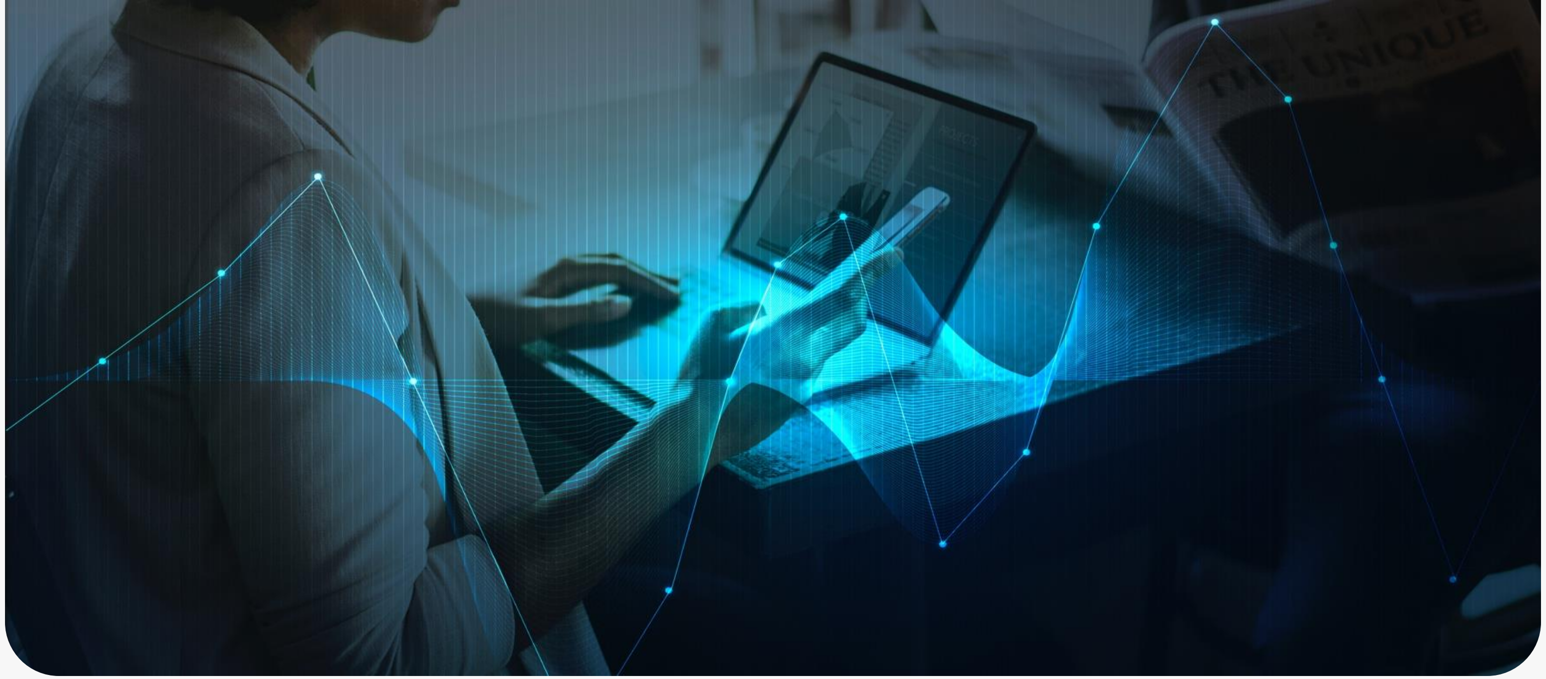
COMPREHENSIVE ANALYSIS OF

THERA BANK LIABILITY CUSTOMERS

Charles Bryant



Context

Objective

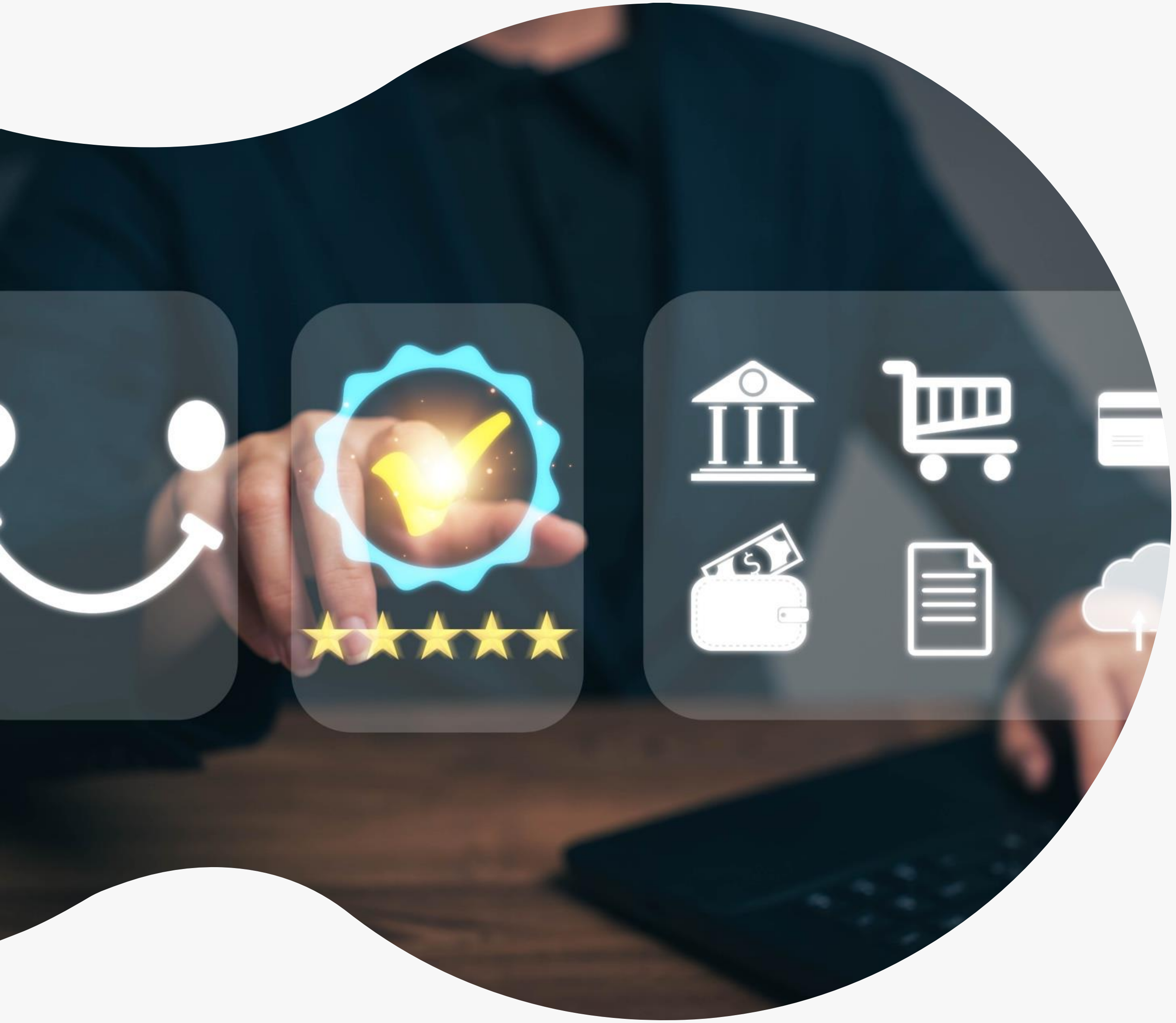
Thera Bank's management aims to convert liability customers into personal loan customers while ensuring they remain depositors.

Opportunity

A previous campaign achieved a promising 9% conversion rate, indicating potential for growth in this segment.

Strategy

The retail marketing team plans to develop targeted marketing campaigns to improve conversion rates further, focusing on cost efficiency and maximizing success.



Defining The Problem

In a recent conversion campaign, 9% of liability customers transitioned to personal loan customers. Despite a broad demographic approach, the campaign achieved its goals. To better serve our customers, we require a more targeted strategy to enhance customer satisfaction and improve conversion efficiency.

Dataset information

- Available on Kaggle
- Contains 5000 customers

Includes

- Demographic information
- The customer’s relationship with the bank
- Customer response to the last campaign

Dataset Attributes

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: Years of professional experience
- Income: Annual income of the customer ($000)
- ZIP Code: Home Address ZIP code
- Family: Family size of the customer
- CCAvg: Average spending on credit cards per month ($000)
- Education Level

  1 Undergrad 2. Graduate 3. Advanced/Professional
- Mortgage: Value of house mortgage if any ($000)
- Personal Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities Account: Does the customer have a securities account with the bank?
- CD Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Does the customer use internet banking facilities?
- Credit card: Does the customer use a credit card issued by the bank?

