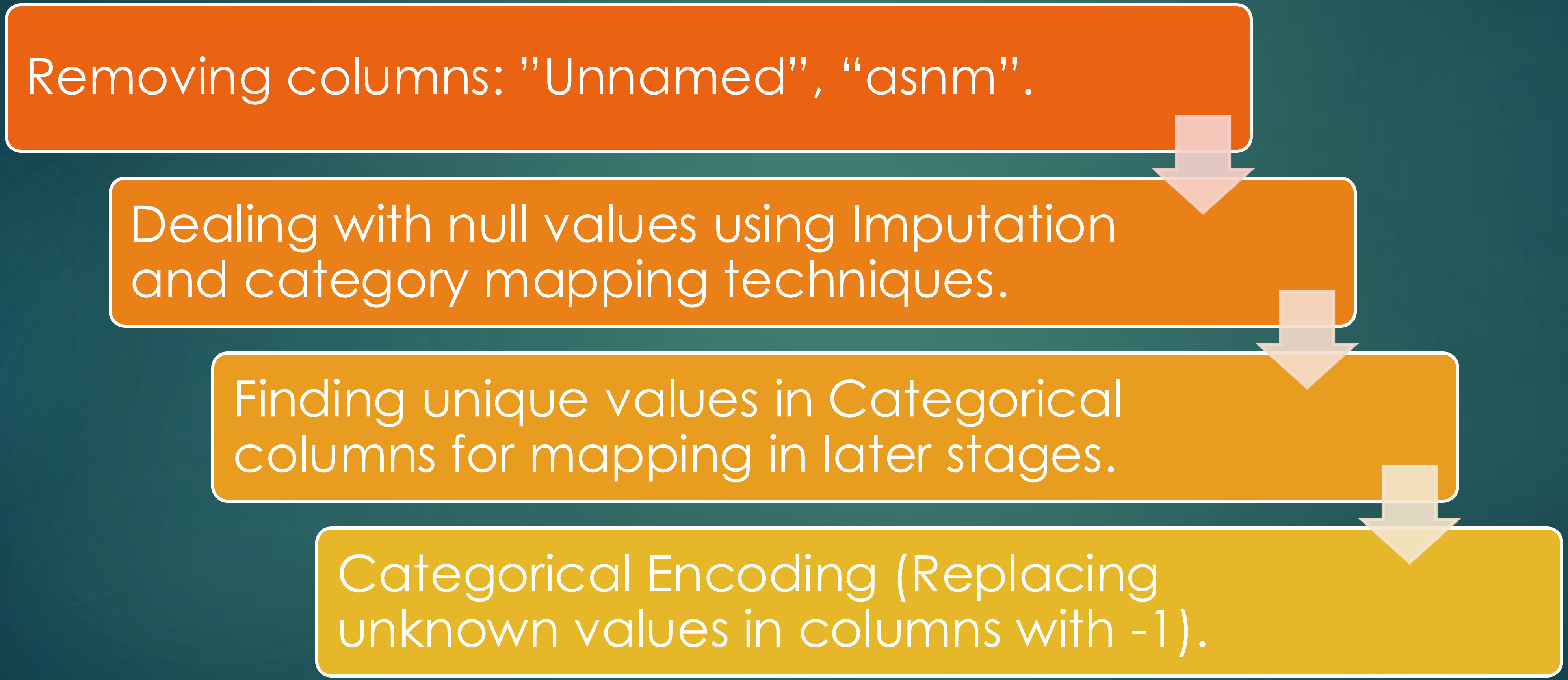
Feature Engineering and Data Imputations

Hypothesis Testing(T-Test, Chi-Square test)

Feature Selection(Feature Importance)

Data Modeling

Data cleaning and Pre-processing



Removing columns: ”Unnamed”, “asnm”.

Dealing with null values using Imputation and category mapping techniques.

Finding unique values in Categorical columns for mapping in later stages.

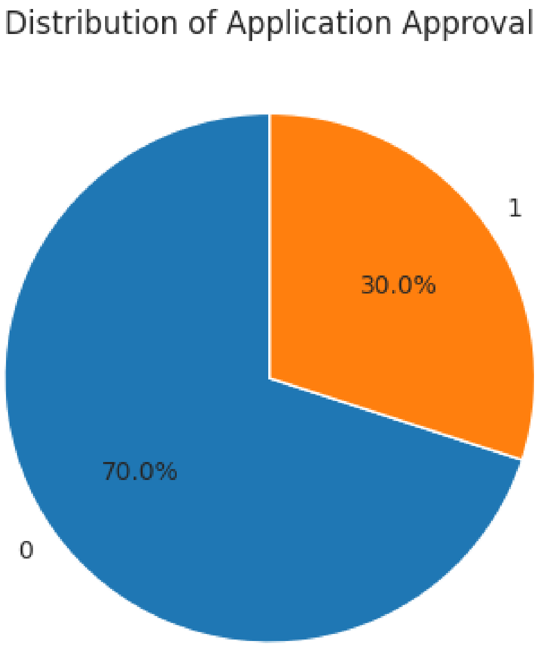
Categorical Encoding (Replacing unknown values in columns with -1).

