Loan Approval Prediction

This is a Binary Classification problem. We must correctly classify whether a loan will be approved or denied based on the given features.

Processed Data

Modified Data had a shape of (514,17) along with:

• Number of Integer-Categorical Columns	= 7
• Number of String-Categorical Columns	= 6
• Number of String-Boolean Columns	= 1
• Number of Numeric-Boolean Columns	= 1
• Number of ID Columns	= 1

The Features are grouped in Numerical and Categorical variable:

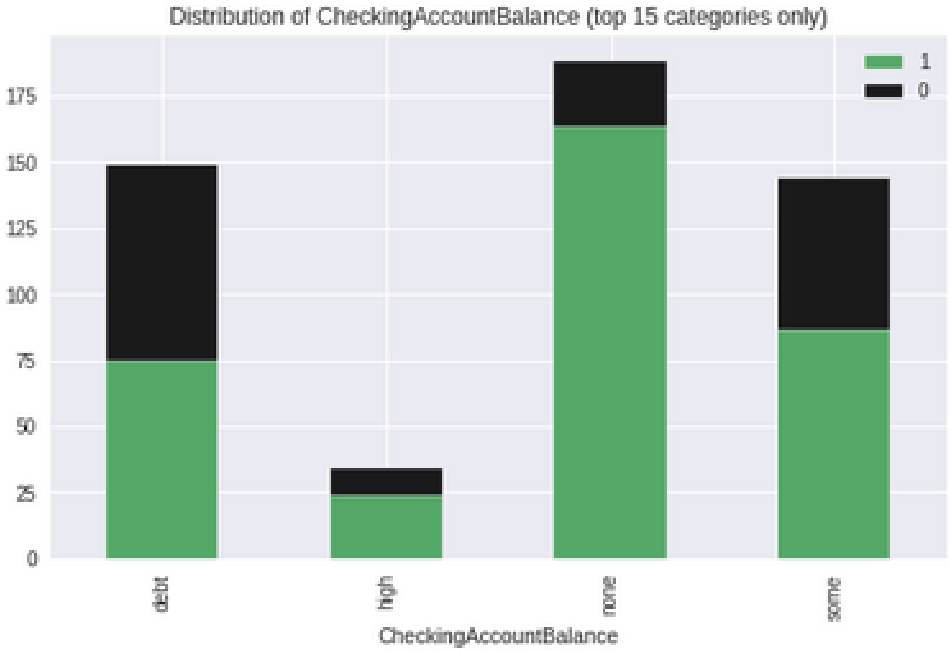
Numerical Features	• LoanPayoffPeriodInMonths • RequestedAmount • InterestRate • YearsAtCurrentEmployer • YearsInCurrentResidence • Age • NumberOfDependantsIncludingSelf • CurrentOpenLoanApplications
Categorical Features	• LoanReason • Co-Applicant • RentOrOwnHome • TypeOfCurrentEmployment • CheckingAccountBalance • DebtsPaid • SavingsAccountBalance
Target	• WasTheLoanApproved

