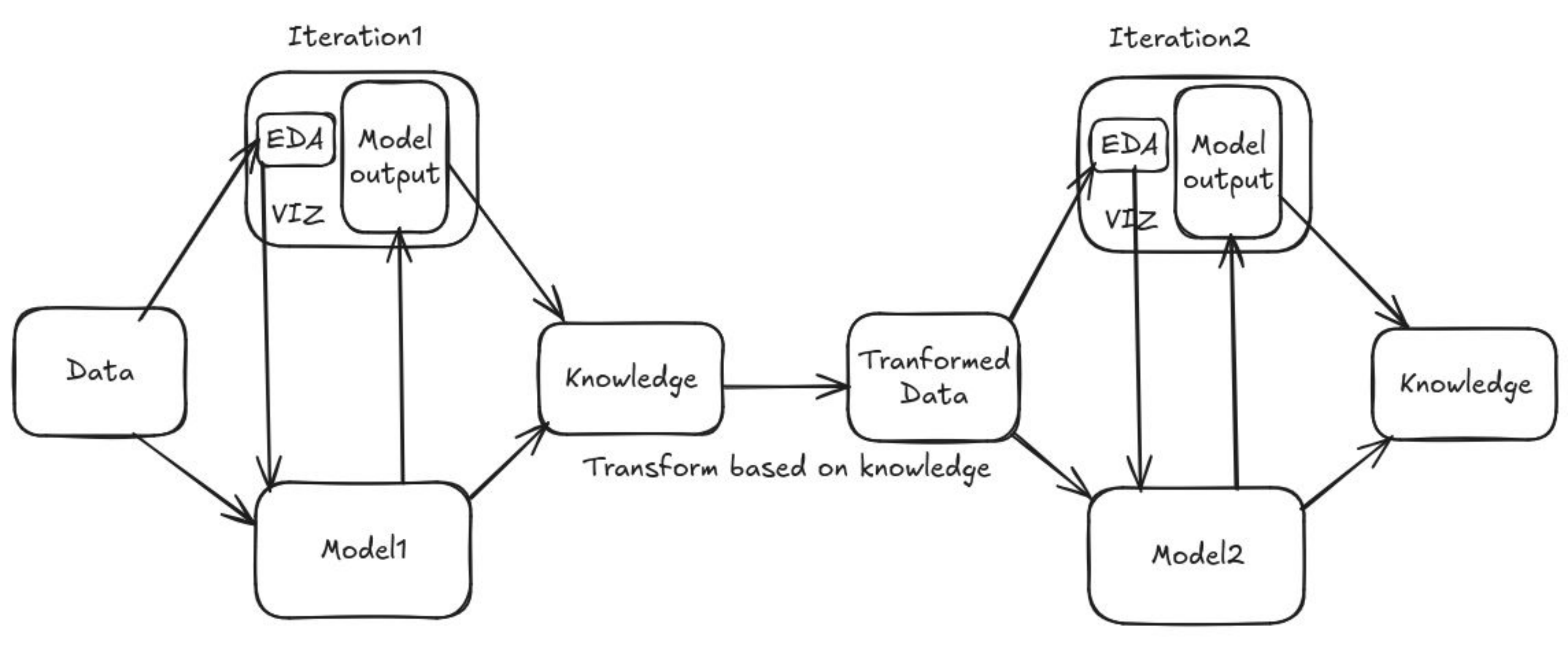
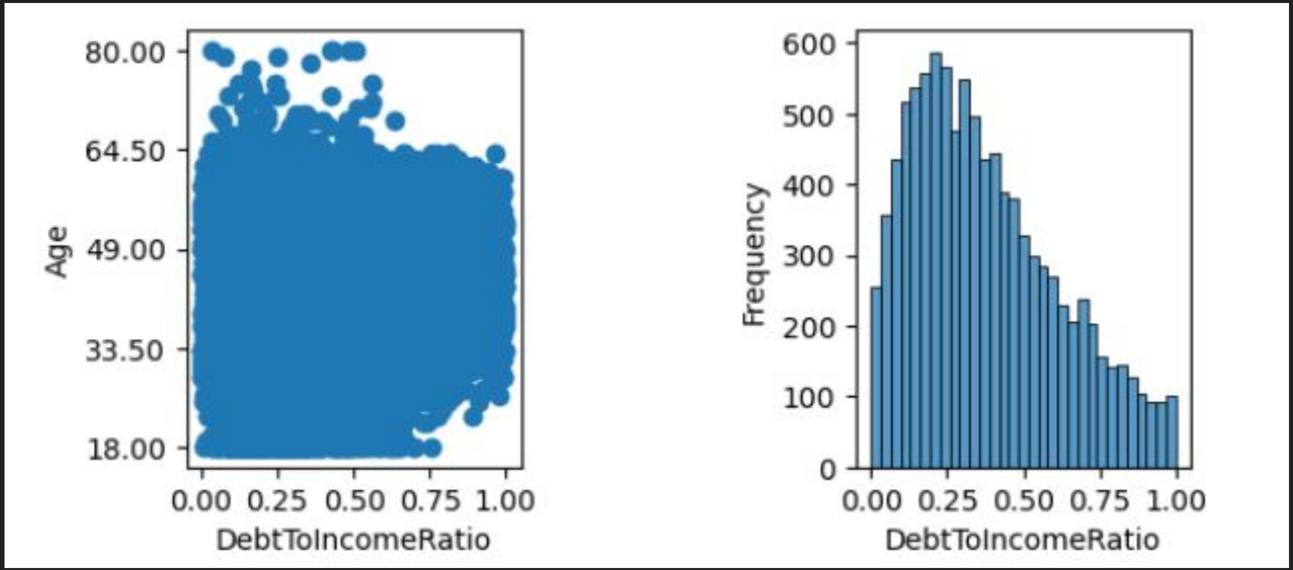
DV Assignment 3