Data Exploration

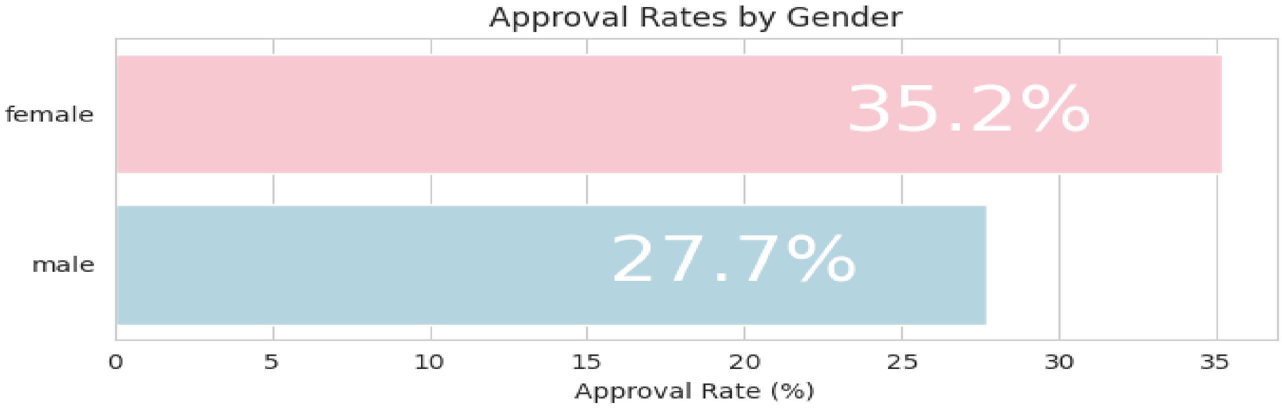
DATA TYPES

VISUALIZATIONS

DATA CLEANING & PRE-PROCESSING



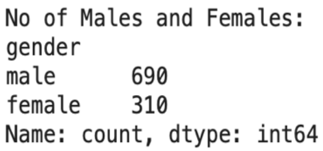
1= Male, 0= Female



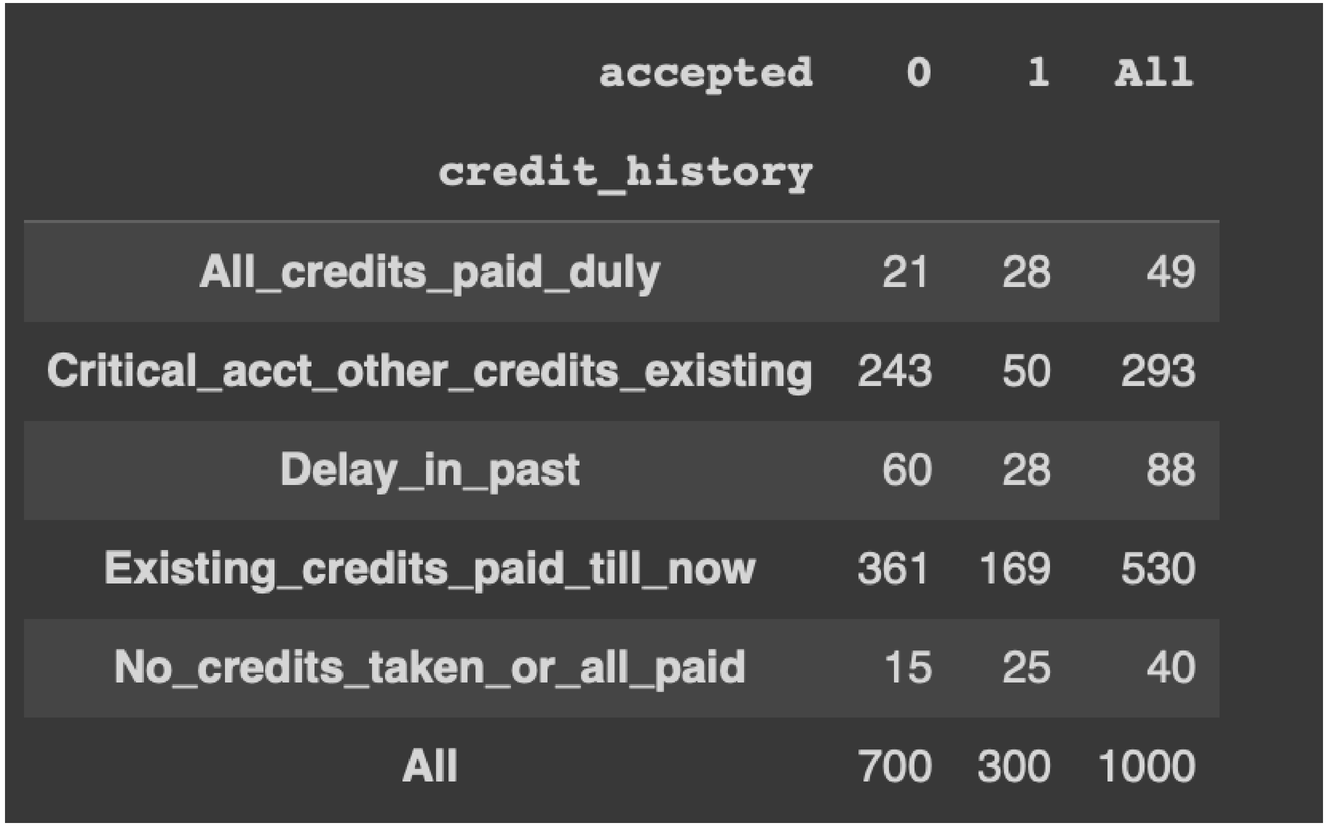
Loan Approval Status: About 1/3rd of applicants have been granted loan.

Sex: There are more Men than Women (approx. 2x) and more females are getting the loans approved.

Martial Status: About 50% of the population in the dataset is Single men; Single men are more likely to be granted loans.



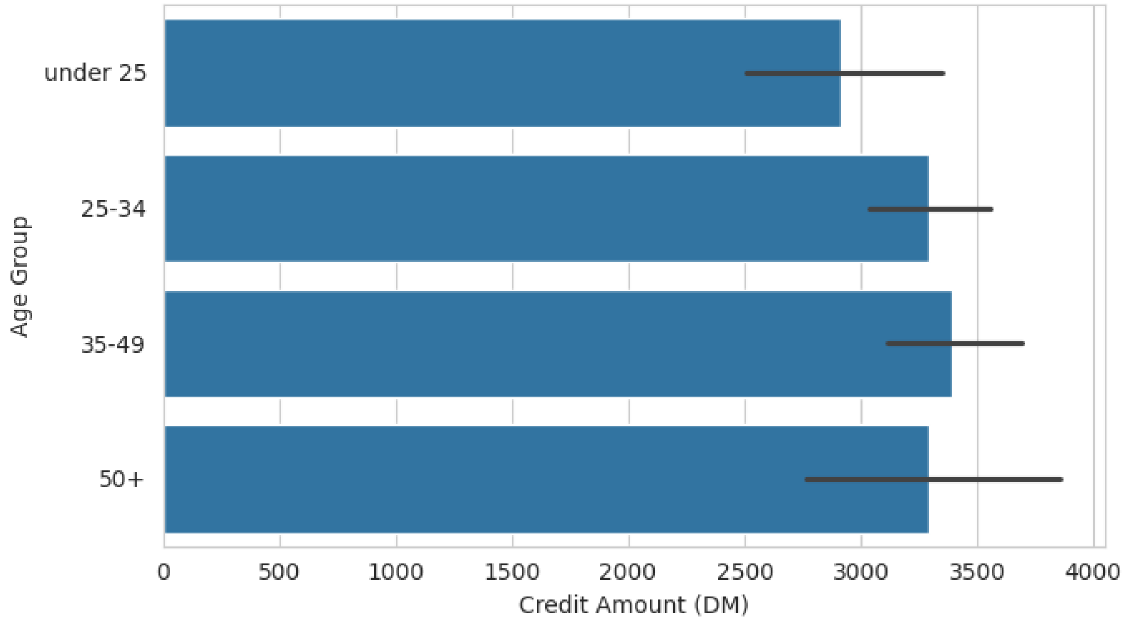
```
No of Males and Females:
gender
male      690
female    310
Name: count, dtype: int64
```



