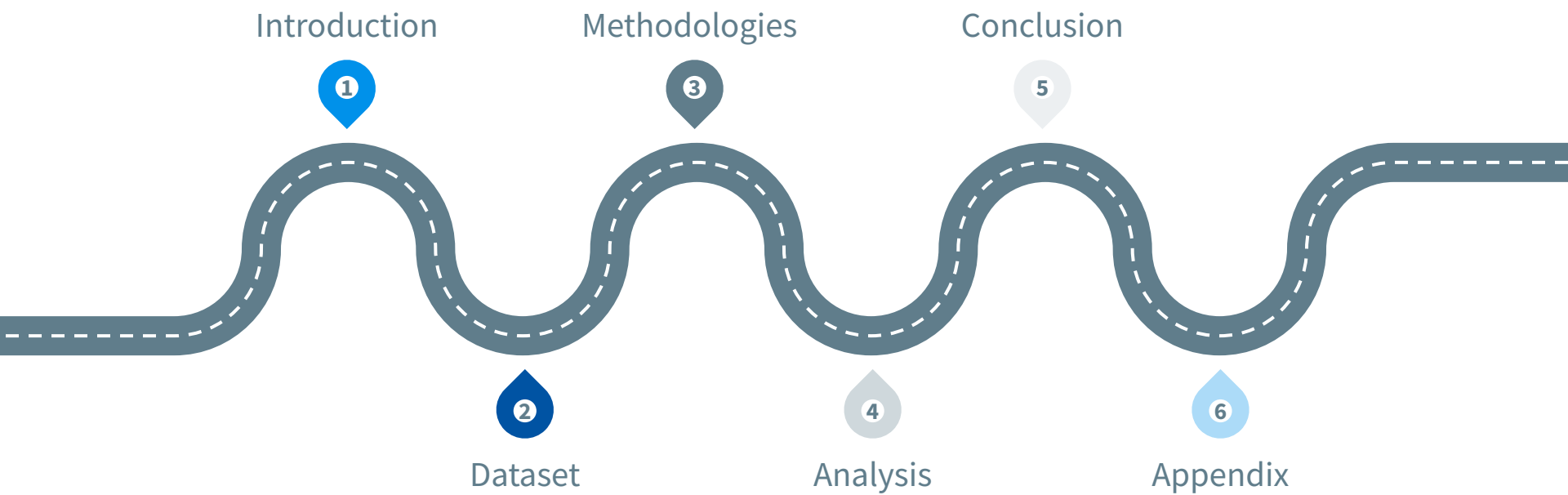
Customer Credit Risk Analysis

Bank of Questrom

Table of Contents

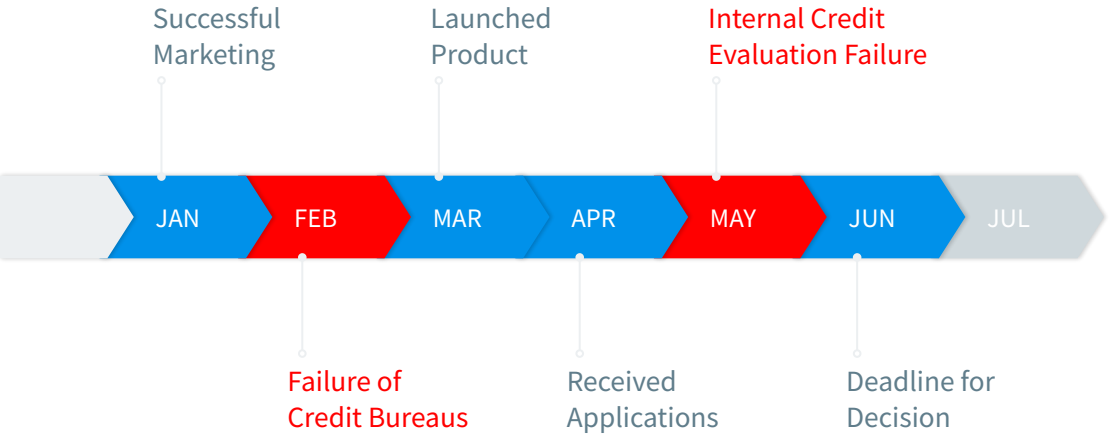


1.
Introduction

Bank of Questrom

Bank of Questrom

Financial Product ‘BA305’ Timeline

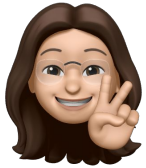


Urgent Need for Credit Risk Evaluation!

Credit Risk Analyst Team

BA 305 Team Only One

Anshi Mittal
Data Scientist



Michelle M. Lu
Business Analyst

Neri A. Arreaga
Data Analyst

David E. Kim
B.I. Engineer

1. *What are the **most important features** in classifying someone with high credit risk?*

2. *What are the **characteristics of an average customer** with low or high credit risk?*

3. *What are the **optimal benchmarks for the application** approval and rejection?*