Team Axis Powers

Workflow 1 - Prateek Rath IMT2022017

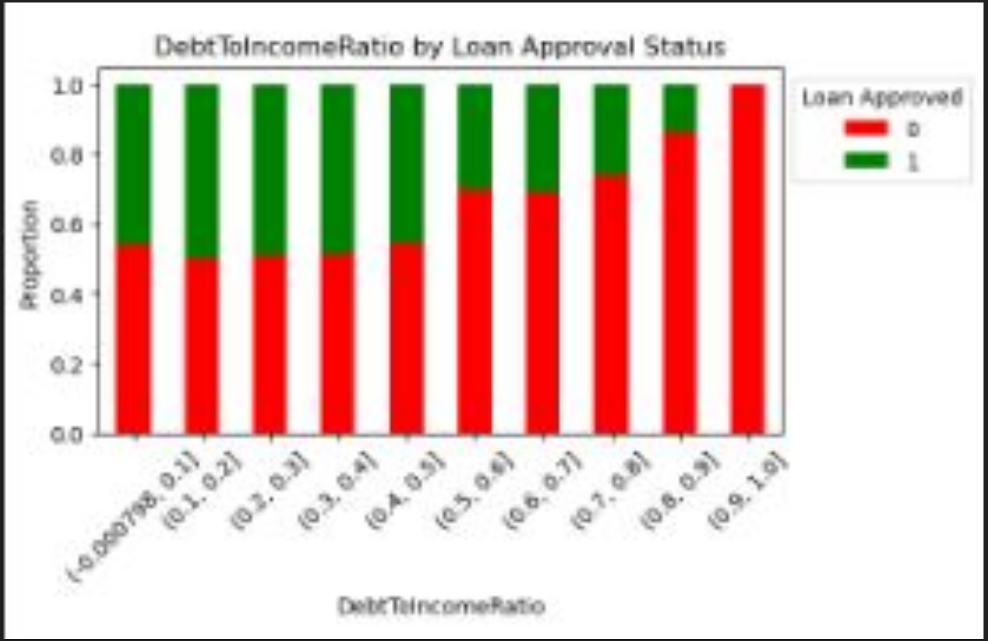


EDA1 - GPLOM



- Uncorrelated features
- Skewed Normal Distributions

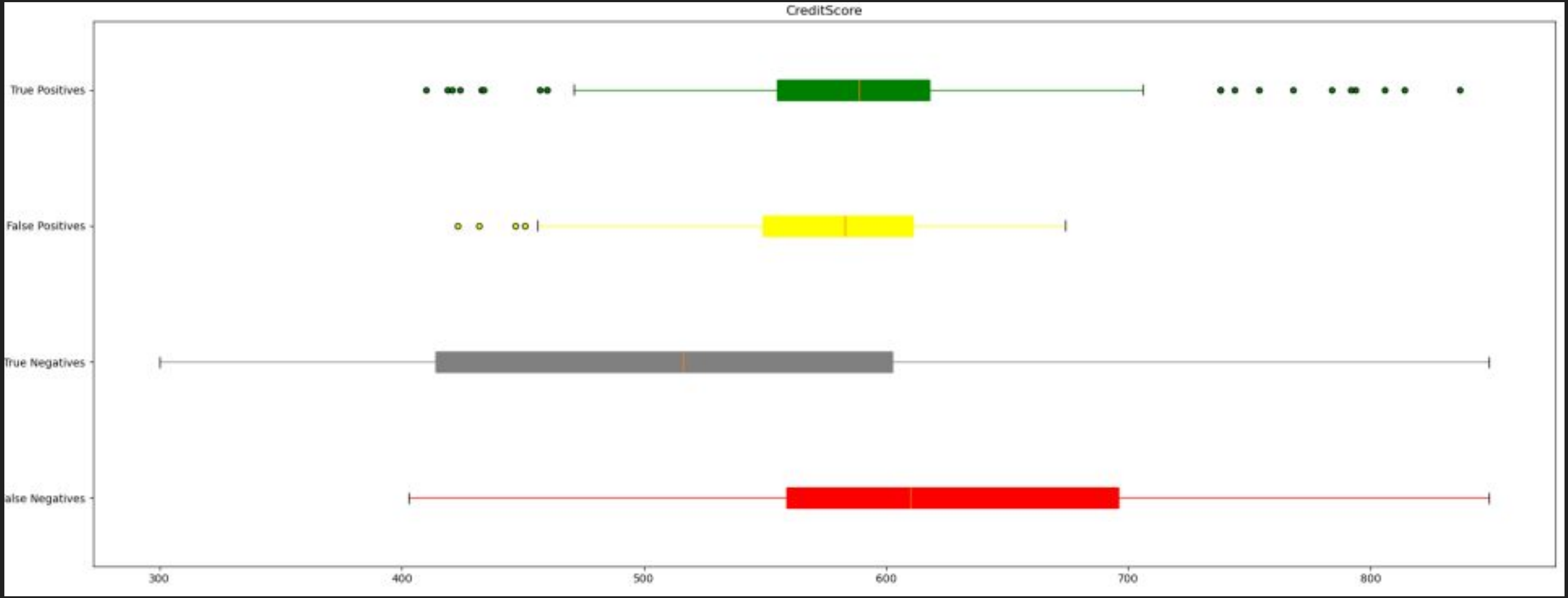
EDA1 - Proportion bar charts



- Higher DTI => No chance of loan approval

Model1 : Gaussian Naive Bayes; Diagnosis

Low Accuracy : 0.67

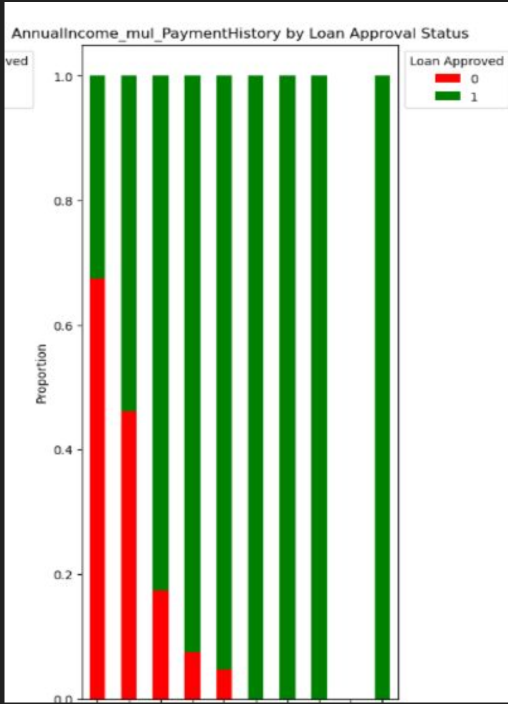


Data Transformations: Feature Engineering

- LoanAmountPerIncome
- LoanDurationToAge
- LoanAmountToCreditScore
- NetWorth: Assests - Liabitilites
- CredRatio: NumberOfCreditInquiries / LengthOfCreditHistory
- Familia: NumberOfDependents + (Marital Status == 'Married')
- AnnualIncome_mul_PaymentHistory
- CreditScore_and_DTI

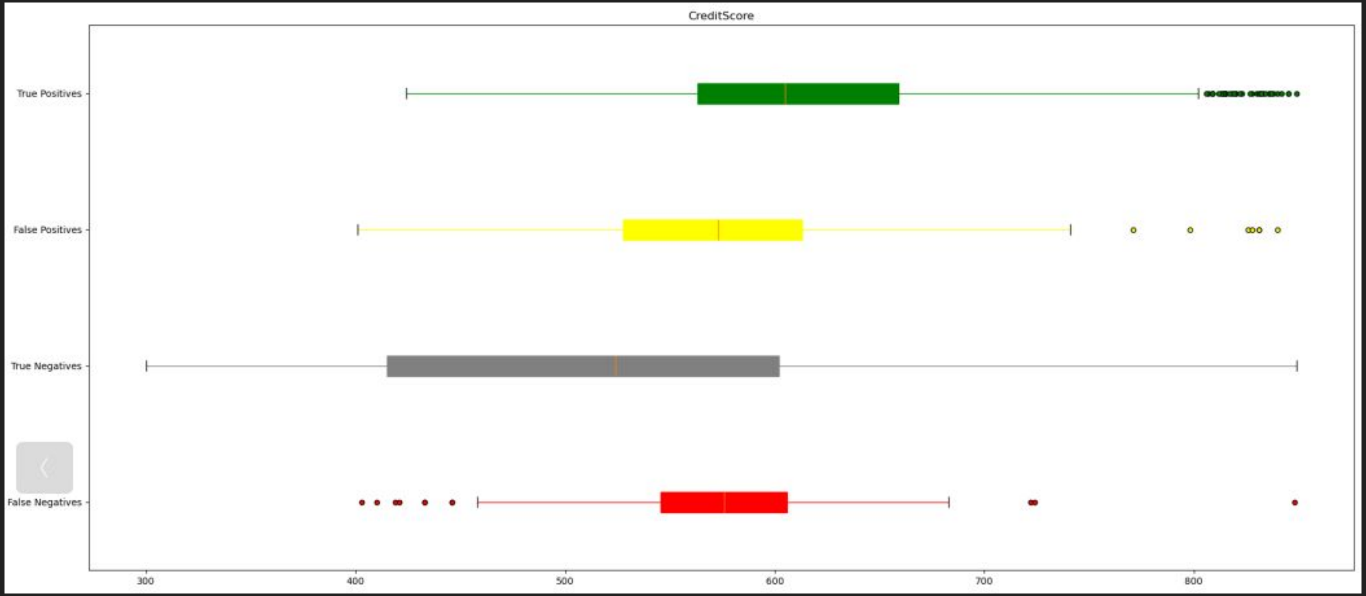
EDA2

- Similar, do proportion charts for new features



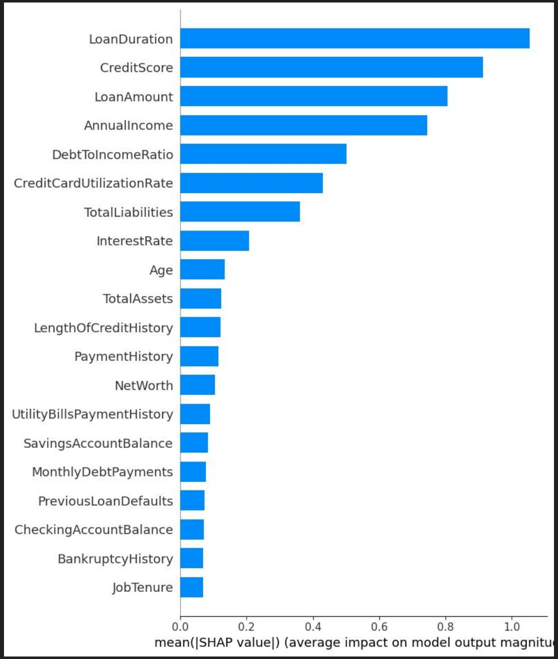
Model2: Xtreme Gradient Boost; Improvement