The given data set was imbalanced with ‘Loan Approved’ being the majority class

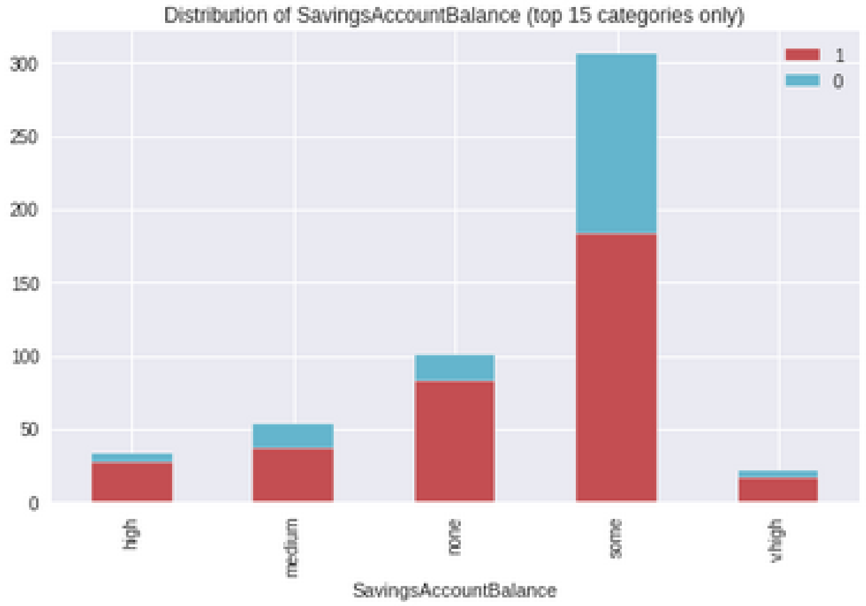
Univariate and Bivariate Analysis

Variable	**Type**	**Comments**
WasTheLoanApproved	Dependent variable	67% of people had their loans approved
RequestedAmount	Independent variable	Mean value of 4000, values lie in the range of (1024-18400)
InterestRate	Independent variable	Mode value of 2, values lie in the range of range (0-4)
Age	Independent variable	Mean value of 36 and values lie in the range of (19 to 75)

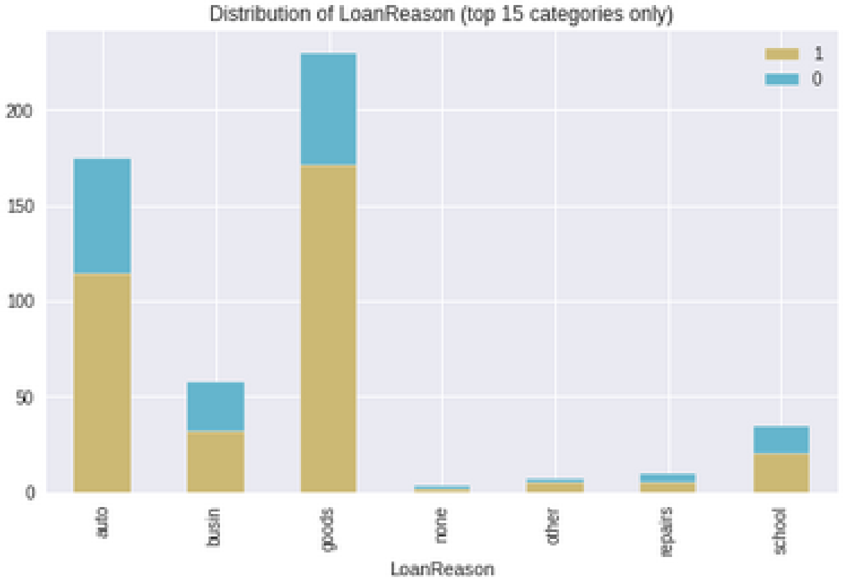
Analysis of 'CheckingAccountBalance' with Target Variable: 'WasTheLoanApproved'