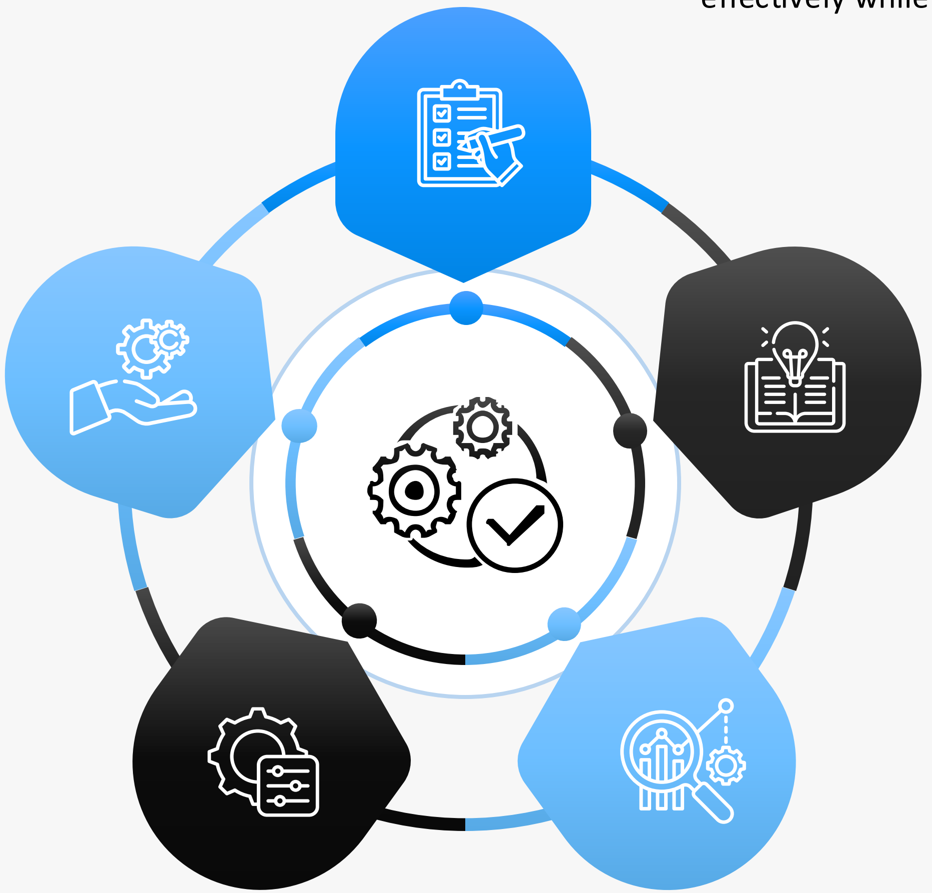
Data analysis Methodologies

Evaluation Metrics

Prioritized precision and recall to capture positive instances effectively while minimizing false positives.

Feature Selection

Identified the most predictive variables using variance inflation factor (VIF) and feature importance rankings.



Ensemble Learning

Leveraged ensemble learning methods to boost overall model performance.

Threshold Adjustment

Refined decision thresholds to better distinguish between majority and minority classes.

Interpretability

Used partial dependence plots to analyze non-linear relationships between variables.

DATA PRE-PROCESSING

Data View

Before Processing

ID	Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0

Pre-Processing

Dropping Irrelevant Columns

Removed the ID column to ensure the data focuses on meaningful attributes. The ID column served as a unique identifier.

Handling Anomalies in Experience Column

Removed negative values in the experience column and transformed into positive values to ensure data integrity and alignment with expectations.

Standardizing Scales

Converted CCAvg to an annual scale to ensure consistency with income data.

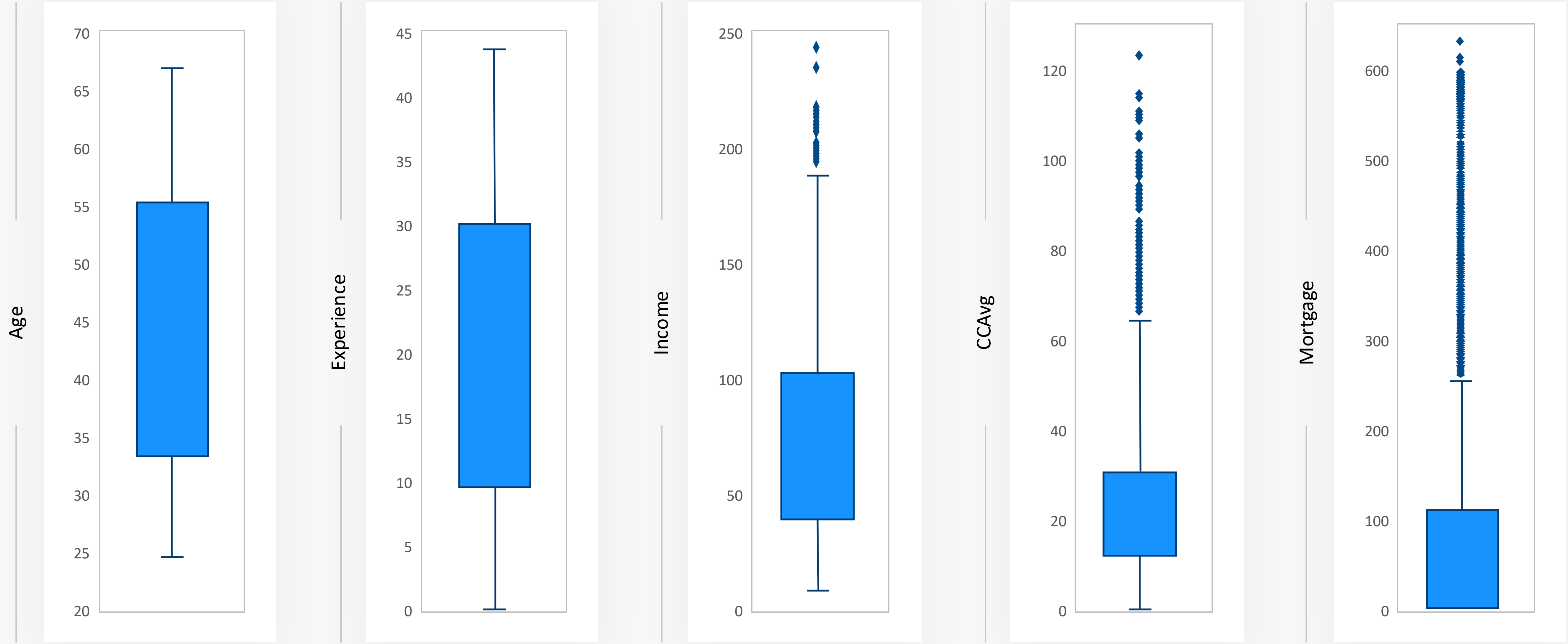
Data View

After initial processing

Age	Experience	Income	ZIP Code	Family	CCAvg	Education	Mortgage	Personal Loan	Securities Account	CD Account	Online	Credit Card
25	1	49	91107	4	19.2	1	0	0	0	1	0	0
45	19	34	90089	3	18.0	1	0	0	0	1	0	0
39	15	11	94720	1	12.0	1	0	0	0	0	0	0
35	9	100	94112	1	32.4	2	0	0	0	0	0	0
35	8	45	91330	4	12.0	2	0	0	0	0	0	1