accepted	0	1	All
credit_history			
All_credits_paid_duly	21	28	49
Critical_acct_other_credits_existing	243	50	293
Delay_in_past	60	28	88
Existing_credits_paid_till_now	361	169	530
No_credits_taken_or_all_paid	15	25	40
All	700	300	1000

It can be seen that if the customer has paid all the credit till date, then the chances for getting loan approved is higher.

Hypothesis: Applicants under the age of 25 request for less Loans



Null Hypothesis (H0): There is no difference in the loan amounts taken by people under 25 and those over 25.

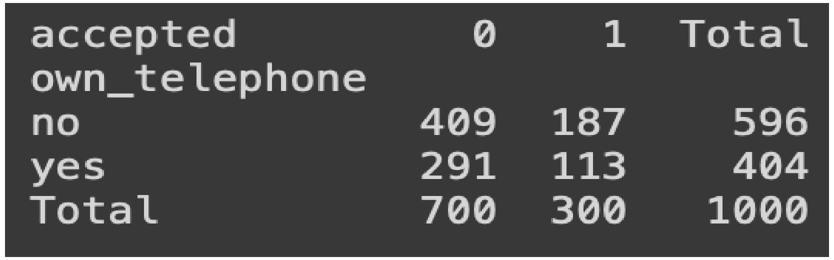
Alternative Hypothesis (H1): There is a difference in the loan amounts taken by people under 25 and those over 25.

According to the T-test, P-Value> 0.05. i.e There is no significant evidence stating that Age group under 25 and over 25 take almost the same loan amount. So, the Null hypothesis cannot be rejected.

What’s the effect of owning a telephone on the likelihood of a credit application being accepted? (Chi-square test)

H0: Owning a phone has no effect on getting the credit approval.

H1: Owning a phone affects getting the credit approval.



```
accepted          0    1  Total
own_telephone
no              409  187    596
yes             291  113    404
Total           700  300   1000
```

Out of 404 participants owning a phone only 113 got the loan accepted(27%). and Out of 596 not owning a phone and 187 getting it accepted (31%).

This shows that this data is probably from the 90's when people didn't own phones.

Note: The chi-square test is for categorical data, while the t-test is for continuous data.
The chi-square test assesses relationships or associations between categorical variables, while the t-test compares means of continuous data.