2.
Dataset

Bank of Questrom

100,000 Observations

8 Months For All 12,500 Unique Clients

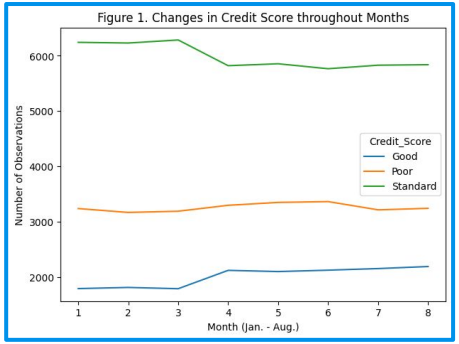
Data Type

Before

Pooled Data

*Time Series + Cross Section

	ID	Customer_ID	Month	Name
0	0x1602	CUS_0xd40	January	Aaron Maashoh
1	0x1603	CUS_0xd40	February	Aaron Maashoh
2	0x1604	CUS_0xd40	March	Aaron Maashoh
3	0x1605	CUS_0xd40	April	Aaron Maashoh
4	0x1606	CUS_0xd40	May	Aaron Maashoh



After

Cross Sectional Data

*Remove Time Series (Months)

	ID	Customer_ID	Month	Name
0	0x1609	CUS_0xd40	8	NaN
1	0x1615	CUS_0x21b1	8	Rick Rothackerj
2	0x1621	CUS_0x2dbc	8	Langep
3	0x162d	CUS_0xb891	8	Jasond
4	0x1639	CUS_0x1cdb	8	Deepaa

Data Pre-processing

Feature Selection

- Drop
  - Identification variables
  - Time related variables
  - ‘Occupation’ variable

ID	Customer_ID	Month	Name	Age	SSN	Occupation
0x1609	CUS_0xd40	8	NaN	23	#F%$D@*&8	Scientist
0x1615	CUS_0x21b1	8	Rick Rothackerj	28	004-07-5839	Teacher

```
['Monthly_Inhand_Salary', 'Changed_Credit_Limit', 'Payment_of_Min_Amount',
 'Total_EMI_per_month', 'Amount_invested_monthly', 'Payment_Behaviour', 'Monthly_Balance'],
inplace=True)
```

Missing Values

- Removed
  - NaN values
    - .dropna()
  - ‘_’ values
    - manual removal

```
df.Credit_Mix.unique()
```

```
array(['Good', '_', 'Standard', 'Bad'], dtype=object)
```

Data Pre-processing

Reformatting

- String to Integer and Floats
  - Remove non-numeric chars.
- 00 Years and 00 Months
  - Recalculation to Months

Outliers

- Removed the outliers that represent approximately **1%** of the data in the columns.

```
df.Num_of_Delayed_Payment.unique()
```

```
array(['4', '6', '2', '14', '11', '8',
       '13', '18', '17', '10', '23', '2
       '3', '0', '5', '14_', '538', '25
       '20_', '3539', '3684', '19_', '-
       '10_', '23_', '3253', '3858', '1
       '733', '2323', '7_', '25_', '28_
       '3_', '15_', '4042', '9_', '2589
       '960', '0_', '24_', '1673', '4_
```

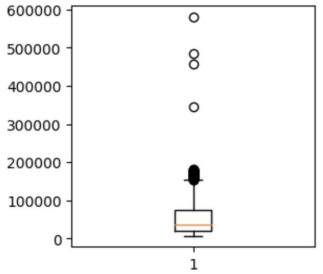
```
df.Credit_History_Age.head()
```

```
0     27 Years and 2 Months
1     18 Years and 4 Months
2     31 Years and 2 Months
3    21 Years and 11 Months
4     27 Years and 5 Months
Name: Credit_History_Age, dtype: object
```

```
# Outliers

df.Annual_Income.value_counts(normalize=True, bins=5)
```

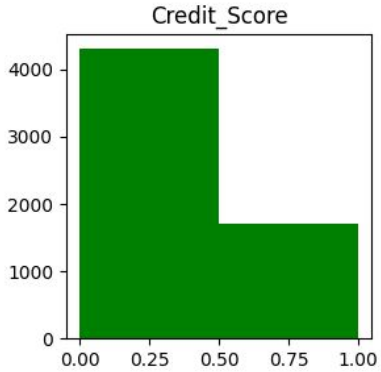
```
(-16821.693, 4772544.4]      0.992134
(14303621.2, 19069159.6]     0.003078
(9538082.8, 14303621.2]      0.001710
(4772544.4, 9538082.8]       0.001596
(19069159.6, 23834698.0]     0.001482
Name: Annual_Income, dtype: float64
```