Boxplots

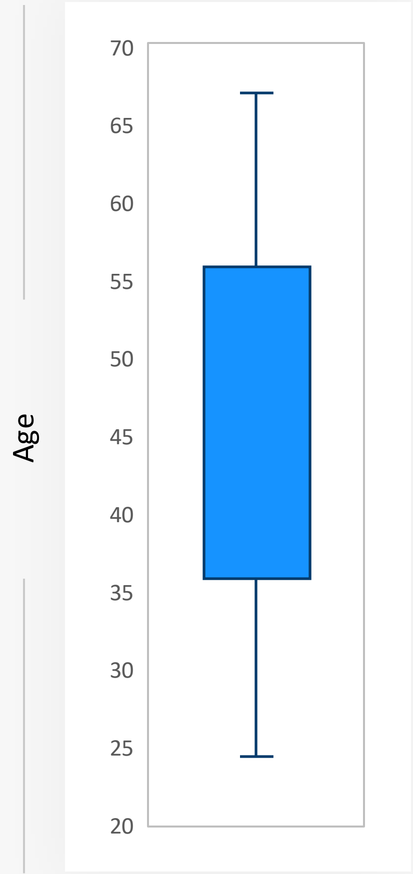
Before outlier removal

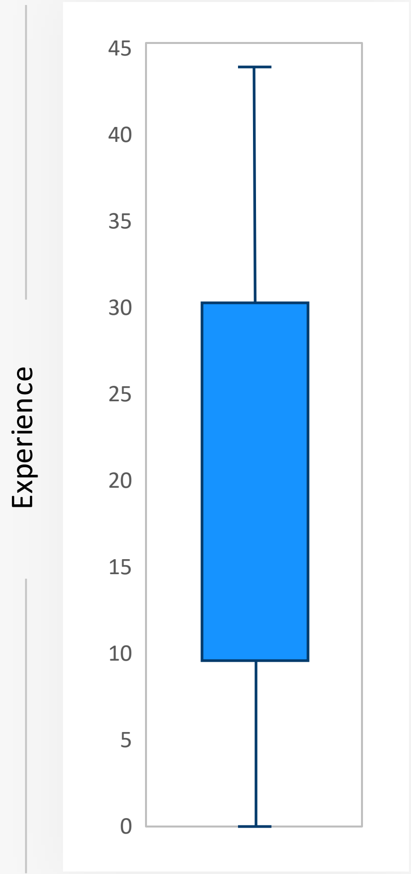


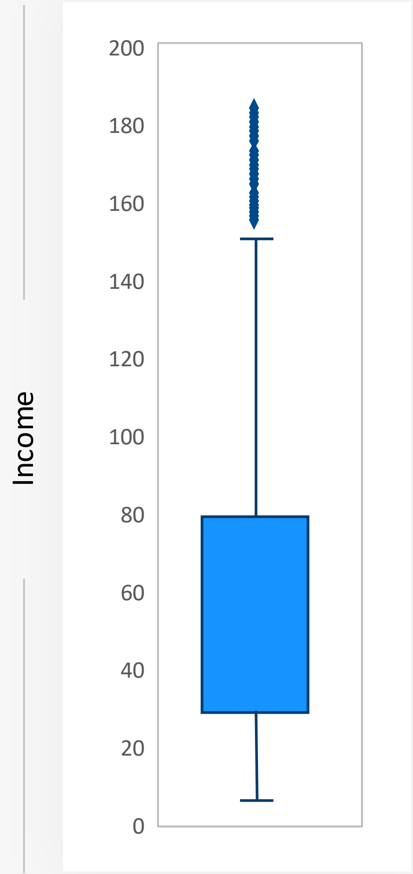
Boxplots

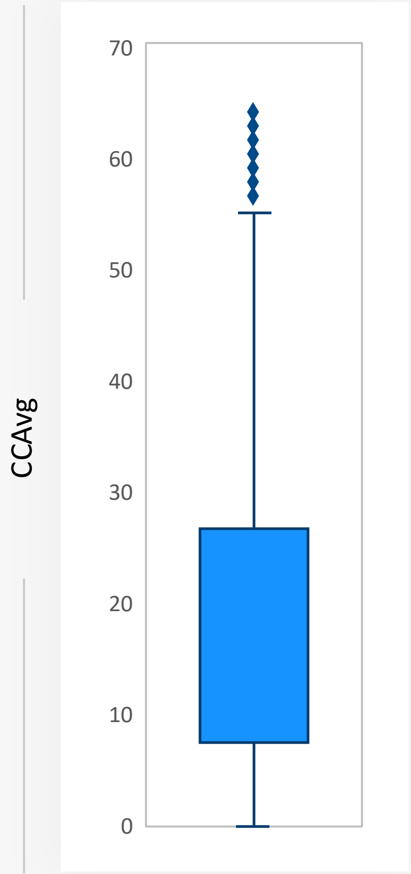
Income, CCAvg, and Mortgage had significant outliers. Using the IQR method, we removed them and analyzed summary statistics to assess their impact, ensuring a more reliable dataset for modeling.