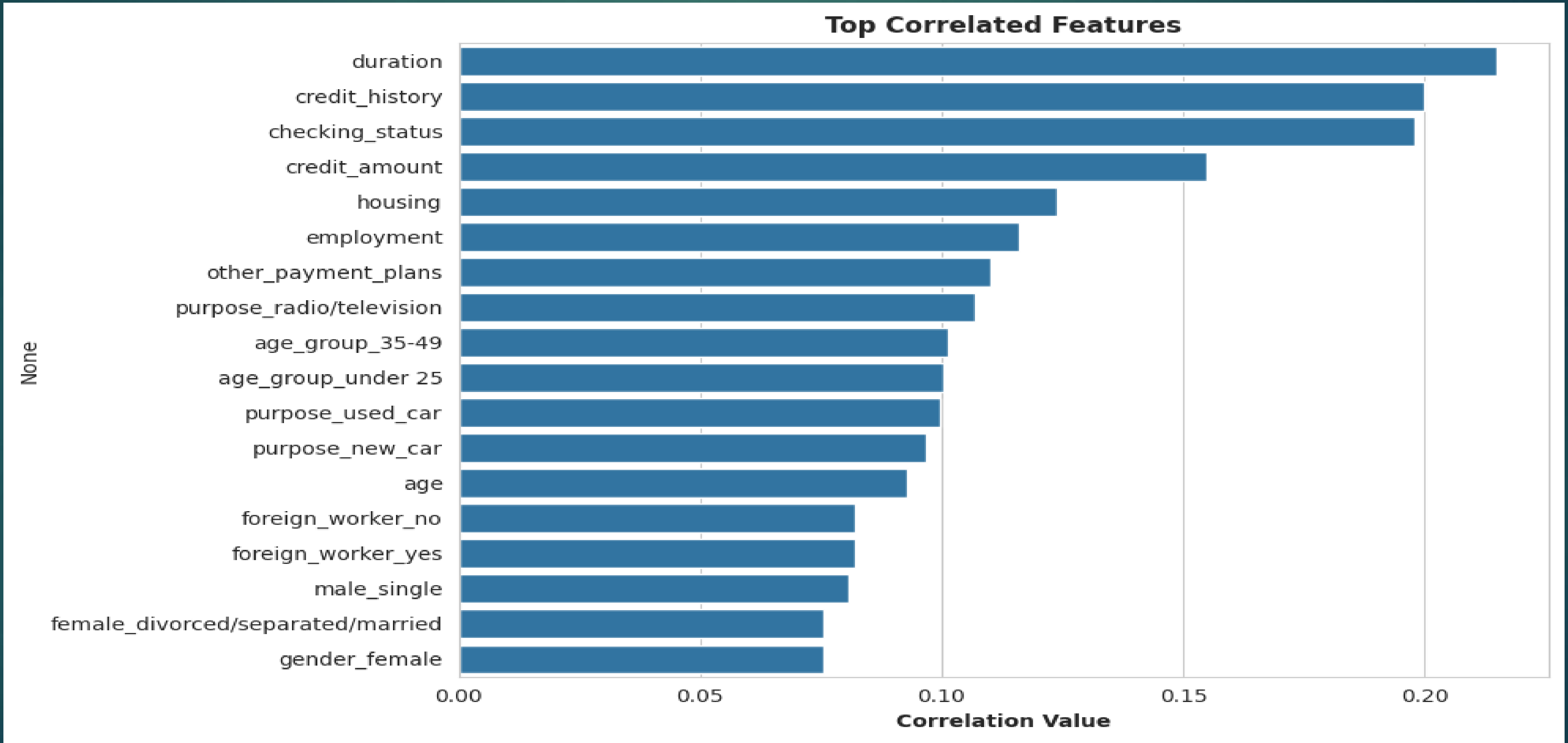
Feature Enginnering

- Techniques used (e.g., creating new features, handling missing values, encoding)
- Feature selection method (Feature Importance)

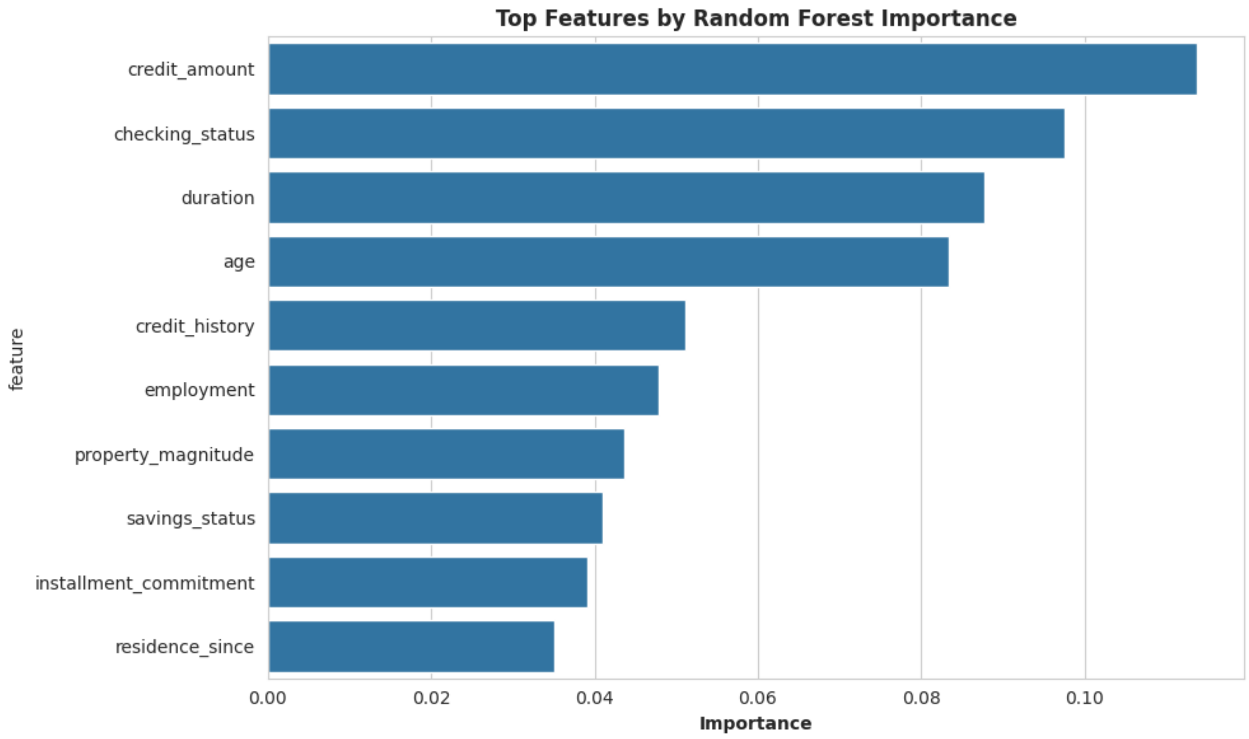
Model Interpretation and Insights

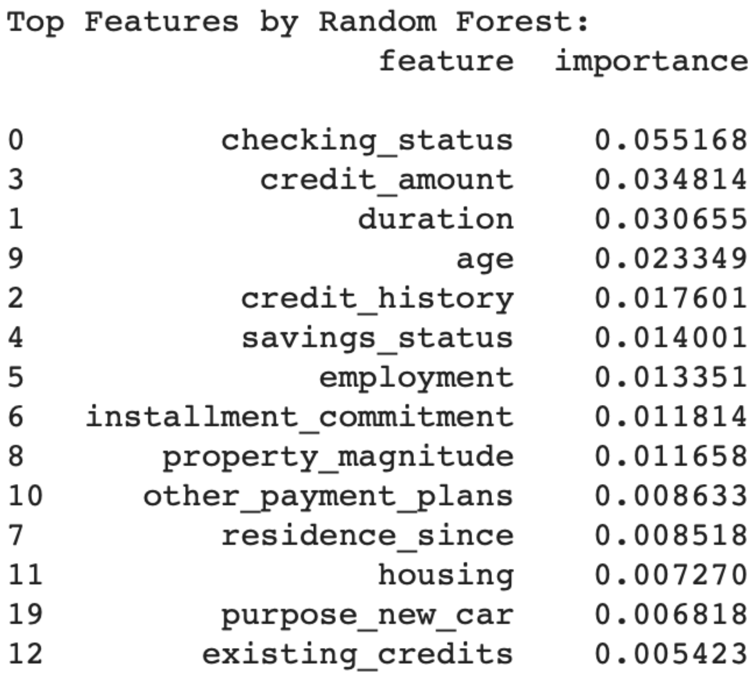
- Feature importance, partial dependence plots, SHAP values
- Key insights and patterns discovered