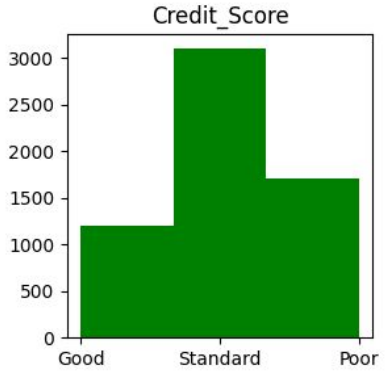


Data Pre-processing

Dependent Variable

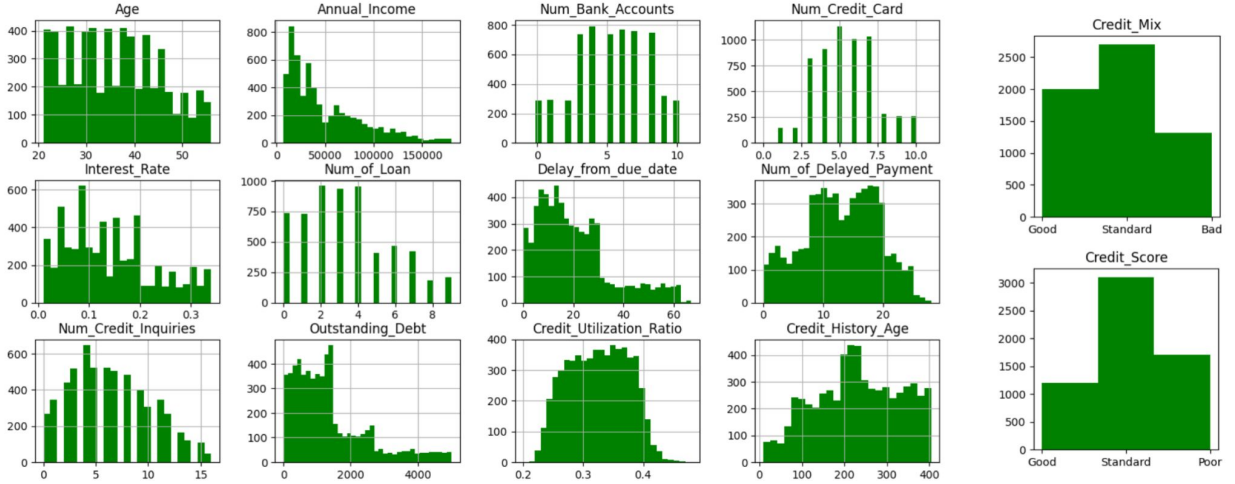
- Decision Tree
  - 0: Good & Standard
  - 1: Poor
- K-Means Clustering
  - Good
  - Standard
  - Poor





Data Pre-processing

```
RangeIndex: 6015 entries, 0 to 6014
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       6015 non-null   int64
 1   Annual_Income             6015 non-null   int64
 2   Num_Bank_Accounts         6015 non-null   int64
 3   Num_Credit_Card           6015 non-null   int64
 4   Interest_Rate             6015 non-null   float64
 5   Num_of_Loan               6015 non-null   int64
 6   Delay_from_due_date       6015 non-null   int64
 7   Num_of_Delayed_Payment    6015 non-null   int64
 8   Num_Credit_Inquiries      6015 non-null   int64
 9   Credit_Mix                6015 non-null   object
 10  Outstanding_Debt          6015 non-null   int64
 11  Credit_Utilization_Ratio  6015 non-null   float64
 12  Credit_History_Age        6015 non-null   int64
 13  Credit_Score              6015 non-null   int64
dtypes: float64(2), int64(11), object(1)
memory usage: 658.0+ KB
```