- Much better accuracy: 83 percent



Model2: Feature Importance

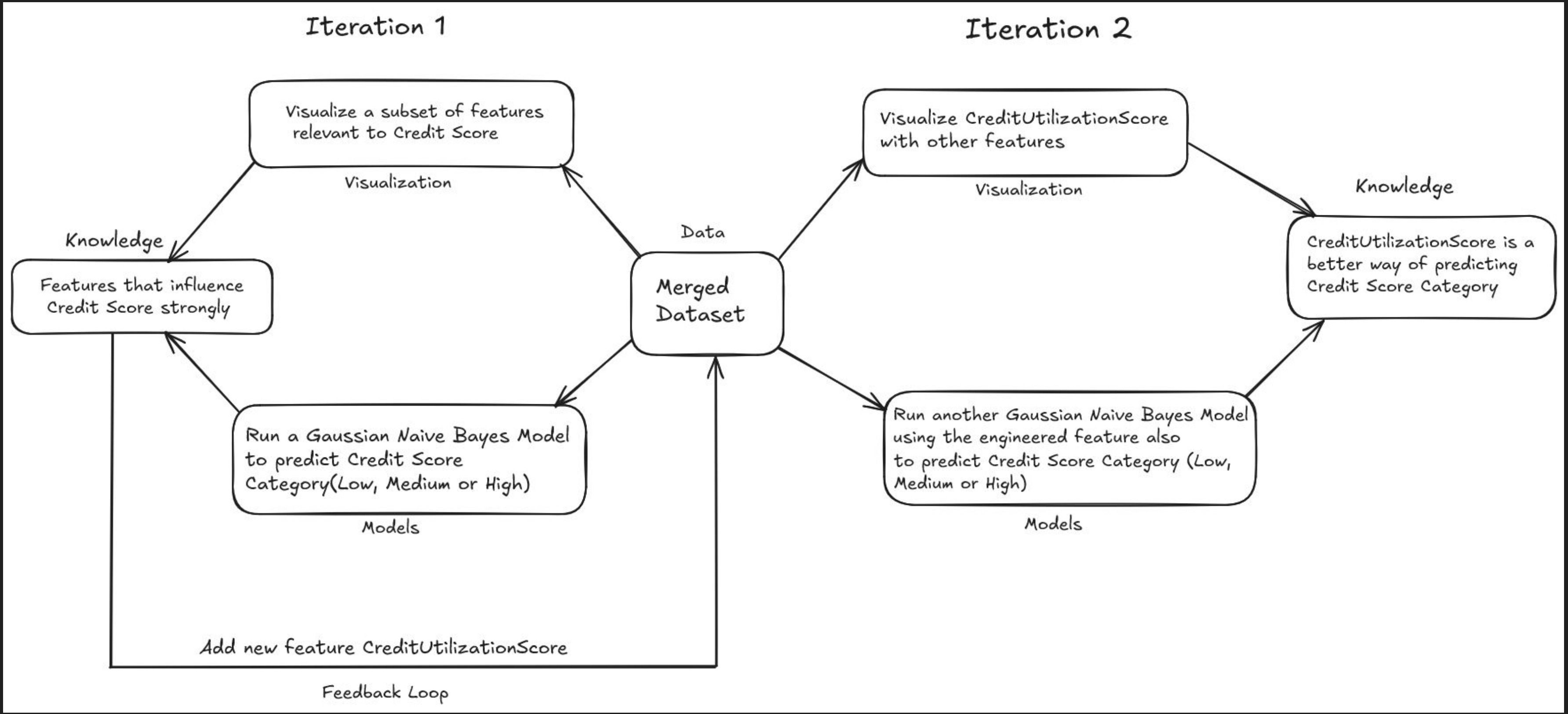
- Much better accuracy: 83 percent



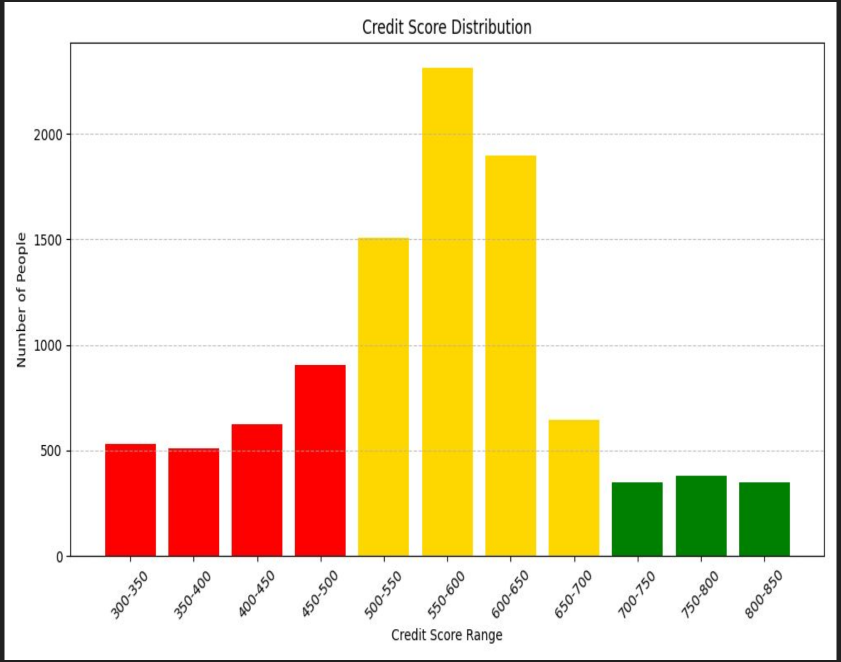
Conclusion and further iterations

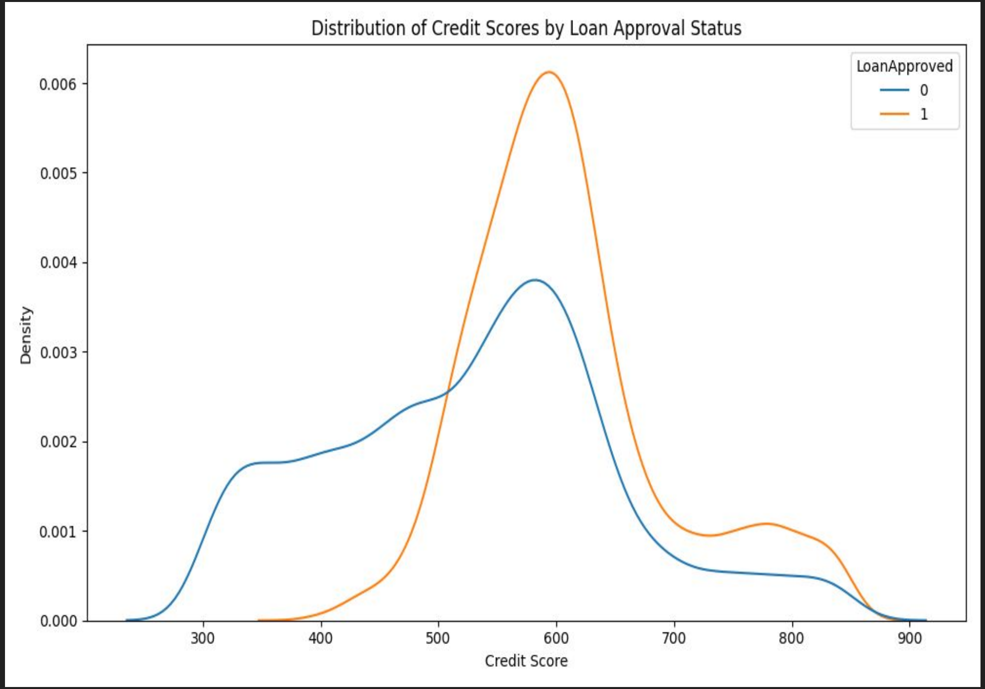
- Our model does pretty well but we don’t have to stop here
- Engineer new features
- Use business rules
- Post classification rules