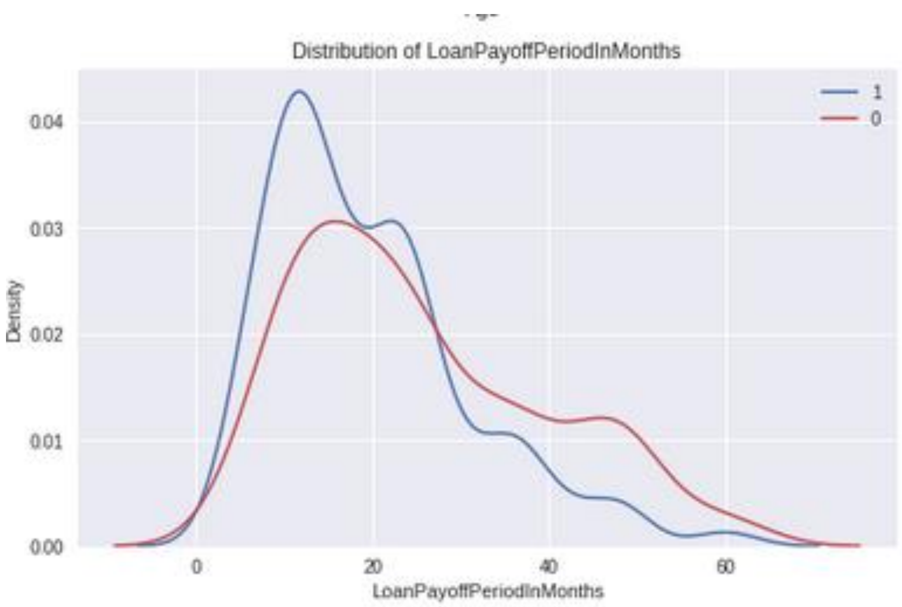
Analysis of 'SavingsAccountBalance' with Target Variable: 'WasTheLoanApproved'



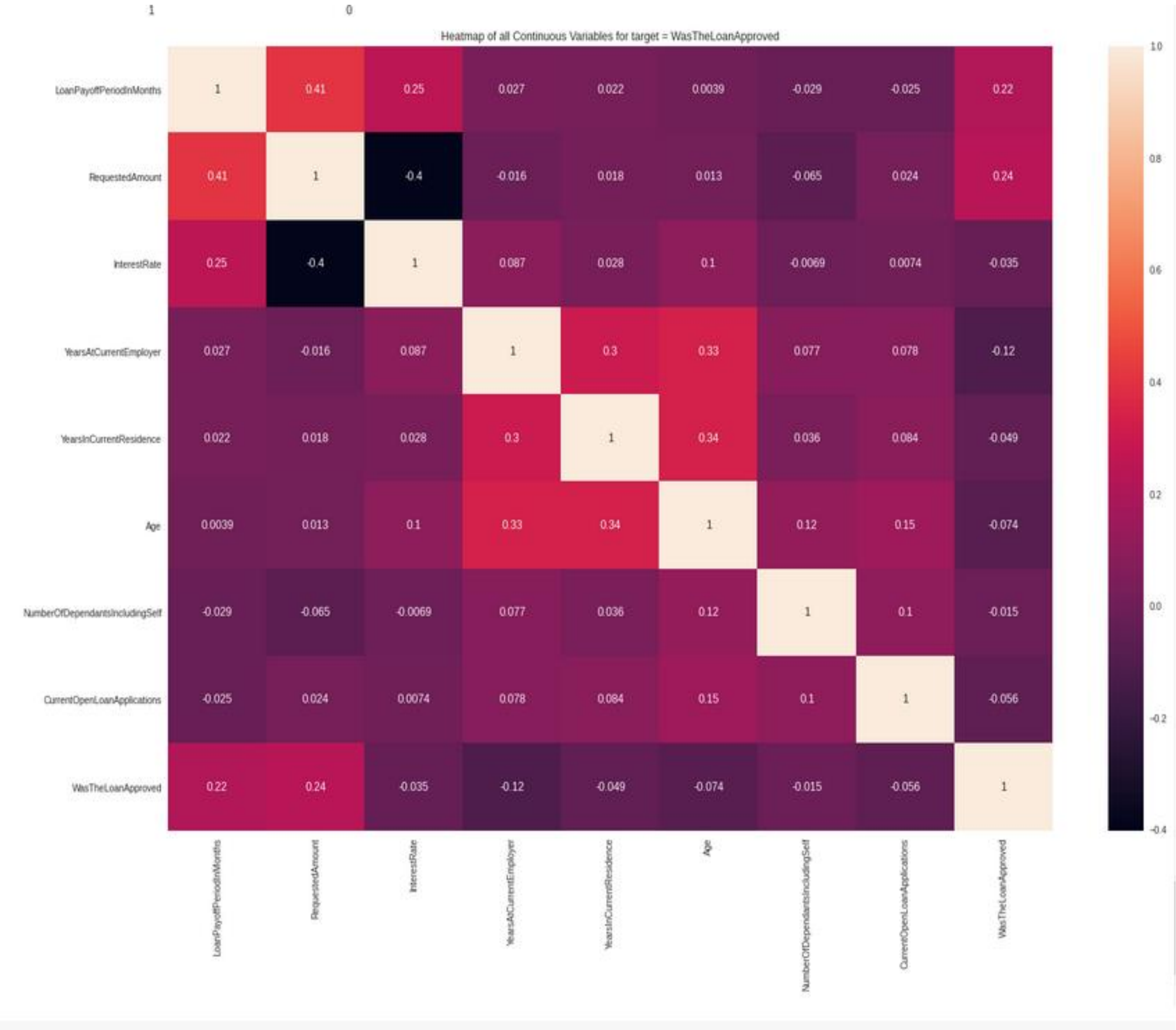
Analysis of 'LoanReason' with Target Variable : ‘WasTheLoanApproved’