Pearsons Correlation for feature selection



Feature Importance using Random Forest





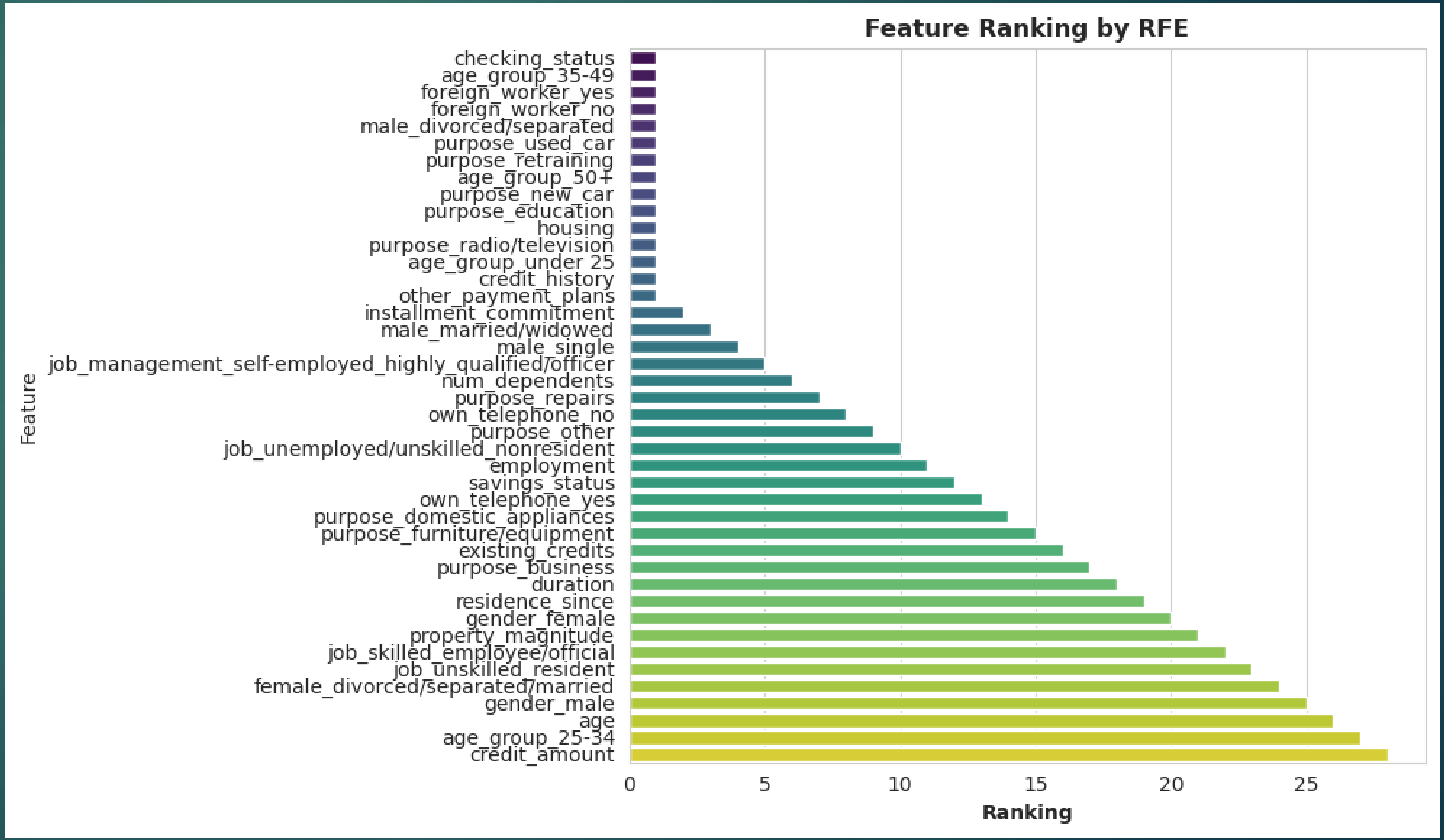
```
Top Features by Random Forest:
                  feature  importance

0         checking_status    0.055168
3           credit_amount    0.034814
1                duration    0.030655
9                     age    0.023349
2          credit_history    0.017601
4          savings_status    0.014001
5              employment    0.013351
6  installment_commitment    0.011814
8      property_magnitude    0.011658
10    other_payment_plans    0.008633
7         residence_since    0.008518
11                housing    0.007270
19        purpose_new_car    0.006818
12       existing_credits    0.005423
```

Selection by RFE (Recursive Feature Elimination)

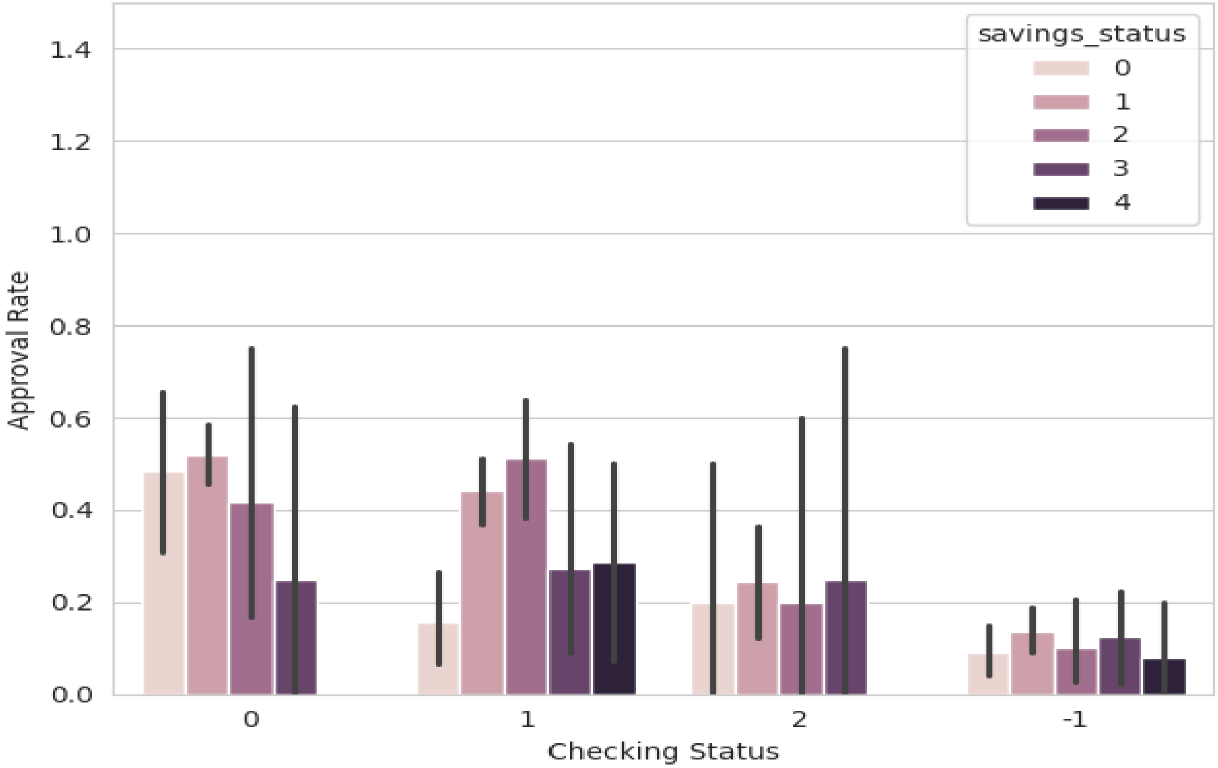
Selected Features by RFE:

1)'checking_status'
2) 'credit_history'
3) 'other_payment_plans',
4) 'housing'
5) 'purpose_education'
6) 'purpose_new_car'
7) 'purpose_radio/television'
8) 'purpose_retraining'
9) 'purpose_used_car'
10) 'male_divorced/separated'
11) 'foreign_worker_no'
12) 'foreign_worker_yes'
13) 'age_group_35-49'
14) 'age_group_50+'
15) 'age_group_under 25']'



1. Having a higher savings account balance (over 1000DM, represented by savings status = 3 in purple) generally leads to a higher approval rate across most checking status categories.
2. However, the approval rate for savings status = 3 is not uniformly high across all checking status categories. It is highest for checking status 2 (balance > 200DM) and 1 (balance 0-200DM), but noticeably lower for checking status 0 (balance < 0DM) and -1 (no checking account/missing data).
3. This suggests that while having a substantial savings balance is advantageous, it does not guarantee approval if the checking account status is poor (negative balance or no account).
4. The combination of a high savings balance (>1000DM) and a reasonably positive checking account balance (>0DM) likely yields the highest approval rates based on this data.

- So, in summary, having over 1000DM in savings does increase the approval rate substantially, but the checking account status still plays an important role. A very high savings balance alone may not overcome the negative impact of a poor or non-existent checking account. Both savings and checking account balances seem to be considered for the approval decision based on the patterns in this chart.



The checking status categories on the x-axis are:
0: indicates a checking account balance of less than 0 DM (Deutsche Mark).
1: indicates a checking account balance between 0 and 200 DM.
2: indicates a checking account balance greater than 200 DM.
-1: indicates no checking account or missing data.

The savings status categories in the legend are:
0: indicates a savings account balance of less than 100 DM.
1: indicates a savings account balance between 100 and 500 DM.
2: indicates a savings account balance between 500 and 1000 DM.
3: indicates a savings account balance greater than 1000 DM. 4: indicates no savings account or missing data.

Modeling Approach

Categorical column values are converted to Numerical values for modeling.

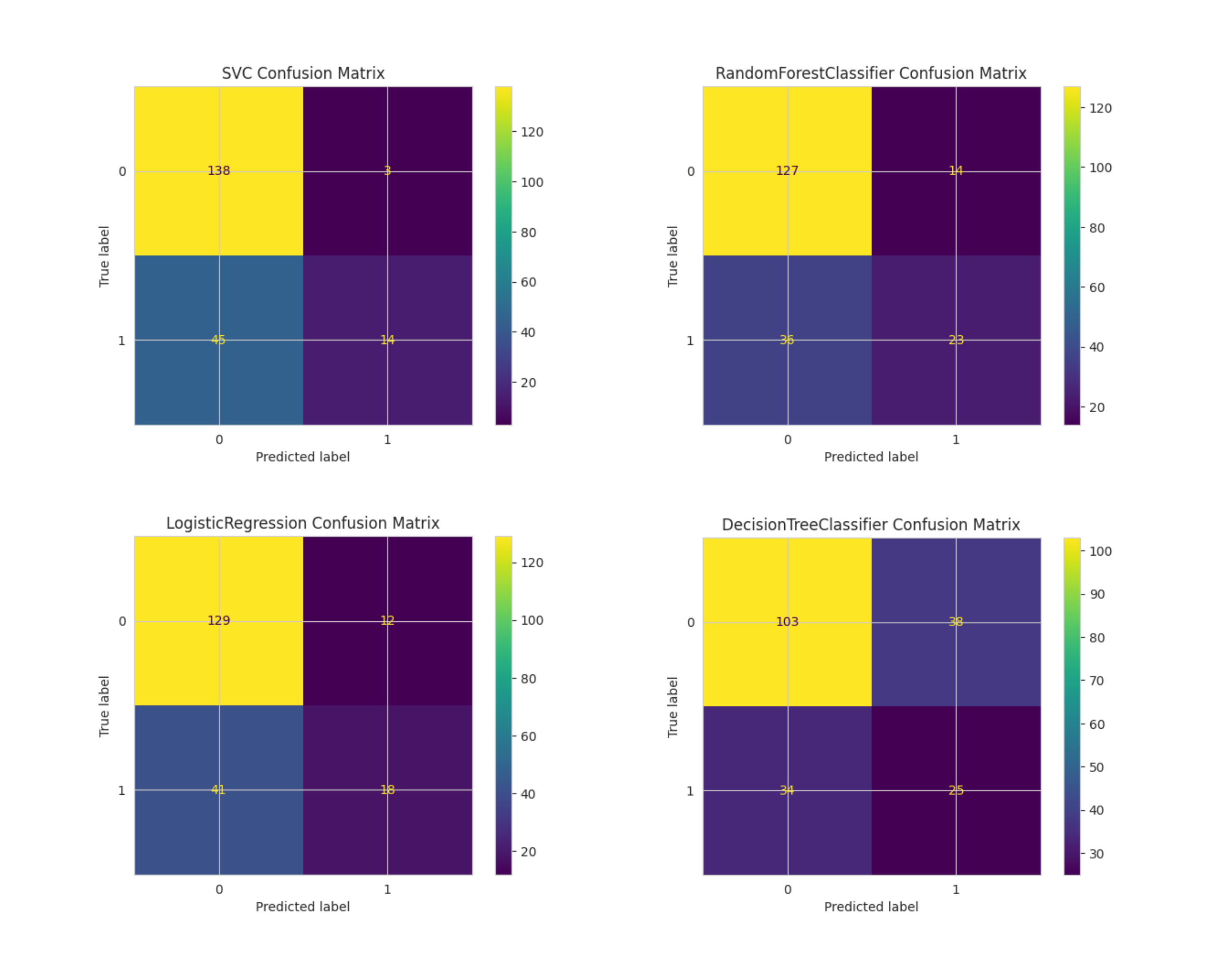
Columns that aren’t important are dropped to reduce complexity and increase computational time.