Workflow 2 - Anurag Ramaswamy IMT2022103

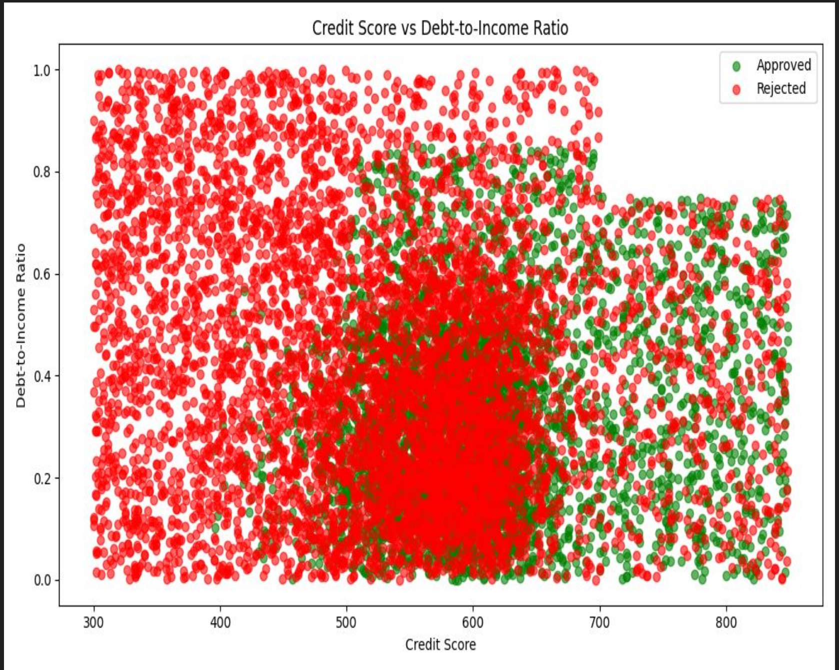


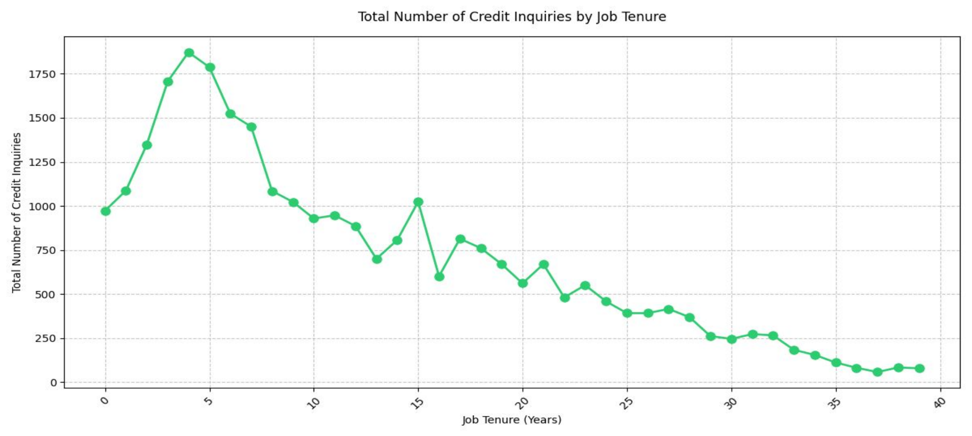
Iteration 1 visualizations

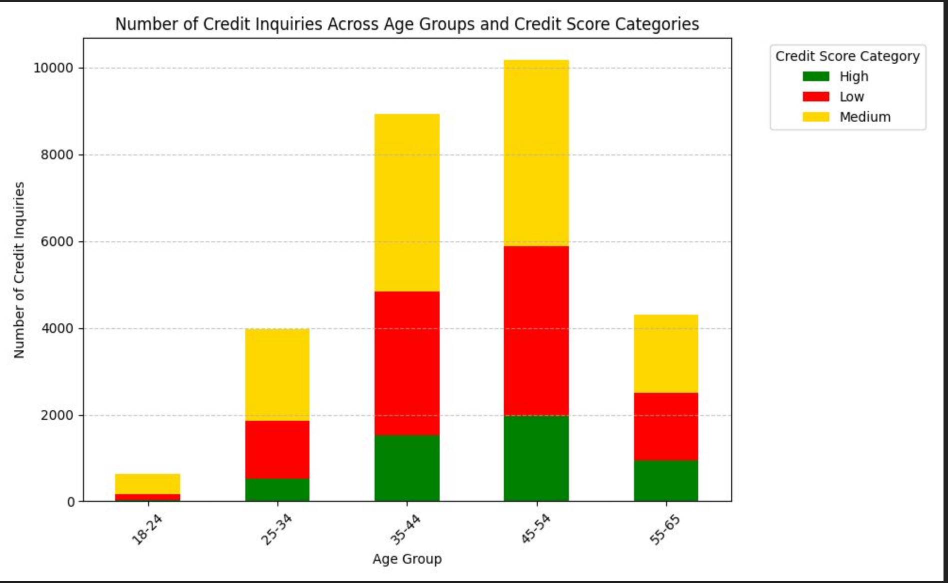




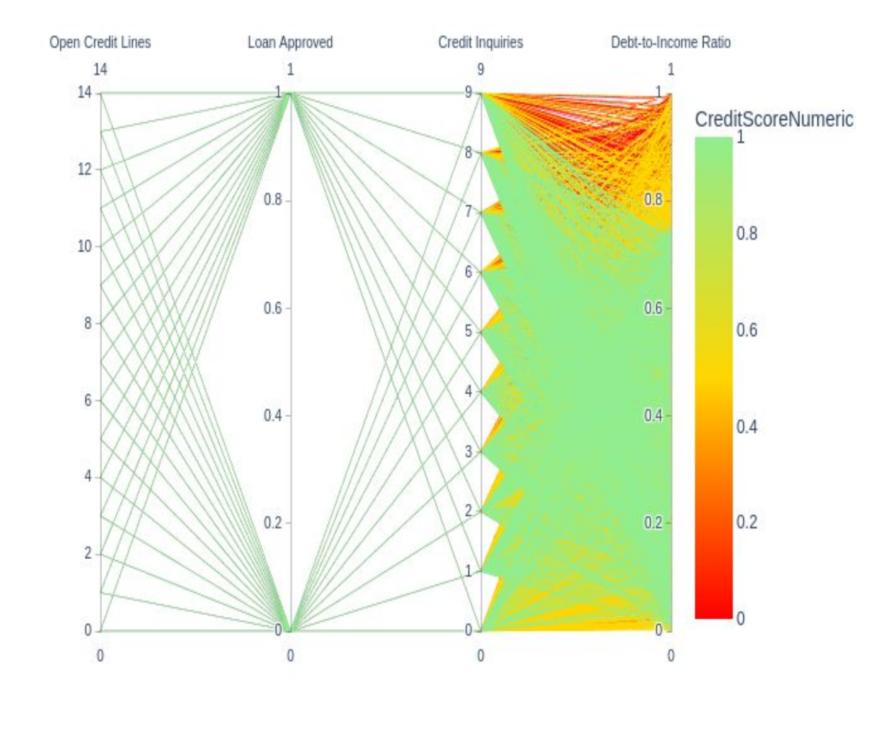
Iteration 1 visualizations

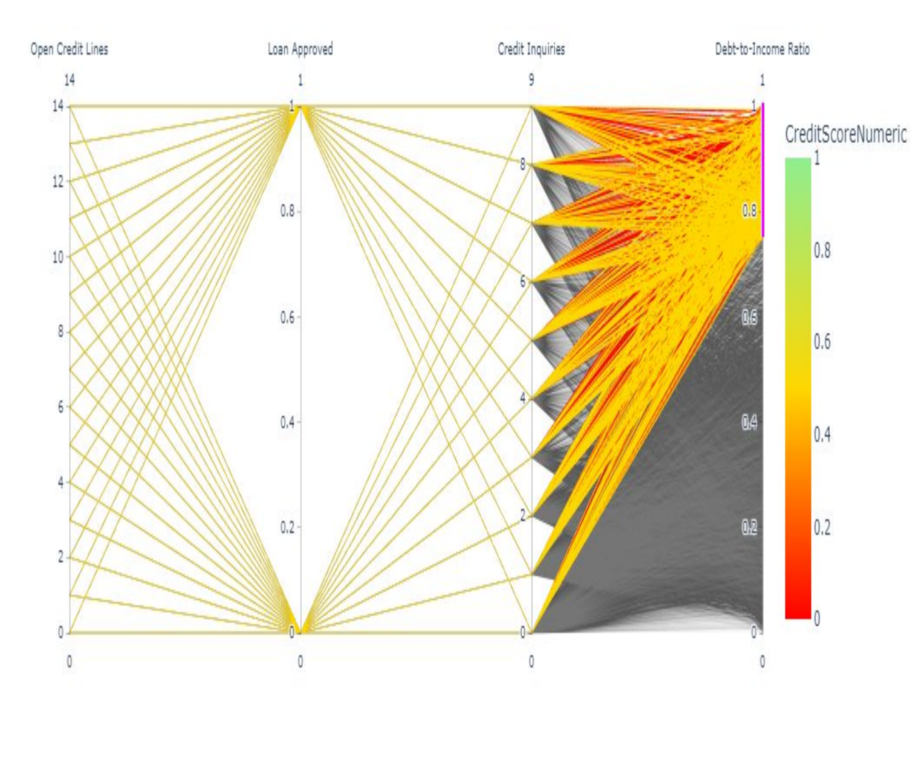






Iteration 1 visualizations