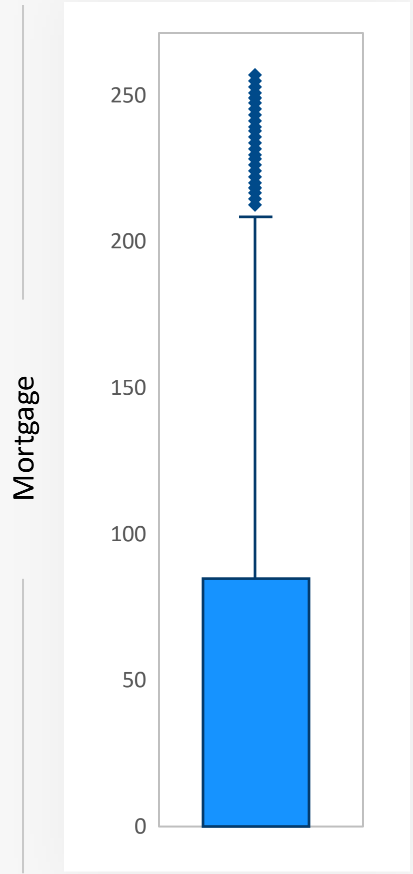
After outlier removal











Outliers comparison

Removing outliers significantly reduced the standard deviations for Mortgage, CCAvg, and Income, indicating less variability in these features. This improvement highlights a more consistent data distribution, which contributes to building a more robust and precise predictive model. Additionally, the removal of outliers decreased the maximum values of these features, aligning them more closely with typical observations and reducing the impact of extreme values.

Summary with Outliers

Statistic	Age	Experience	Income	CCAvg	Mortgage
Count	5000	5000	5000	5000	5000
Mean	45.3384	20.1346	73.7742	23.25526	56.4988
Std Dev	11.46317	11.41519	46.03373	20.97191	101.7138
Min	23	0	8	0	0
25%	35	10	39	8.4	0
50%	45	20	64	18	0
75%	55	30	98	30	101
Max	67	43	224	120	635

Summary without Outliers

Statistic	Age	Experience	Income	CCAvg	Mortgage
Count	4398	4398	4398	4398	4398
Mean	45.53661	20.3098	64.08458	18.61375	38.49068
Std Dev	11.49029	11.45877	38.02465	13.89401	68.10812
Min	23	0	8	0	0
25%	35	10	35	7.2	0
50%	46	20	58	16.8	0
75%	56	30	83	26.4	83
Max	67	43	185	62.4	252

Percentage of Outliers: 8.8%

Correlation Heat MAP

Before removal of insignificant features

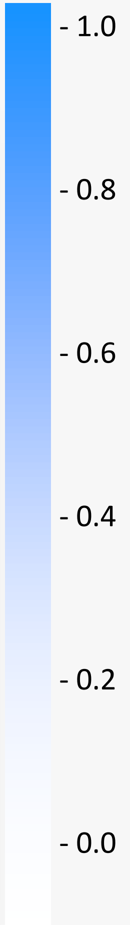
Age	1	1	-0.04	-0.03	-0.06	-0.02	0.03	-0.008	-0.02	0.006	0.006	0.02	0.01
Experience	1	1	-0.03	-0.03	-0.06	-0.03	0.008	-0.007	-0.01	0.004	0.008	0.02	0.01
Income	-0.04	-0.03	1	-0.1	-0.1	0.5	-0.1	-0.1	0.5	0.002	0.1	0.004	0.004
ZIP Code	-0.03	-0.03	-0.1	1	0.008	-0.008	-0.02	0.01	0.003	0.0003	0.02	0.003	0.02
Family	-0.06	-0.06	-0.1	0.008	1	-0.06	0.02	0.03	0.01	-8E-04	0.01	0.002	0.02
CCAvg	-0.02	-0.03	0.5	-0.008	-0.06	1	-0.07	0.3	0.02	0.08	-0.02	-0.002	-0.002
Education	0.03	0.008	-0.1	0.02	0.02	-0.07	1	-0.01	-0.02	-0.006	-0.02	-0.02	-0.02
Mortgage	-0.008	-0.007	-0.1	0.01	0.03	0.3	-0.01	1	-0.03	-0.009	0.01	-0.01	-0.002
Personal Loan	-0.02	-0.01	0.5	0.003	0.01	0.3	-0.02	-0.03	1	0.02	0.3	0.02	0.3
Securities Account	0.006	0.004	0.002	0.0003	−0.0008	0.08	-0.006	-0.009	0.02	1	0.005	0.02	-0.006
CD Account	0.006	0.008	0.1	0.02	0.01	-0.02	-0.02	0.01	0.3	0.005	1	0.2	0.3
Online	0.02	0.02	0.004	0.02	0.002	-0.002	-0.02	-0.01	0.02	0.02	0.2	1	-0.02
CreditCard	0.01	0.01	0.004	0.003	0.002	-0.002	-0.02	-0.002	0.3	-0.006	0.3	-0.02	1

Correlation Heat MAP

After removal of insignificant features

Age, Experience, and Zip Code had correlation coefficients below 0.1, indicating negligible predictive power; these features were removed. Several predictors showed high correlations, suggesting potential multicollinearity. To confirm this, the Variance Inflation Factor (VIF) was calculated to quantify and address multicollinearity among the features.