Analysis of 'LoanPayoffPeriodInMonths' with Target Variable : ‘WasTheLoanApproved’



Correaltion Matrix between different features



	LoanPayoffPeriodInMonths	RequestedAmount	InterestRate	YearsAtCurrentEmployer	YearsInCurrentResidence	Age	NumberOfDependantsIncludingSelf	CurrentOpenLoanApplications	WasTheLoanApproved
LoanPayoffPeriodInMonths	1	0.41	0.25	0.027	0.022	0.0039	-0.029	-0.025	0.22
RequestedAmount	0.41	1	-0.4	-0.016	0.018	0.013	-0.065	0.024	0.24
InterestRate	0.25	-0.4	1	0.087	0.028	0.1	-0.0069	0.0074	-0.035
YearsAtCurrentEmployer	0.027	-0.016	0.087	1	0.3	0.33	0.077	0.078	-0.12
YearsInCurrentResidence	0.022	0.018	0.028	0.3	1	0.34	0.036	0.084	-0.049
Age	0.0039	0.013	0.1	0.33	0.34	1	0.12	0.15	-0.074
NumberOfDependantsIncludingSelf	-0.029	-0.065	-0.0069	0.077	0.036	0.12	1	0.1	-0.015
CurrentOpenLoanApplications	-0.025	0.024	0.0074	0.078	0.084	0.15	0.1	1	-0.056
WasTheLoanApproved	0.22	0.24	-0.035	-0.12	-0.049	-0.074	-0.015	-0.056	1

From the Correaltion Matrix we can infer that:

- Loan pay off periods in months is correlated with Requested Amount
- Loan pay off periods in months is correlated with Interstate
- Years at Employer is correlated with Years in Current residence
- Years at Employer is correlated with Age
- Years in Current residence is correlated with Age

Model Building and Feature Engineering

From Feature selection algorithm the *important features* were:

- CheckingAccountBalance
- LoanPayoffPeriodInMonths
- RequestedAmount
- SavingsAccountBalance
- CurrentOpenLoanApplication
- Age
- YearsAtCurrentEmployer
- LoanReason

For model validation, I used accuracy, precision, and recall.

The best base line accuracy was for XgBoost which had the highest accuracy of close to 69%

After up-sampling and scaling, the ensemble voting model of "XGB","RF","DT","ADB","GB” showed an accuracy of 80%

Submission Files:

1) Data.csv
2) Model_experimentation.ipynb
3) Data_explore.ipynb
4) Preprocessing.ipynb

Assumptions:

I used inner join of the .tsv files to avoid data imputation problem. The data had 515 rows for training/testing.