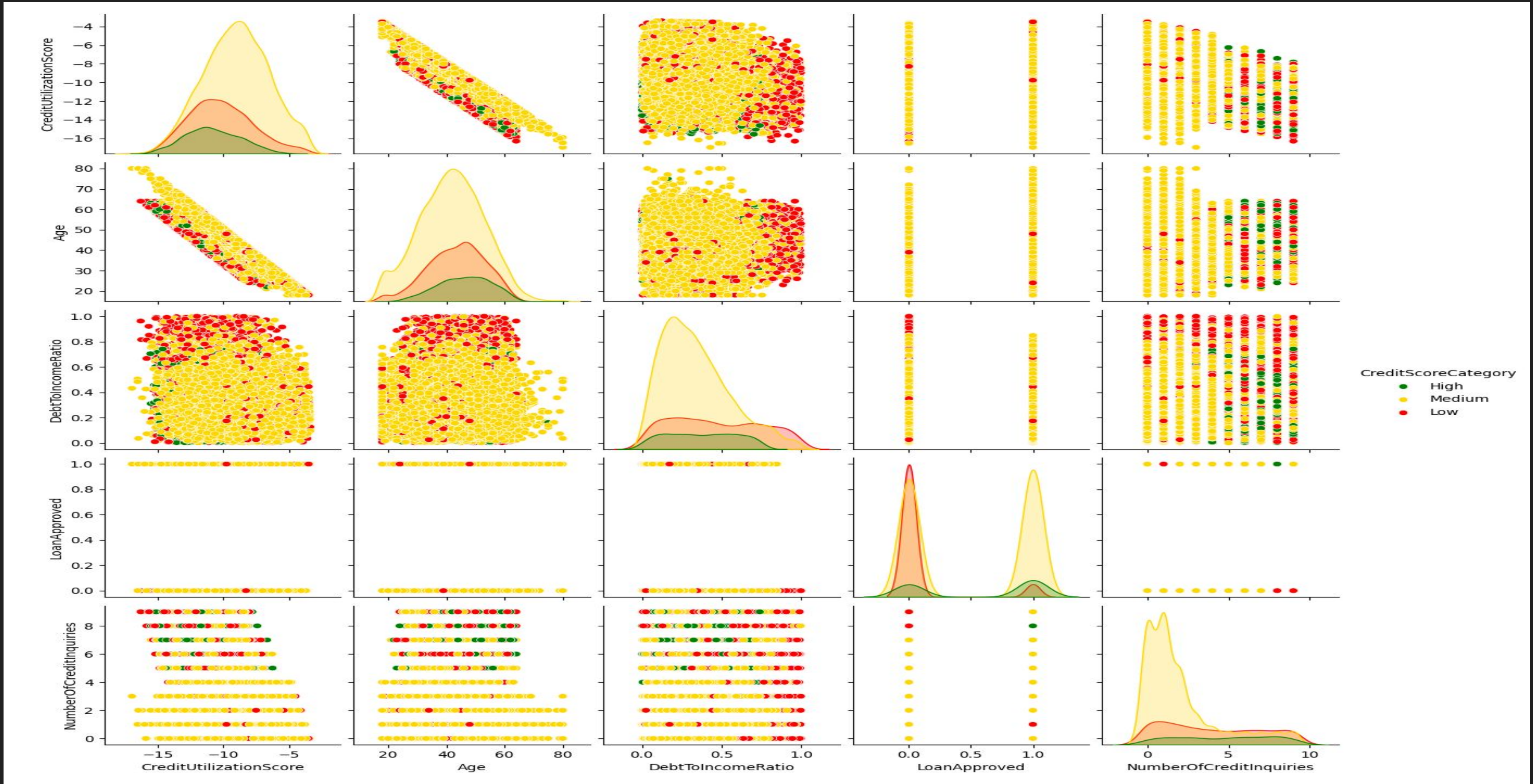
Create new feature CreditUtilizationScore

This feature is a weighted sum of:

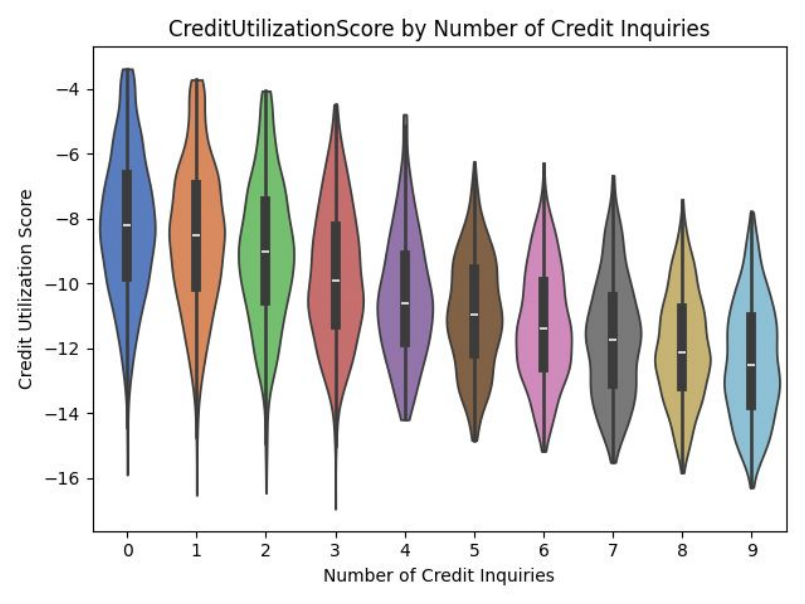
- Age
- NumberOfCreditInquiries
- DebtToIncomeRatio
- LoanApproved

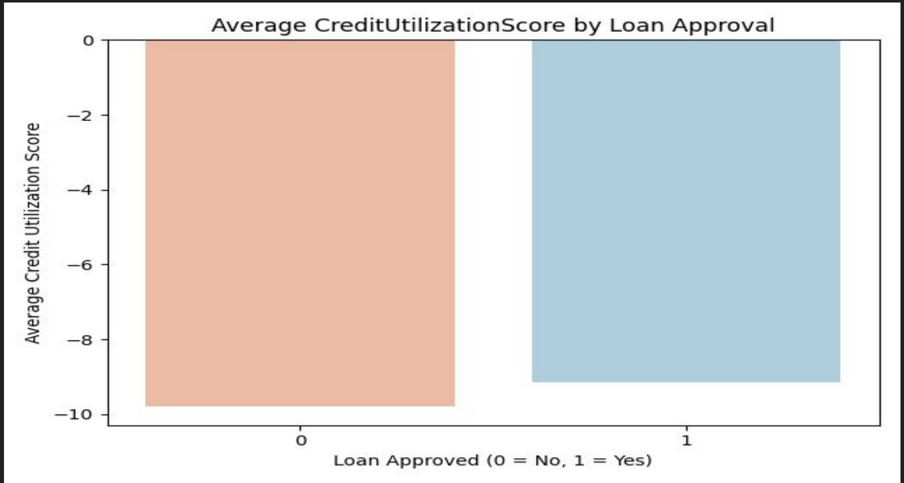
Why this new feature? - mainly to improve accuracy in predicting credit score category.