6,015 Observations

2 Categorical & 12 Numerical Variables

Correlations

Assumptions

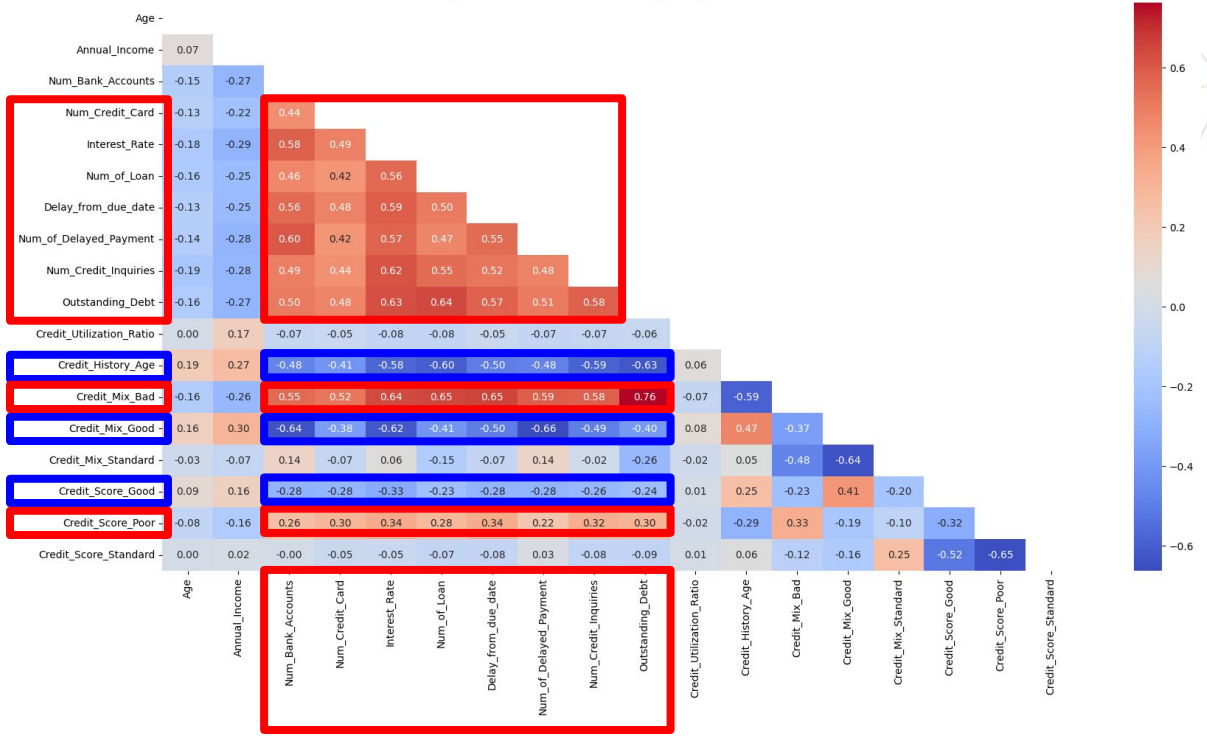
Red Highlights

- Negative effects on Credit Score

Blue Highlights

- Positive effects on Credit Score

Figure 3. Correlation Matrix including Categorical Variables



3.
Methodologies

Bank of Questrom

Decision Trees

1. What are the **most important features** in classifying someone with high credit risk?

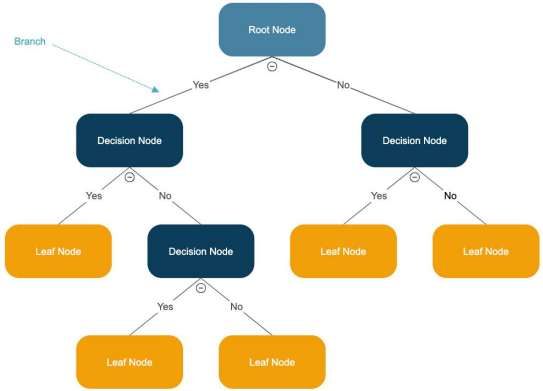


Figure 4.1 Full Decision Tree Model ROC curve

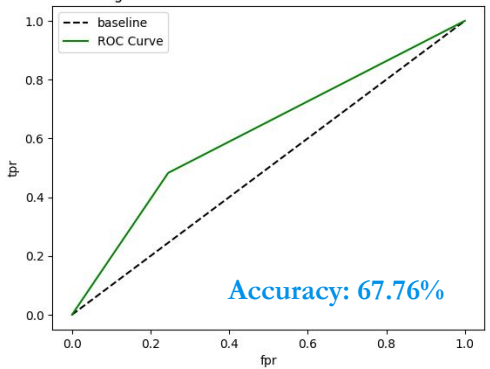
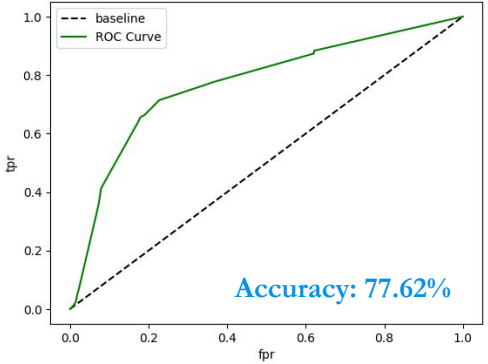
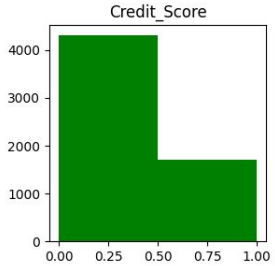
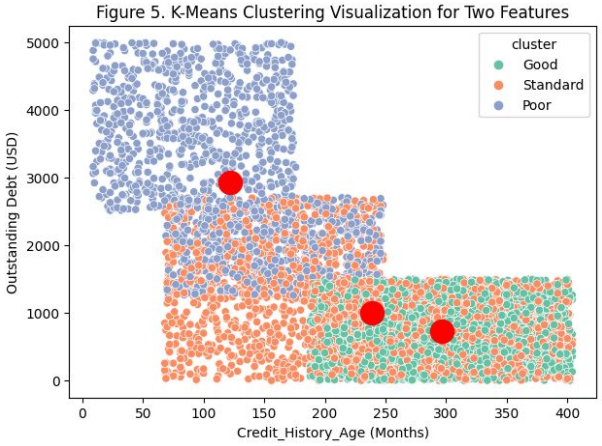
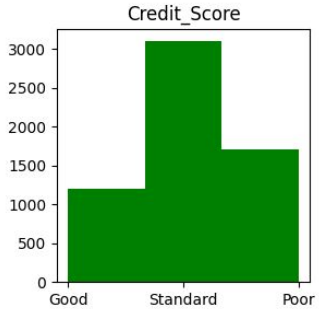


Figure 4.2 Reduced Decision Tree Model ROC curve









K-Means Clustering

2. What are the **characteristics of an average customer** with low credit risk or high credit risk?

4.
Analysis

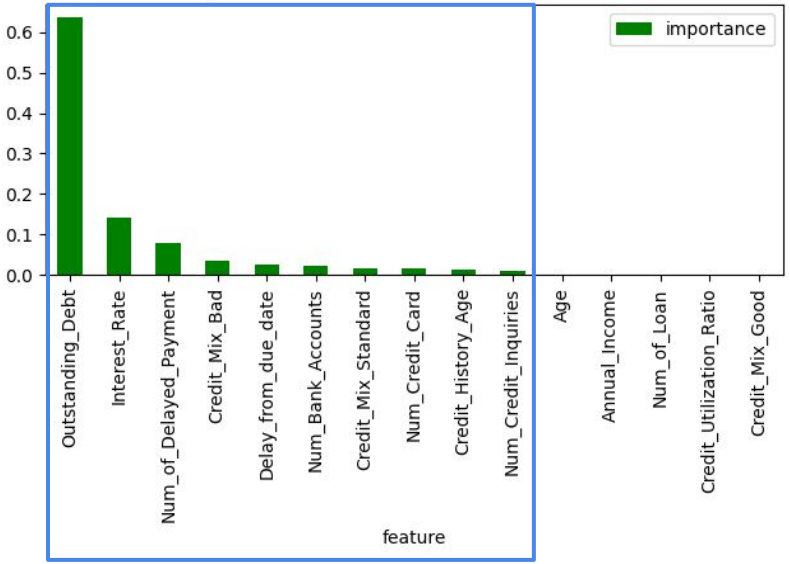
Bank of Questrom

Decision Tree

Feature Importance

- Outstanding_Debt
- Interest_Rate
- Num_of_Delayed_Payment
- Delay_from_due_date
- Num_Bank_Accounts
- Num_Credit_Card
- Num_Credit_Inquiries