Income	1	-0.1	0.5	-0.1	-0.06	0.5	0.002	0.1	0.004	0.004
Family	-0.1	1	-0.06	0.02	0.03	0.01	−0.0008	0.01	0.002	0.002
CCAvg	0.5	-0.06	1	-0.07	-0.02	0.3	0.08	0.02	-0.02	-0.002
Education	-0.1	0.02	-0.07	1	0.01	0.1	-0.006	-0.02	-0.02	-0.02
Mortgage	-0.06	0.03	-0.02	0.01	1	-0.03	-0.009	-0.006	-0.01	-0.002
Personal Loan	0.5	0.01	0.3	0.1	-0.03	1	0.02	0.2	0.02	0.3
Securities Account	0.002	−0.0008	0.08	-0.006	-0.009	0.02	1	0.005	0.02	-0.006
CD Account	0.1	0.01	0.02	-0.02	-0.006	0.2	0.005	1	0.3	0.3
Online	0.004	0.002	-0.02	-0.02	-0.01	0.02	0.02	0.3	1	-0.02
CreditCard	0.004	0.002	-0.002	-0.02	-0.002	0.3	-0.006	0.3	-0.02	1



Finalize feature space

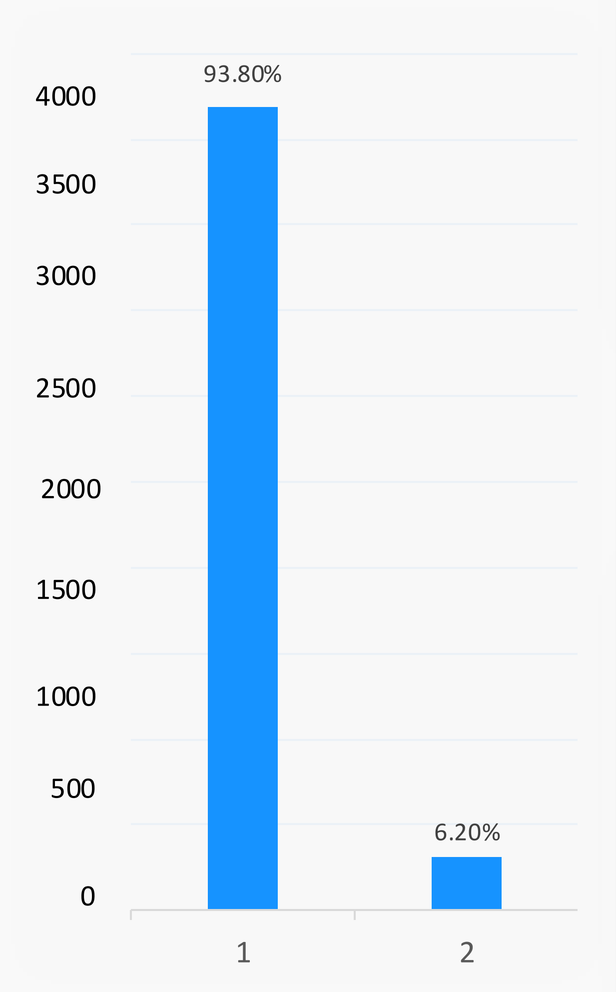
Income, CCAvg, and Education were identified as the most predictive features. Reducing the feature space improved the model's ability to distinguish between classes effectively.

Output from Random Forest Feature Importance

Feature	RF Feature Importance
Income	0.44
CCAvg	0.06
Education	0.5
Family	0.0
Mortgage	0.0
Personal Loan	0.0
Securities Account	0.0
CD Account	0.0
Online	0.0