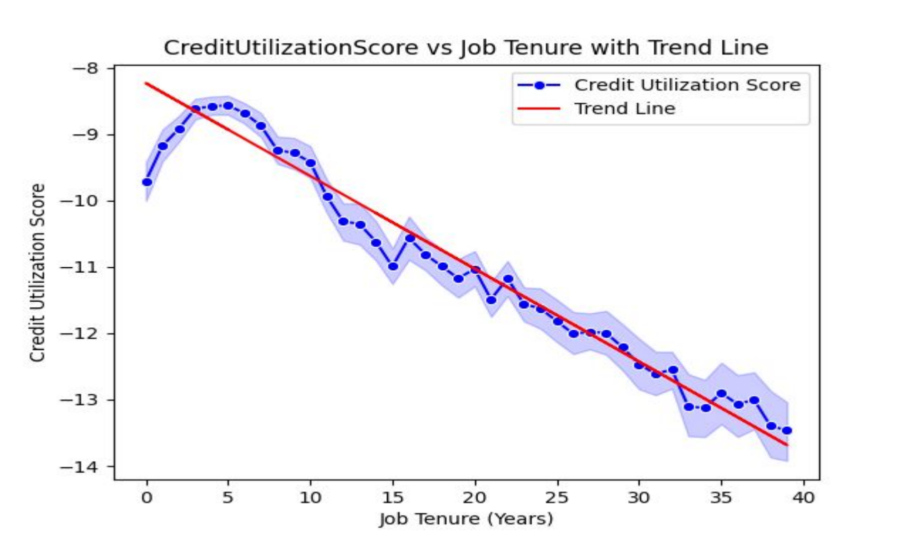
Iteration 2 visualizations



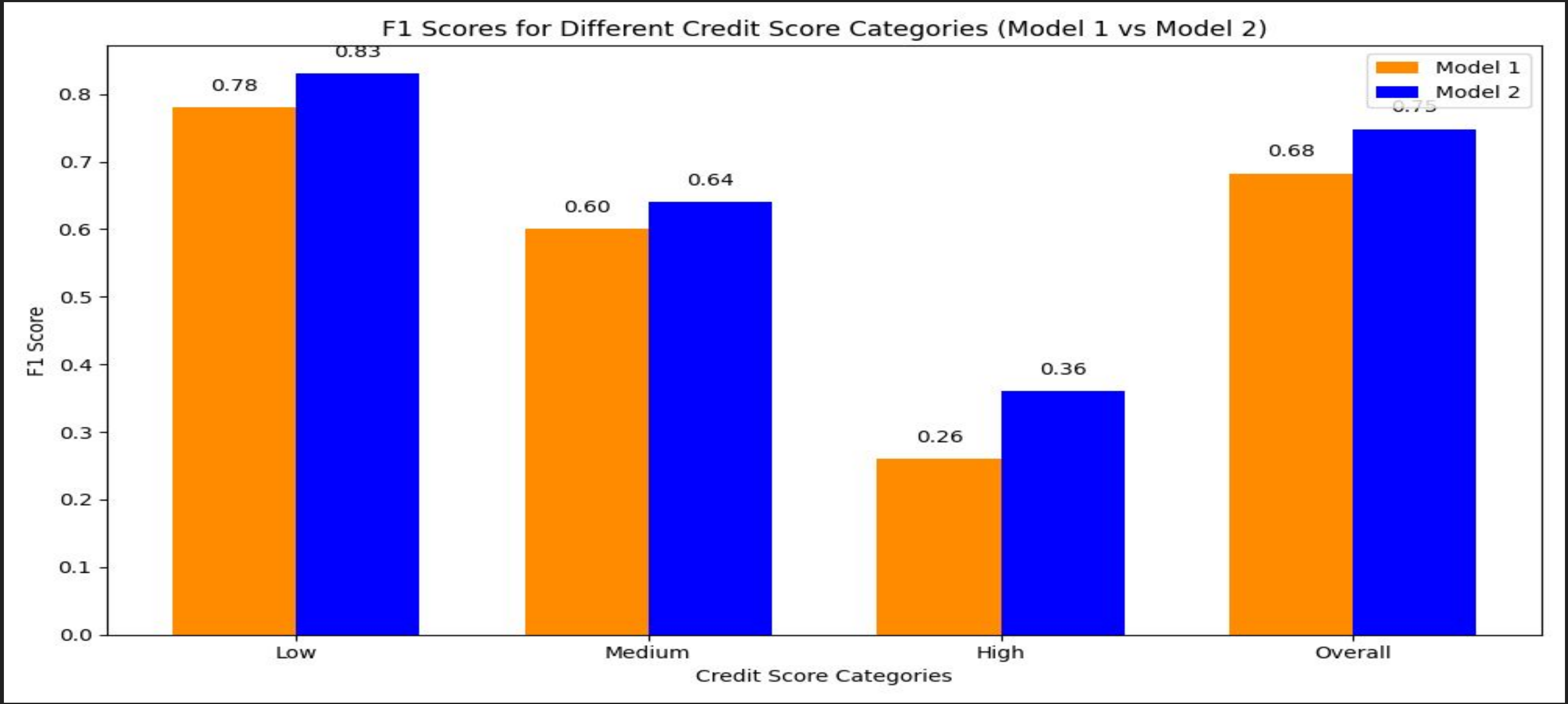
Iteration 2 visualizations







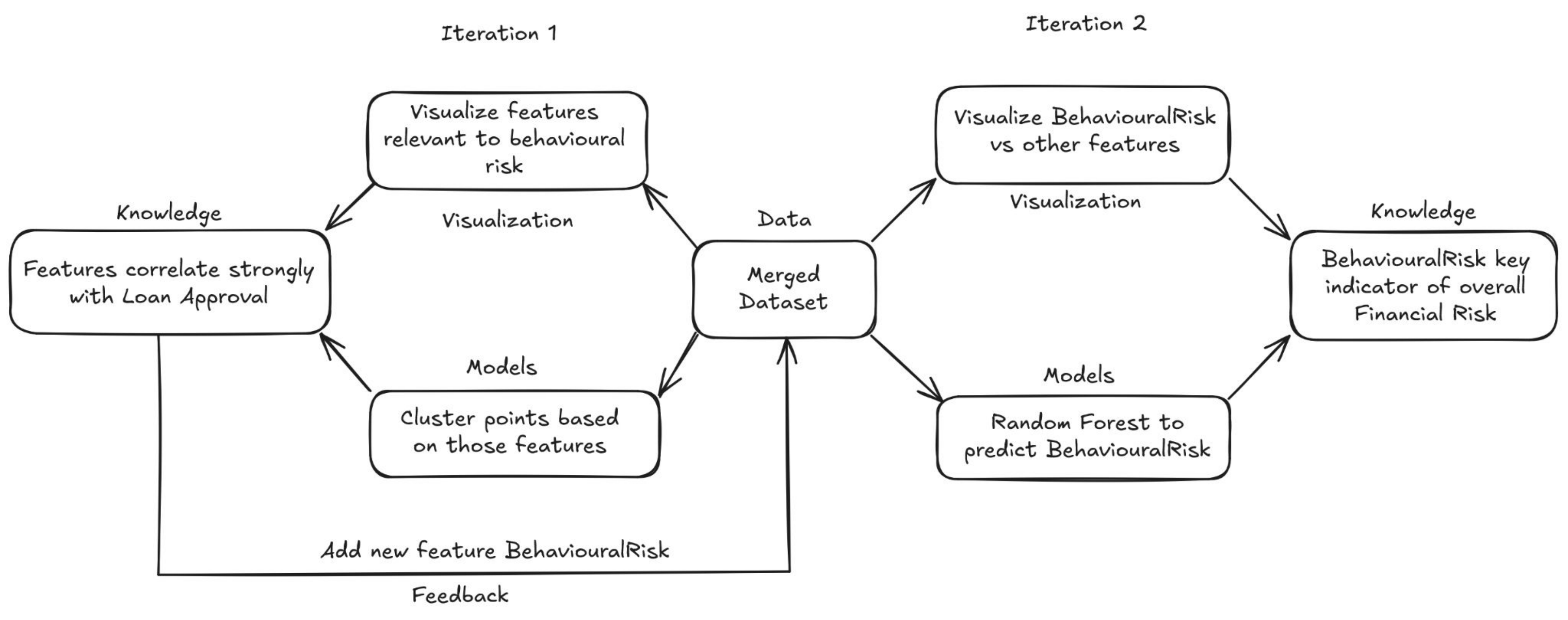
Accuracy comparison

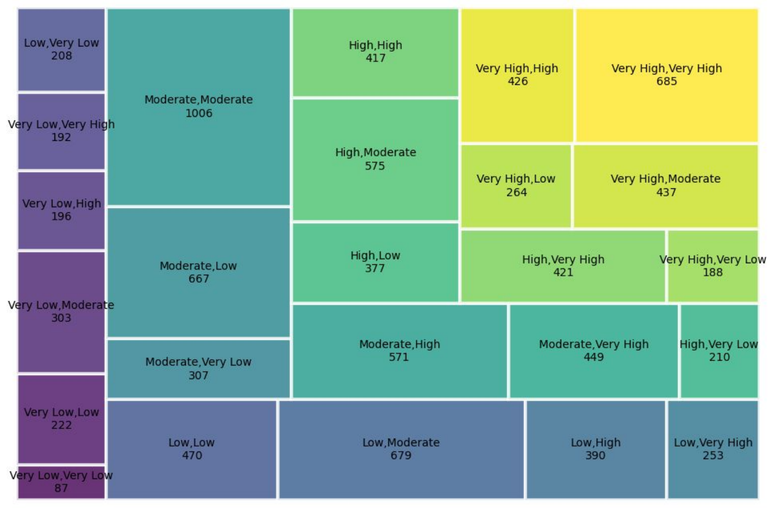


Conclusions

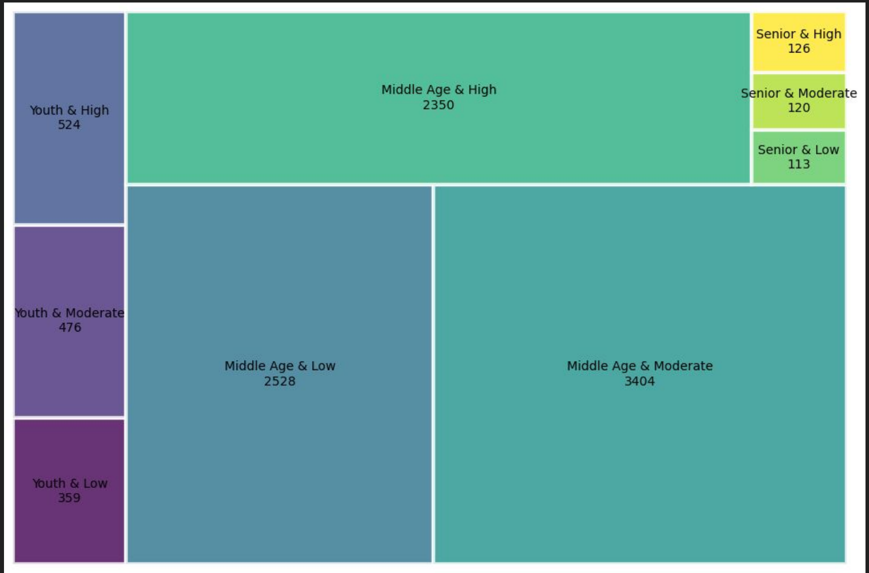
- There were trends in the data based on credit score and related factors
- These patterns were visualized but accuracy of prediction of credit score category was not high.
- New feature CreditUtilizationScore helped in improving accuracy
- Future steps to increase accuracy further may include introducing post-classification rules or by repeating the workflow for more iterations by using some other features