Figure 6. Decision Tree Feature Importance



Decision Tree

Reduced Decision Tree

- Outstanding debt higher than 2488 USD and interest rate higher than 34%.

- Outstanding debt between 1493.5 and 2488 USD and the number of bank accounts higher than 5.5.

Figure 7. Reduced Decision Tree Plot

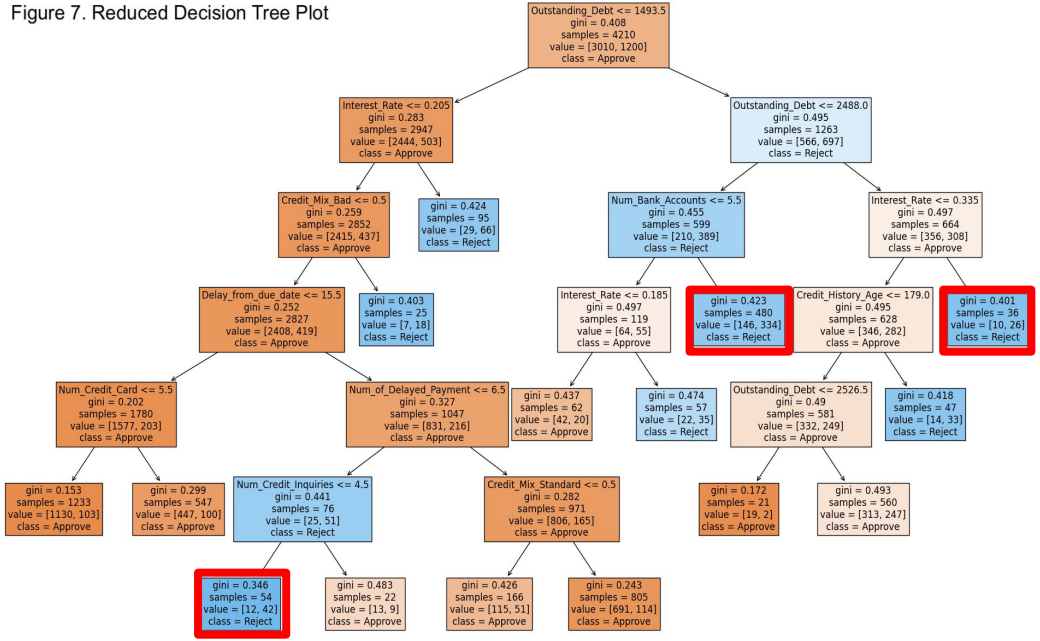
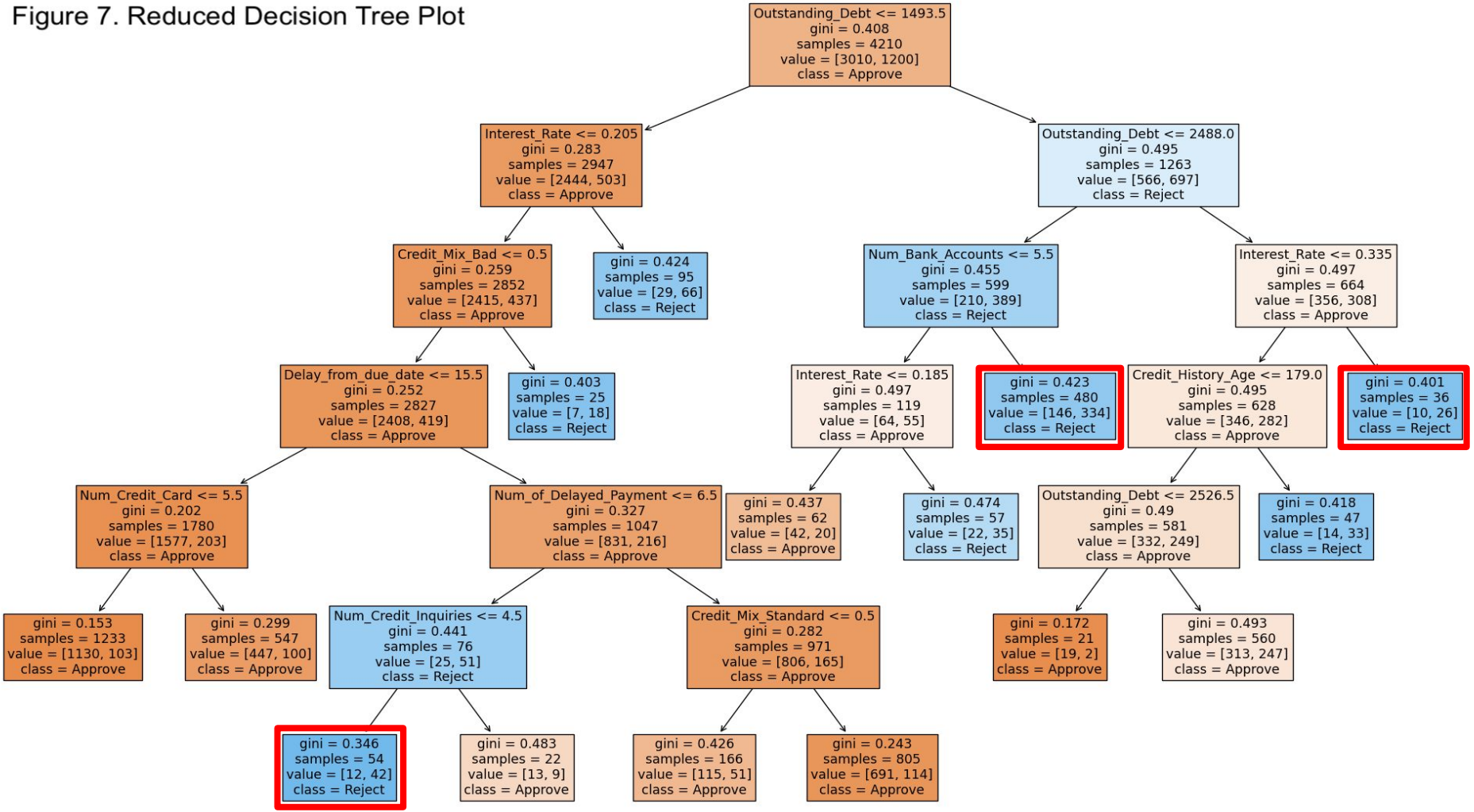