Supervised training algorithms like Logistic regression, Random Forest, Support Vector Machine, Decision tree are tested.

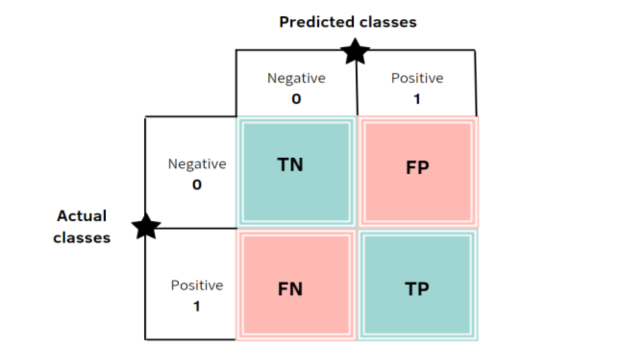
Finding the best model for training and testing.

Training process (cross-validation, hyperparameter tuning)

Model Evaluation & Performance



CONFUSION MATRIX



		Predicted classes	
		Negative 0	Positive 1
Actual classes	Negative 0	TN	FP
	Positive 1	FN	TP

Accuracy and Cross validation scores

- **Random Forest Classifier** performed the best in terms of both test accuracy (75%) and average cross-validation score (73.6%).

- **SVC** had the highest test accuracy (76%) but a slightly lower average cross-validation score (71.3%).

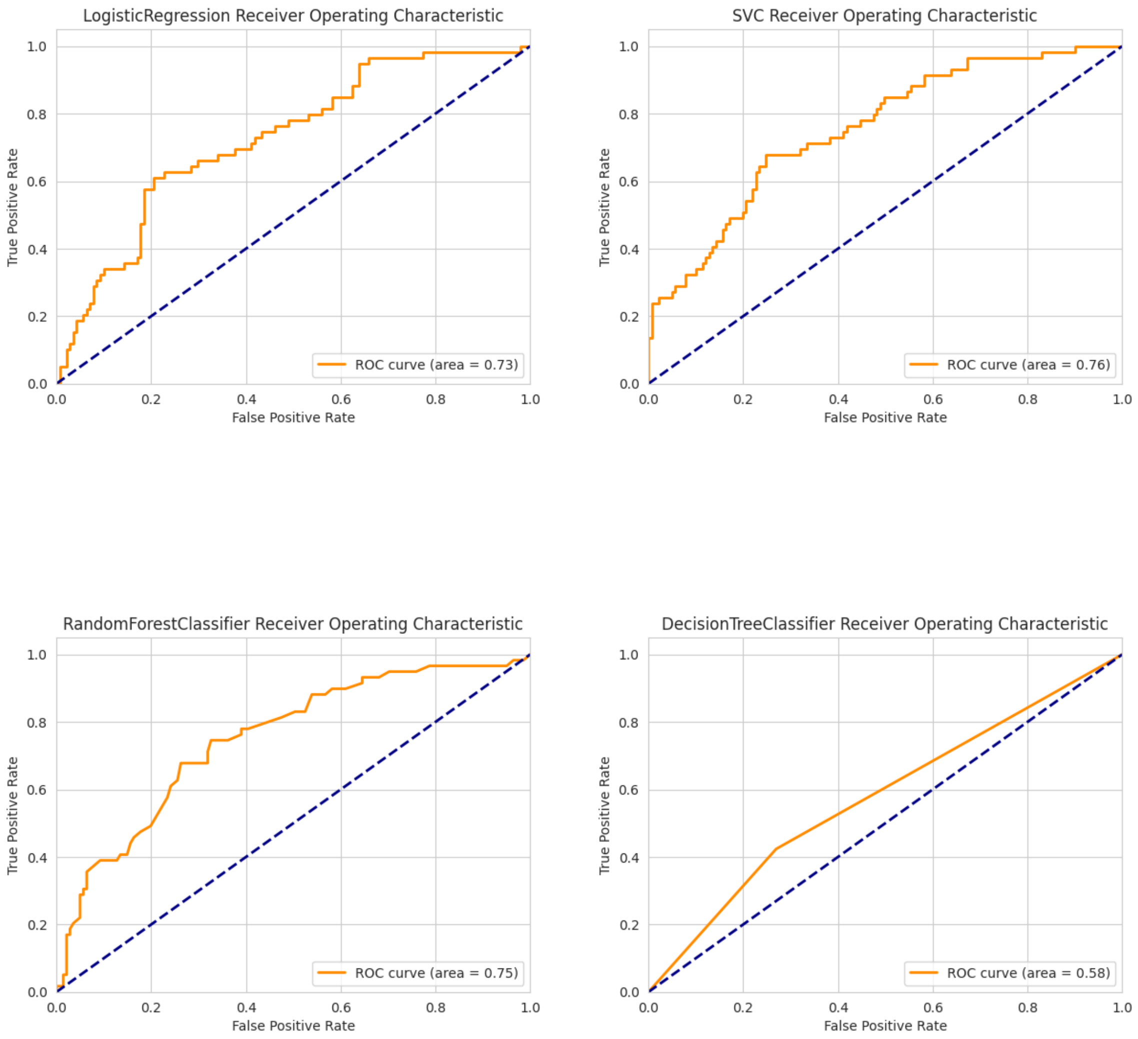
- **Logistic Regression** performed well with a test accuracy of 73.5% and an average cross-validation score of 71.4%.

- **Decision Tree Classifier** had the lowest performance with a test accuracy of 64% and an average cross-validation score of 68.2%.