Workflow 3

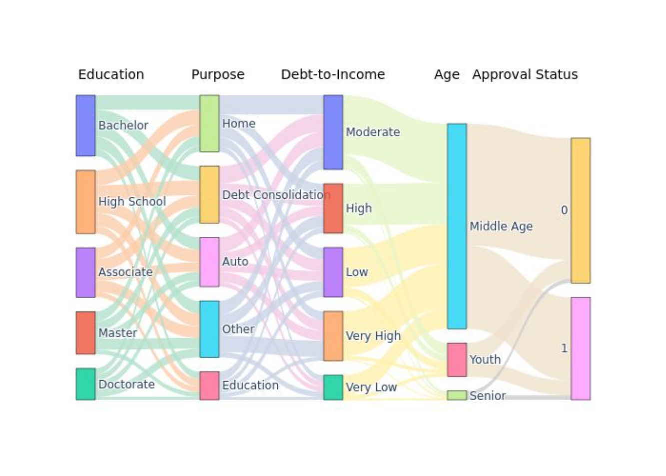

Iteration 1
Iteration 2
Visualize features relevant to behavioural risk
Visualize BehaviouralRisk vs other features
Knowledge
Visualization
Data
Visualization
Knowledge
Features correlate strongly with Loan Approval
Merged Dataset
BehaviouralRisk key indicator of overall Financial Risk
Models
Models
Cluster points based on those features
Random Forest to predict BehaviouralRisk
Add new feature BehaviouralRisk
Feedback



Low,Very Low
208
Very Low,Very High
192
Very Low,High
196
Very Low,Moderate
303
Very Low,Low
222
Very Low,Very Low
87
Moderate,Moderate
1006
Moderate,Low
667
Moderate,Very Low
307
Low,Low
470
High,High
417
High,Moderate
575
High,Low
377
Moderate,High
571
Low,Moderate
679
Very High,High
426
Very High,Very High
685
Very High,Low
264
Very High,Moderate
437
High,Very High
421
Very High,Very Low
188
Moderate,Very High
449
High,Very Low
210
Low,High
390
Low,Very High
253



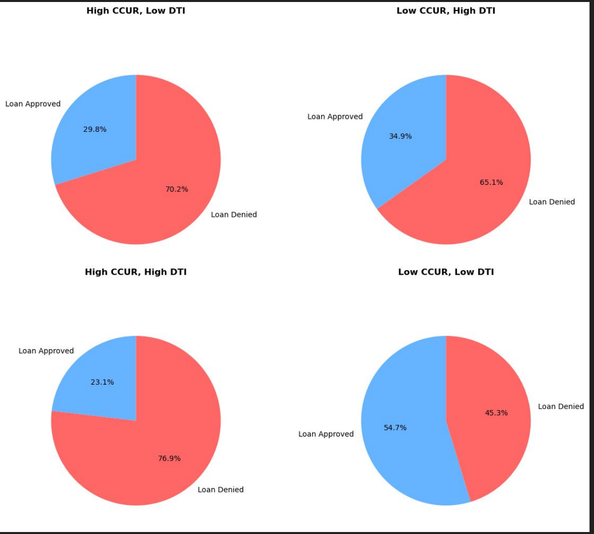
Youth & High
524
Youth & Moderate
476
Youth & Low
359
Middle Age & High
2350
Middle Age & Low
2528
Middle Age & Moderate
3404
Senior & High
126
Senior & Moderate
120
Senior & Low
113

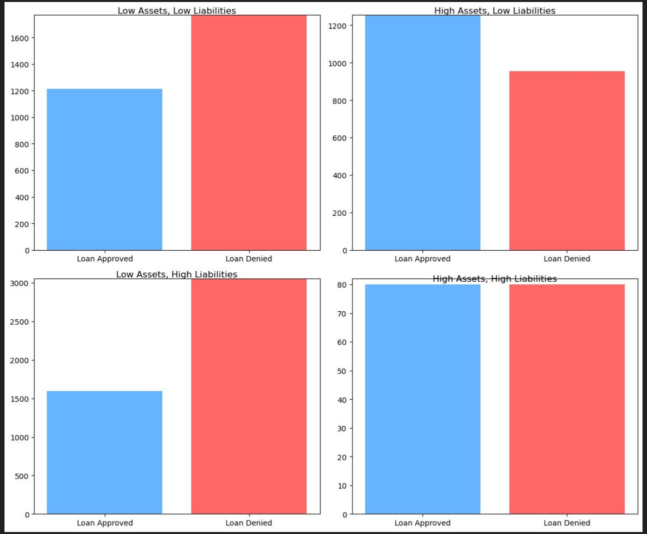


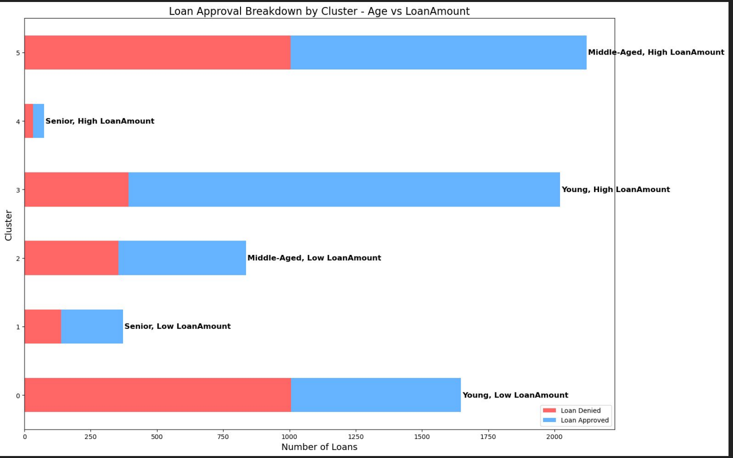
Education
Purpose
Debt-to-Income
Age
Approval Status
Bachelor
High School
Associate
Master
Doctorate
Home
Debt Consolidation
Auto
Other
Education
Moderate
High
Low
Very High
Very Low
Middle Age
Youth
Senior
0
1

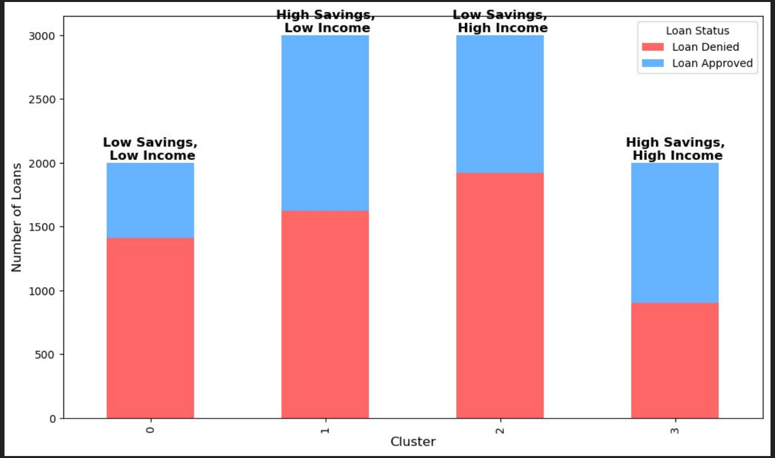
Initial patterns observed

- Most borrowers are middle-aged and middle-class
- A low savings-to-income ratio leads to loans not getting approved
- A decent number of people have high DTI ratio and high credit card utilization


High CCUR, Low DTI
Loan Approved
29.8%
70.2%
Loan Denied
Low CCUR, High DTI
Loan Approved
34.9%
65.1%
Loan Denied
High CCUR, High DTI
Loan Approved
23.1%
76.9%
Loan Denied
Low CCUR, Low DTI
Loan Denied
45.3%
Loan Approved
54.7%


Low Assets, Low Liabilities
High Assets, Low Liabilities
Low Assets, High Liabilities
High Assets, High Liabilities
Loan Approved
Loan Denied


Loan Approval Breakdown by Cluster - Age vs LoanAmount
Middle-Aged, High LoanAmount
Senior, High LoanAmount
Young, High LoanAmount
Middle-Aged, Low LoanAmount
Senior, Low LoanAmount
Young, Low LoanAmount
Loan Denied
Loan Approved
Cluster
Number of Loans


Low Savings, Low Income
High Savings, Low Income
Low Savings, High Income
High Savings, High Income
Loan Status
Loan Denied
Loan Approved
Number of Loans
Cluster

Patterns after clustering

- Low assets increase risk of rejection, regardless of liabilities.

- Higher savings did not always indicate better financial health, especially without high annual income to support them.

- Young borrowers seeking larger loans were often approved.

Create new column - BehaviouralRisk

Why?