Figure 7. Reduced Decision Tree Plot

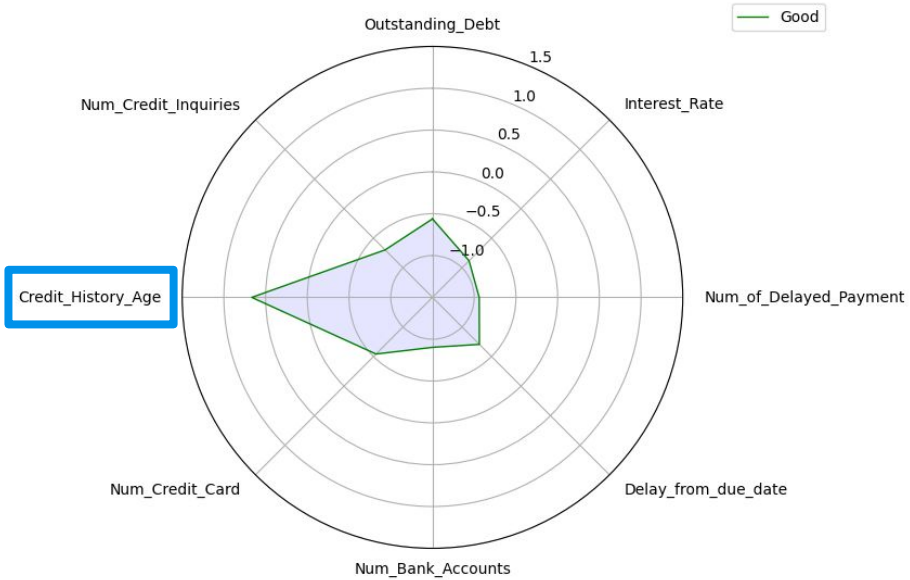


K-Means Clustering

Low Credit Risk

- **Higher average Credit_History_Age**
- Lower average for everything else

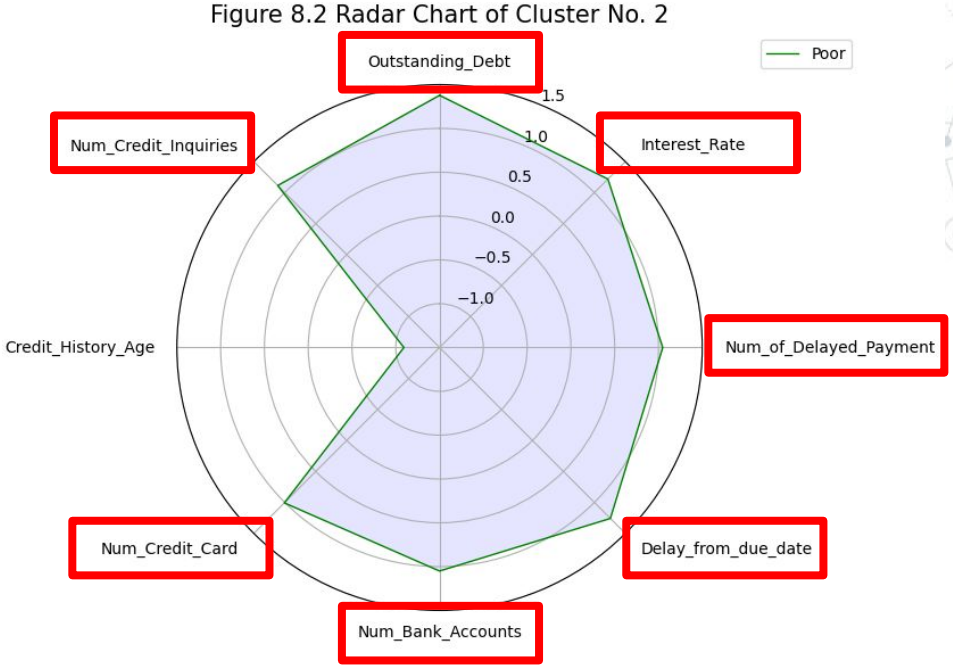
Figure 8.1 Radar Chart of Cluster No. 1



K-Means Clustering

High Credit Risk

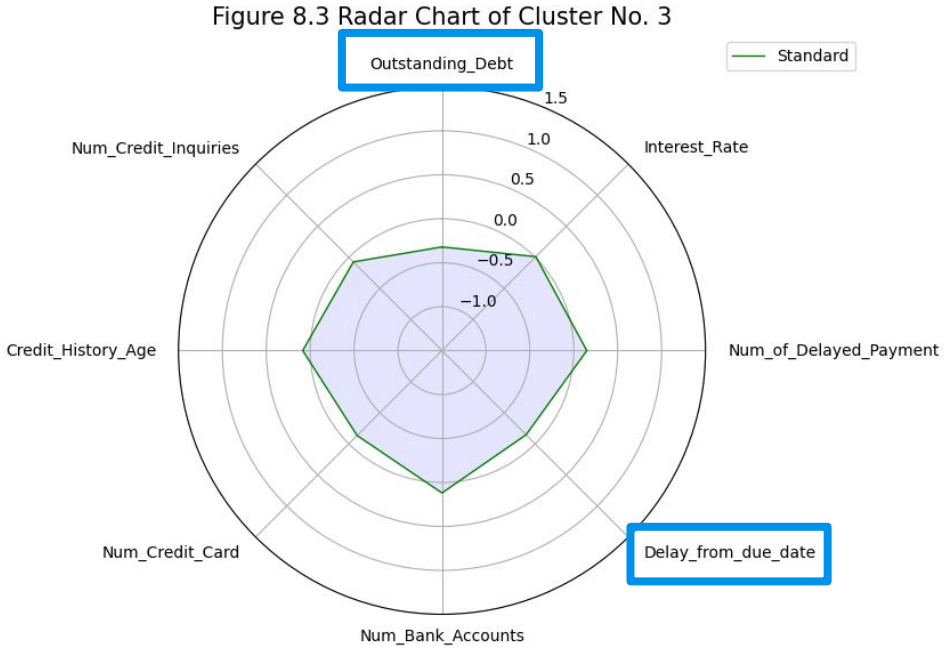
- Opposite of low credit risk
- Lower average Credit_History_Age
- **Higher average for everything else**



K-Means Clustering