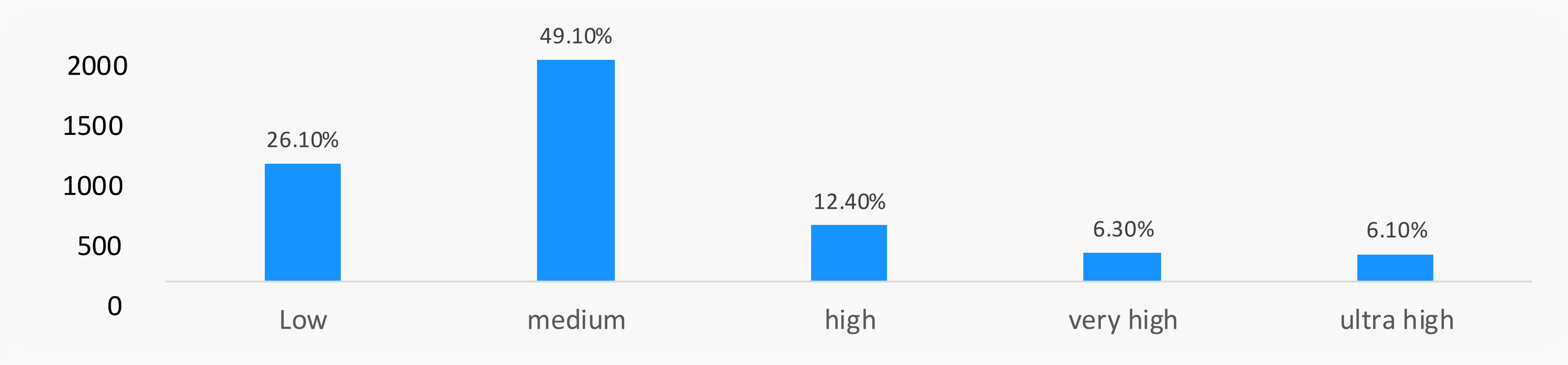
DATA VISUALIZATION

Distribution of Customers by Personal Loan Status



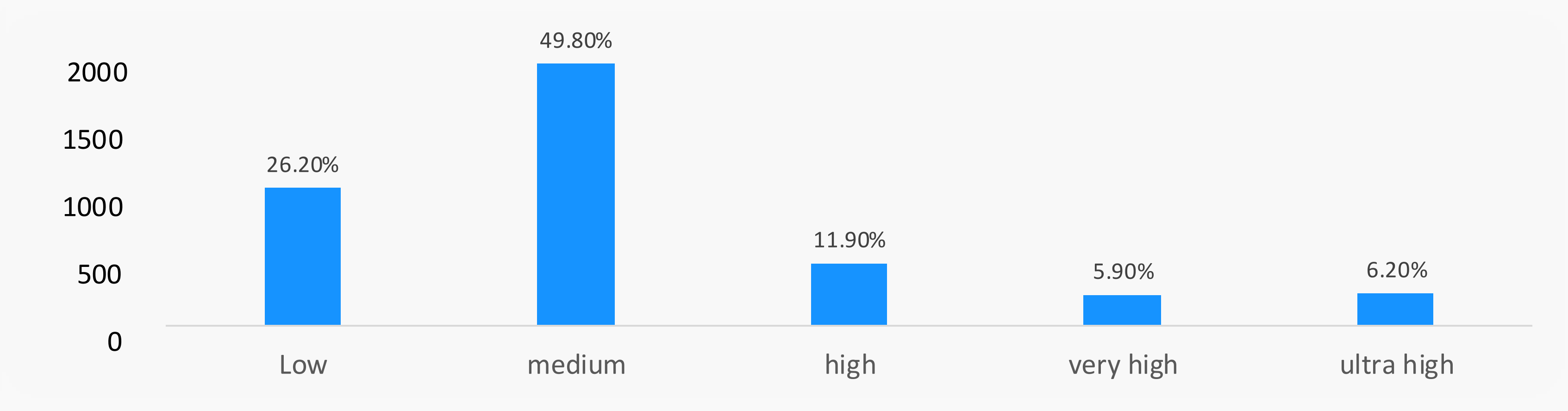
Personal Loan (0 = No, 1 = Yes

Distribution of Customers by income Levels