- Bridging the gap between traditional and behavioural metrics

- To help with simplifying a lot of complex interactions

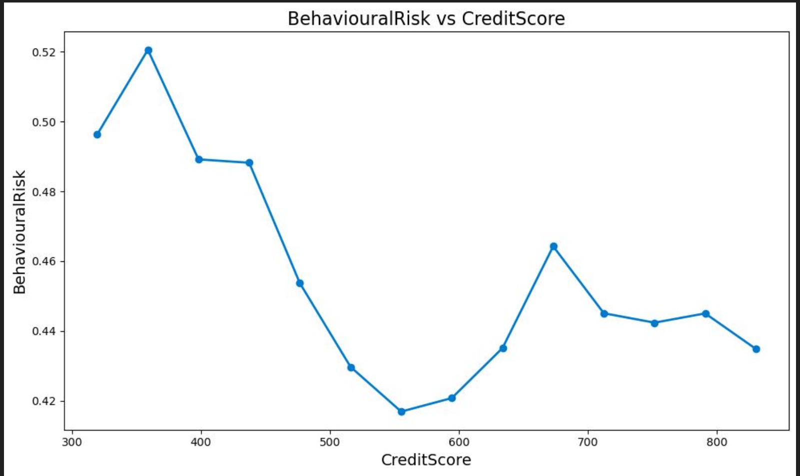
- Hope to explain anomalies/ambiguities

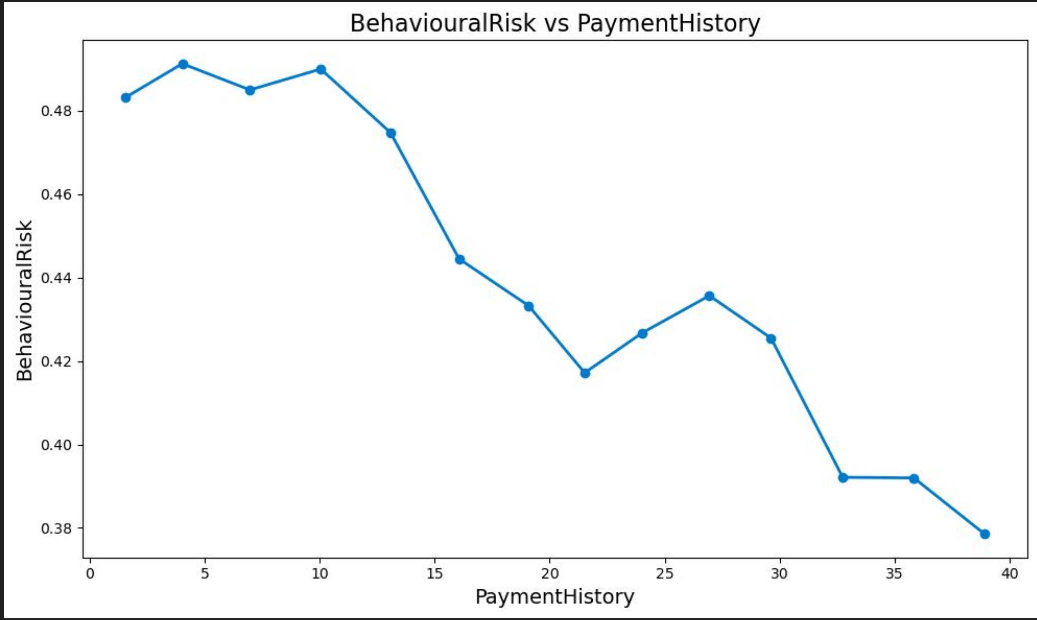
Visualize and Predict BehaviouralRisk

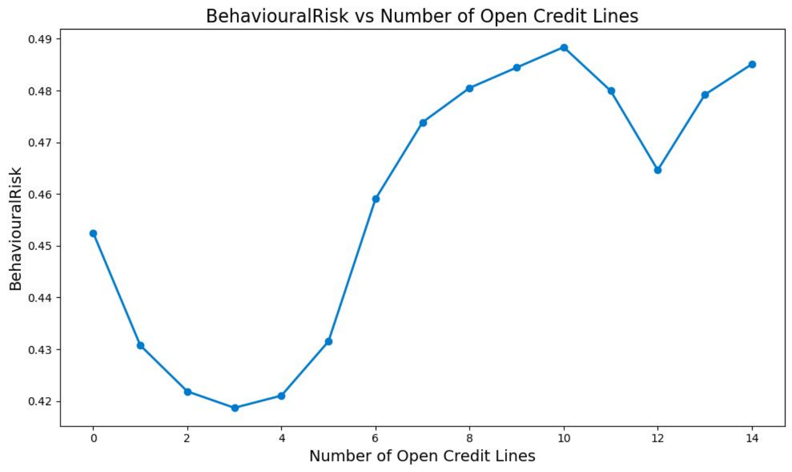
- Should be <0.7 for a good chance of loan approval

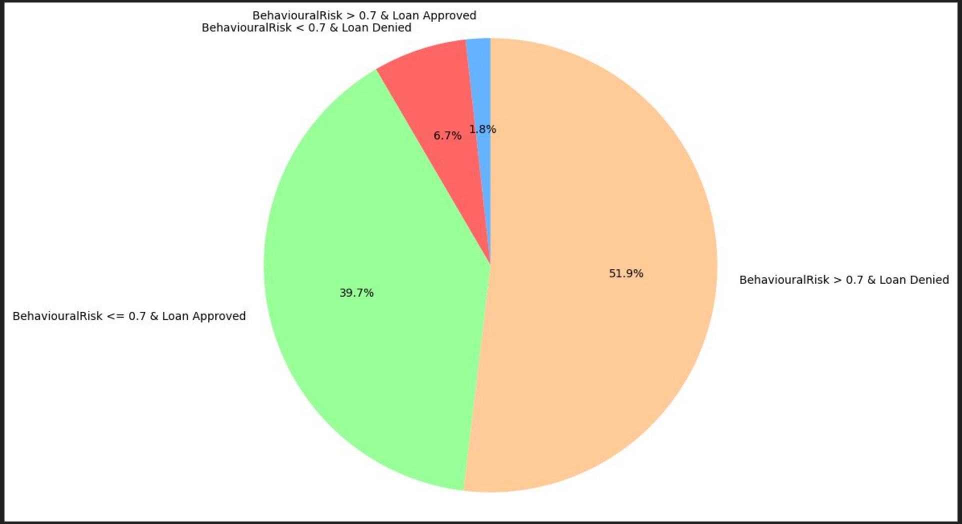
- Correlated with credit scores

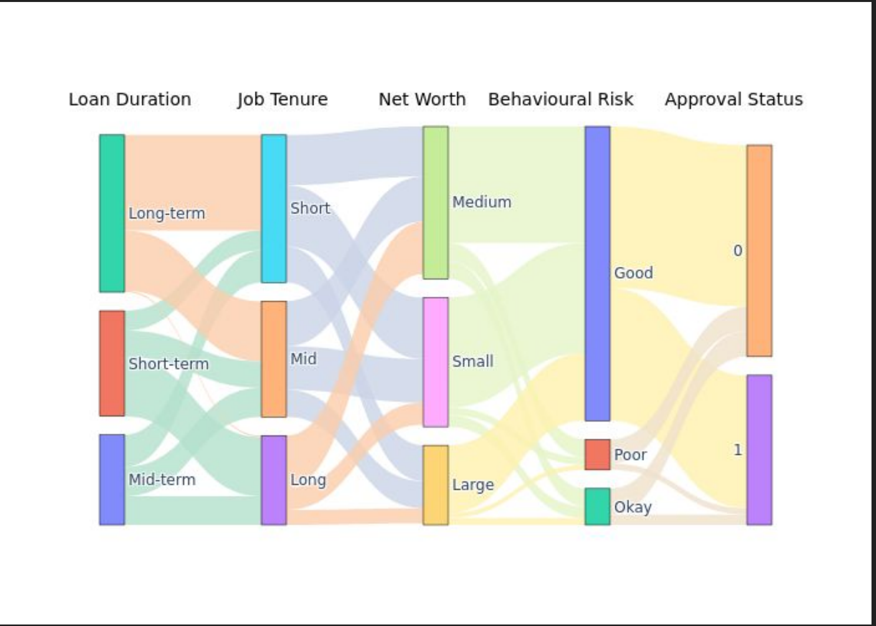
- Predicted BehaviouralRisk with over 80% accuracy


BehaviouralRisk vs CreditScore
BehaviouralRisk
CreditScore
0.52
0.50
0.48
0.46
0.44
0.42
300
400
500
600
700
800


BehaviouralRisk vs PaymentHistory
BehaviouralRisk
PaymentHistory
0.48
0.46
0.44
0.42
0.40
0.38
0
5
10
15
20
25
30
35
40


BehaviouralRisk vs Number of Open Credit Lines
BehaviouralRisk
Number of Open Credit Lines
0.49
0.48
0.47
0.46
0.45
0.44
0.43
0.42
0
2
4
6
8
10
12
14


BehaviouralRisk > 0.7 & Loan Approved
BehaviouralRisk < 0.7 & Loan Denied
6.7%
1.8%
51.9%
BehaviouralRisk > 0.7 & Loan Denied
39.7%
BehaviouralRisk <= 0.7 & Loan Approved


Loan Duration
Job Tenure
Net Worth
Behavioural Risk
Approval Status
Long-term
Short-term
Mid-term
Short
Mid
Long
Medium
Small
Large
Good
Poor
Okay
0
1

Final understandings

- There were patterns in the data based on loan approvals

- These patterns were visualized, but could not be fully explained using existing risk analysis methods

- Introducing behavioural risk allowed a more nuanced but still accurate risk assessment, while being able to explain the underlying patterns.