Standard Credit Risk

- In between low and high credit risk
- **Outstanding_Debt** and **Delay_from_due_date** have the highest difference compared to high credit risk



5.
Conclusion

Bank of Questrom

Result

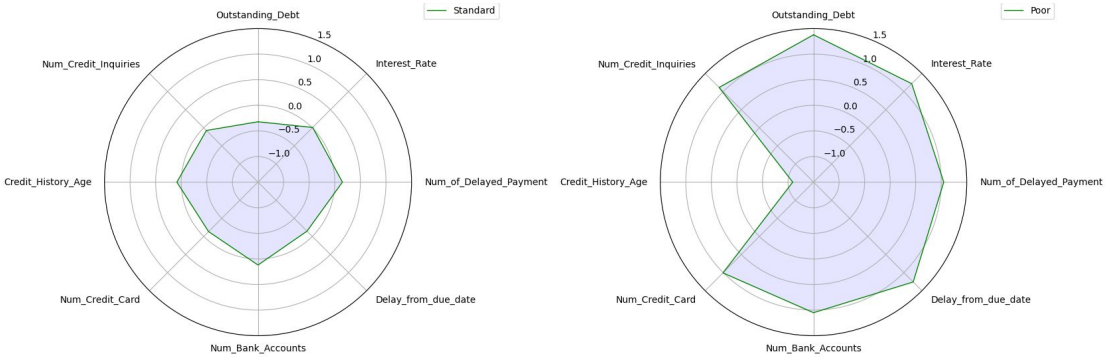
1. Amount of **outstanding debt** + **interest rate** ⇒ Most Important in identifying high credit risk!

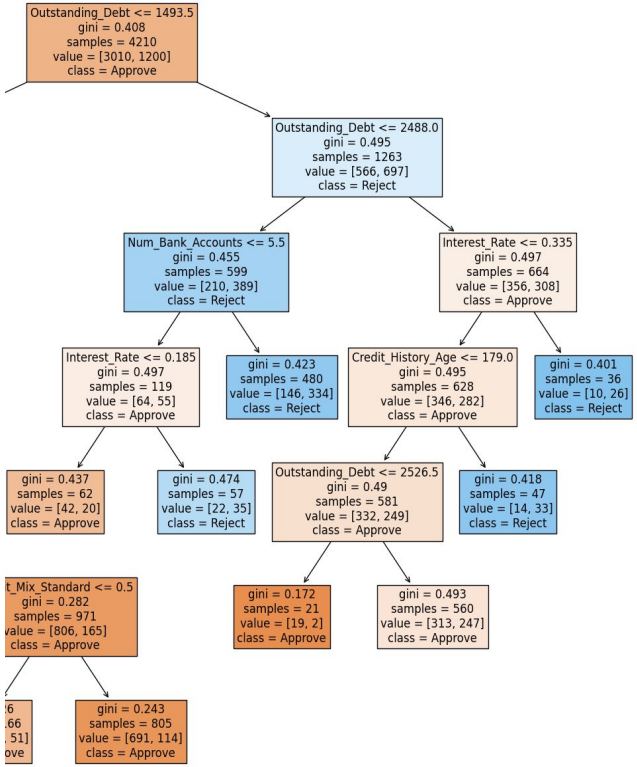
2. Excessive number of **delayed payments**, **accounts**, **credits**, and **inquiries** ⇒ Can be red flag!