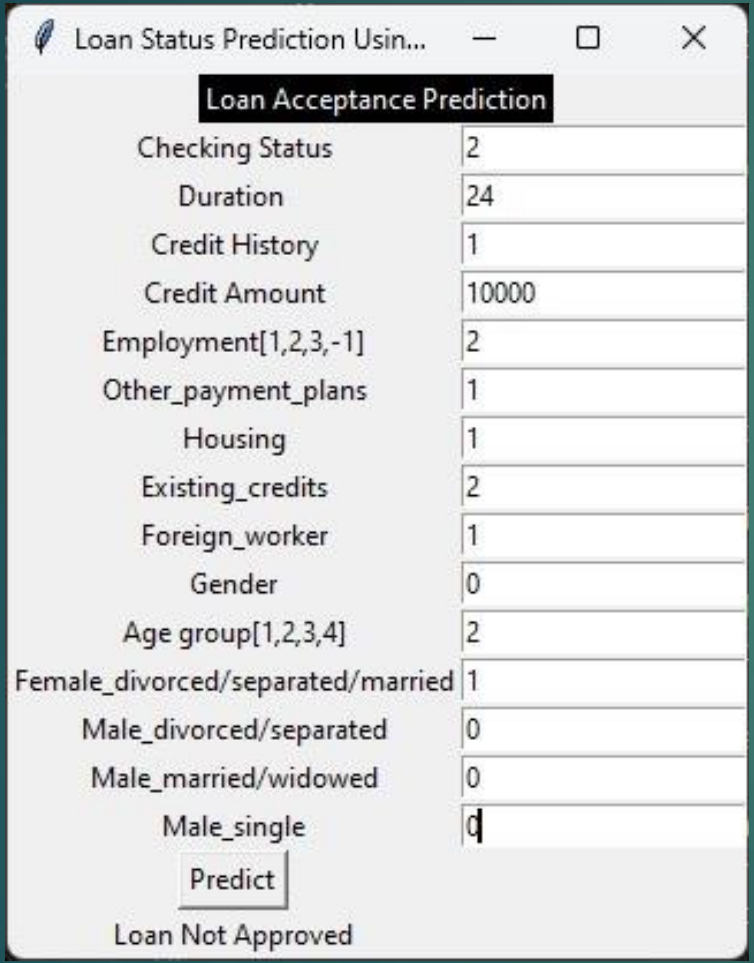
Algorithm	Accuracy	Cross-Val Score	Performance
Logistic regression	73.5%	71.4%	Good
Random Forest	75%	73.6%	Best
SVM	76%	71.3%	High
Decision Tree	64%	68.2%	Lowest

ROC Curve

- ROC curve for Decision Tree classifier suggests that the model is not effective in discriminating between classes as its very close to the diagonal line.
- SVC and RF classifiers on the other hand are close to the top-left corner of the plot indicating high sensitivity and low false positive rate across different thresholds.



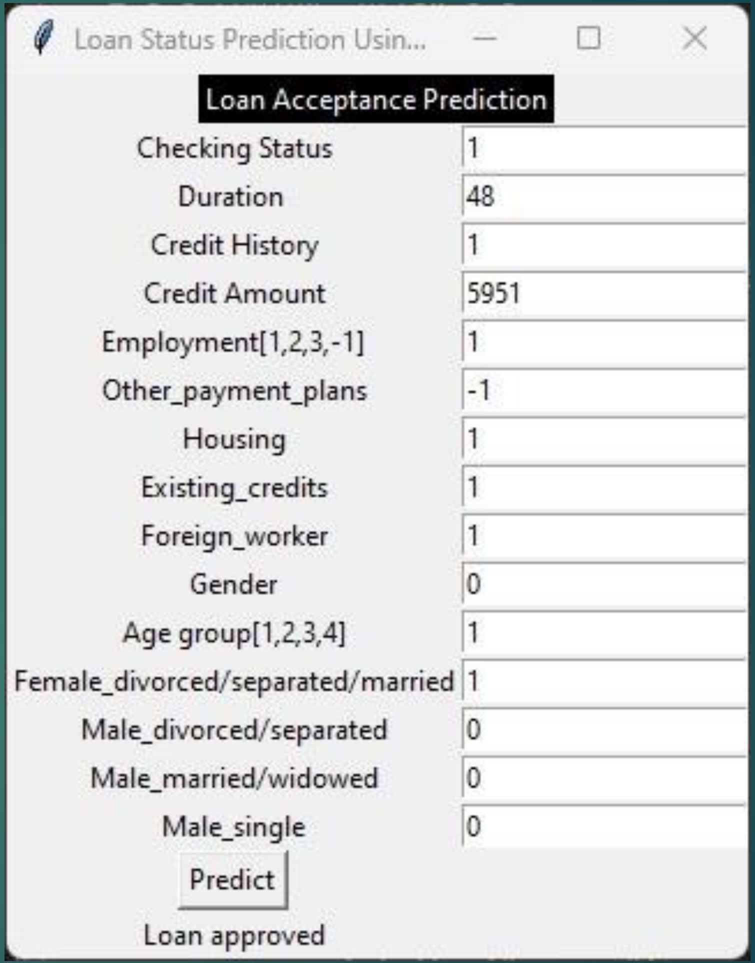
Output



Loan Acceptance Prediction

Checking Status	2
Duration	24
Credit History	1
Credit Amount	10000
Employment[1,2,3,-1]	2
Other_payment_plans	1
Housing	1
Existing_credits	2
Foreign_worker	1
Gender	0
Age group[1,2,3,4]	2
Female_divorced/separated/married	1
Male_divorced/separated	0
Male_married/widowed	0
Male_single	0

Loan Not Approved



Loan Acceptance Prediction

Checking Status	1
Duration	48
Credit History	1
Credit Amount	5951
Employment[1,2,3,-1]	1
Other_payment_plans	-1
Housing	1
Existing_credits	1
Foreign_worker	1
Gender	0
Age group[1,2,3,4]	1
Female_divorced/separated/married	1
Male_divorced/separated	0
Male_married/widowed	0
Male_single	0

Loan approved

Loan Not Approved

Loan Approved

Future Work and Limitations

LIMITATIONS AND
ASSUMPTIONS

POTENTIAL IMPROVEMENTS
AND EXTENSIONS

Thank You