3. **Lengthy credit history age** ⇒ Good indicator of low credit risk.

   But! **Negative factors** may diminish the positive influence of the credit history age.

Optimal Benchmark





3. What are the **optimal benchmarks for the application** approval and rejection?

Reject the following criteria

(Debt amount > 1900 USD) & (Interest rate higher > 20%) & (Delay from due date > 28 days)

Testing the Evaluation Model

- Unfortunately, the decision tree model was **unable to identify** the average person in the **'high credit risk'** cluster.
- Problem: **Low sensitivity score**
- Only identifies **extreme (polar) cases**

```
c_test = pd.read_csv('data/centroid_test.csv')

y_pred = dt.predict(c_test)
y_pred
```

```
array([0, 0, 0])
```

Improvement

Increased Data

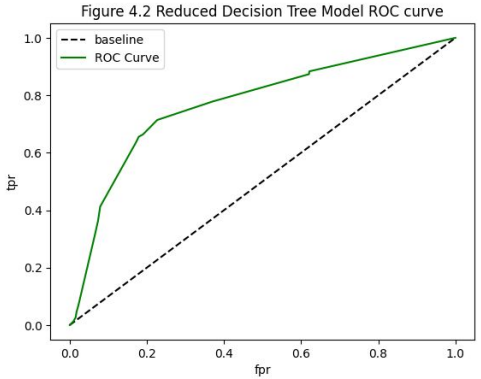
- Include all **8 months**
- 6015 ⇒ **47,900 Observations**

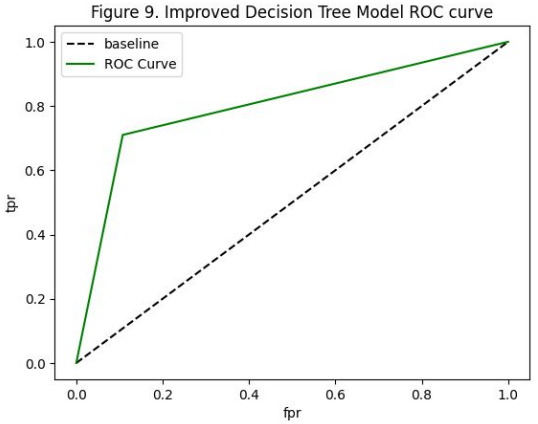
Accuracy

- 77.62% ⇒ **84.18%**

F1 Score & Sensitivity

- 51.21% ⇒ **71.35%**
- 41.25% ⇒ **71.01%**





Success!

The improved evaluation model **identified** the average individual within the **‘high credit risk’** cluster.

Thanks!

Any questions?

Appendix

6.1 Codes

6.1.1 1_credit_data_cleaning.ipynb
Data cleaning file with the dependent variable 'Credit_Score' as a polytomous variable
Unique values of dependent variable: Good, Standard, Poor

6.1.2 1_credit_data_cleaning2.ipynb
Data cleaning file with the dependent variable 'Credit_Score' as a binary variable
Unique values of dependent variable: 0 (Includes Good and Standard), 1 (Poor)

6.1.3 2_credit_data_correlation.ipynb
File for finding correlations between variables via correlation matrix heatmap.

6.1.4 3_credit_knn.ipynb
KNN classification model with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.5 3_credit_knn2.ipynb
KNN classification model with cleaned_credit_data2.csv (Binary dependent variable)

6.1.6 4_credit_decision_tree.ipynb
Decision tree model with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.7 4_credit_decision_tree2.ipynb
Decision tree model with cleaned_credit_data2.csv (Binary dependent variable)

6.1.8 5_credit_kmeans.ipynb
K-means clustering with cleaned_credit_data.csv (Polytomous dependent variable)

6.1.9 experiment.ipynb
File for experimental improvement stage for the study
Increased number of observations by including all months instead of only 'August'
Decision tree model with binary dependent variable

6.2 Data

6.2.1 credit_raw.csv
Original unprocessed data from the following url
https://www.kaggle.com/datasets/parisrohan/credit-score-classification
100,000 Entries with 28 columns

6.2.2 cleaned_credit_data.csv
Processed data with the dependent variable 'Credit_Score' as a polytomous variable
6015 Observations with 2 categorical and 12 numerical variables

6.2.3 cleaned_credit_data2.csv
Processed data with the dependent variable 'Credit_Score' as a binary variable
6015 Observations with 2 categorical and 12 numerical variables

6.2.4 centroids.csv
Centroids dataframe saved from '5_credit_kmeans.ipynb'
Standardized feature values of centroids within the three clusters

6.2.5 centroids_test.csv
Centroids dataframe used for testing the decision tree models
Unstandardized feature values of centroids within the three clusters