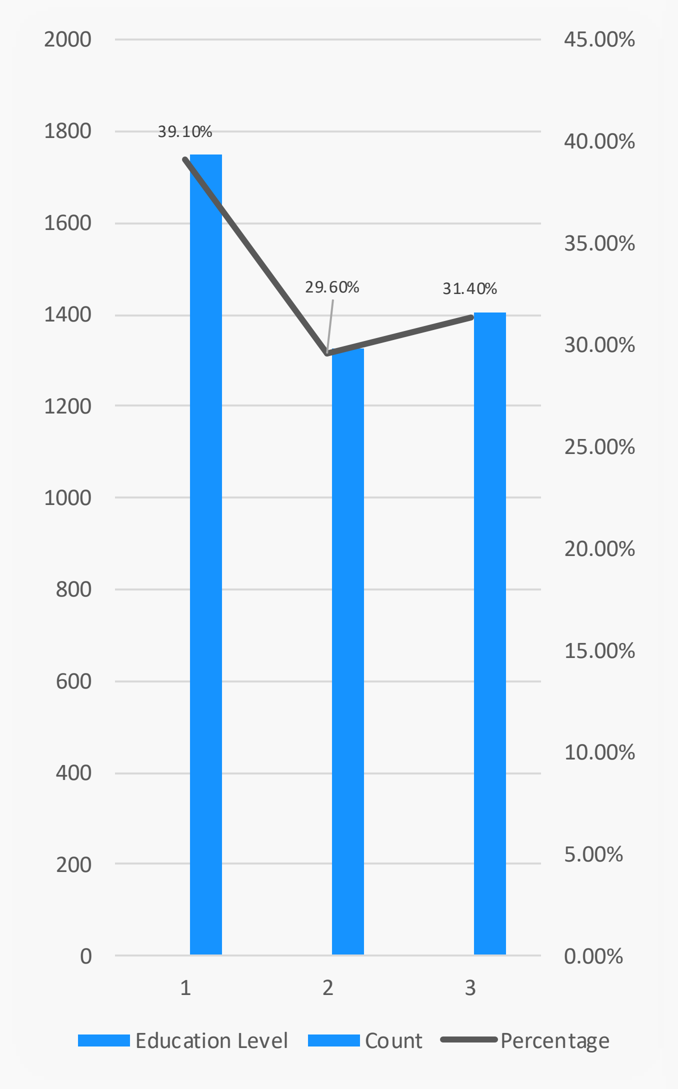
Income Level

Distribution of Customers by CCAvg Levels

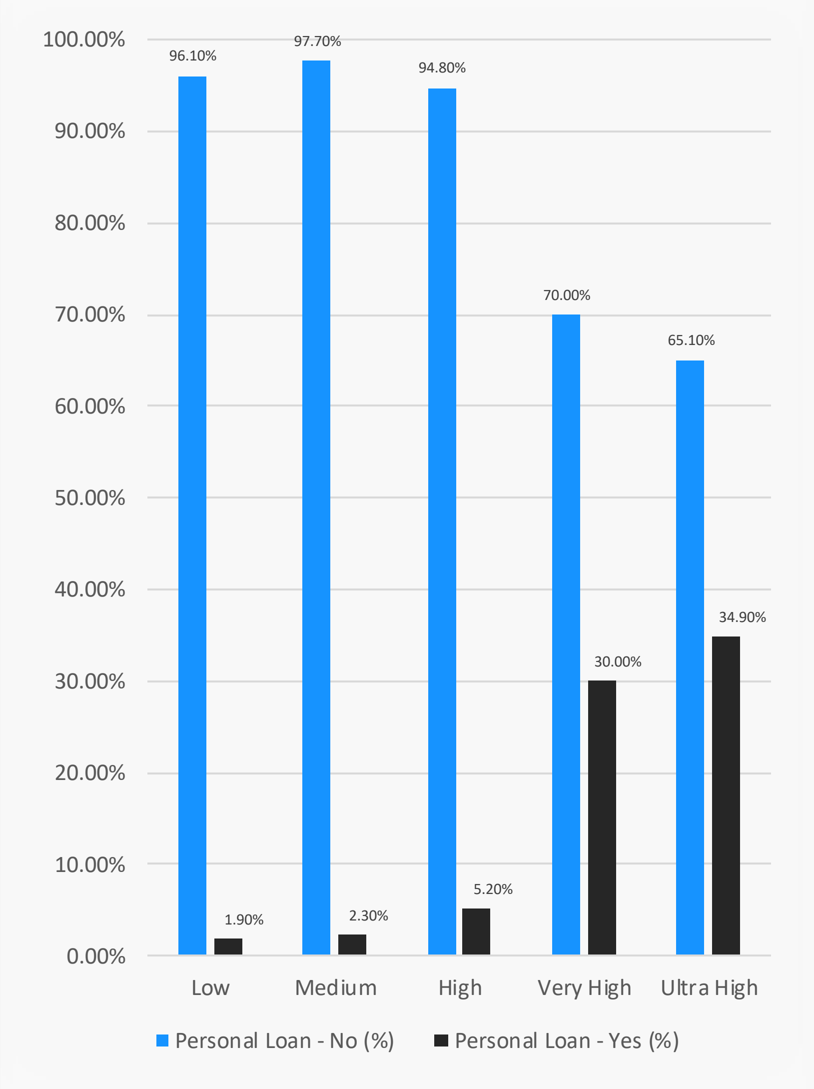


CCAvg Levels

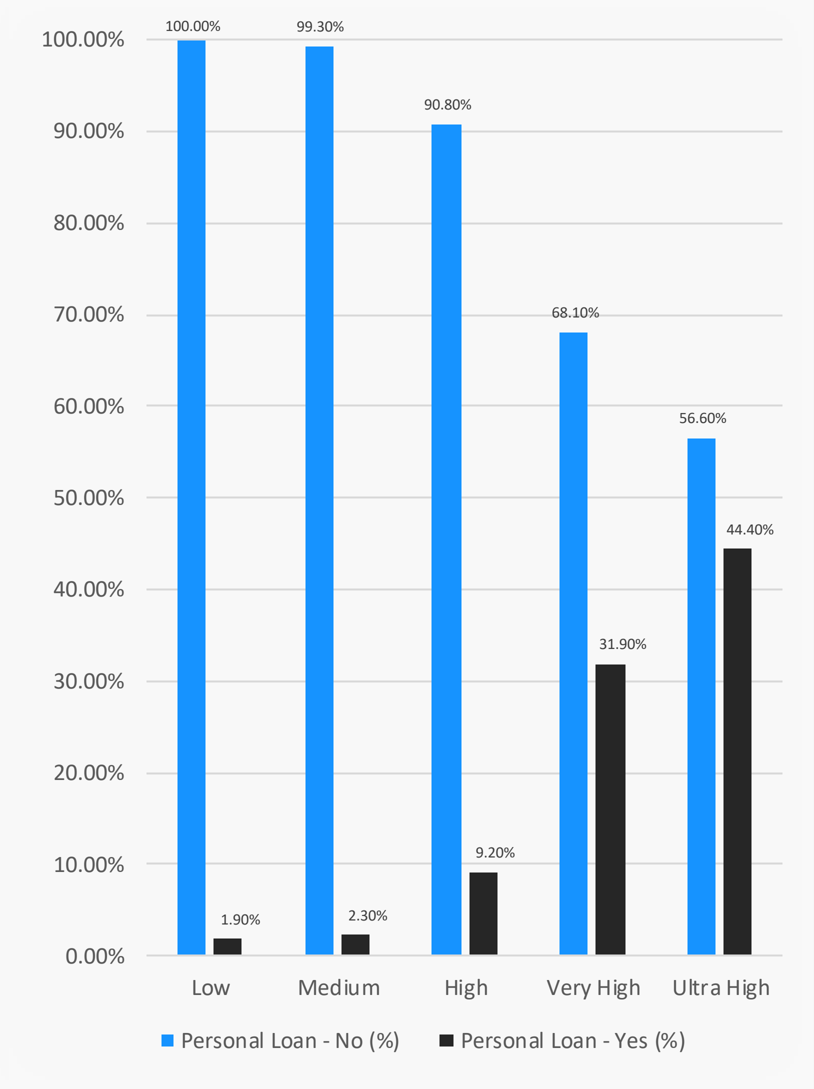
Distribution of Customers by Education



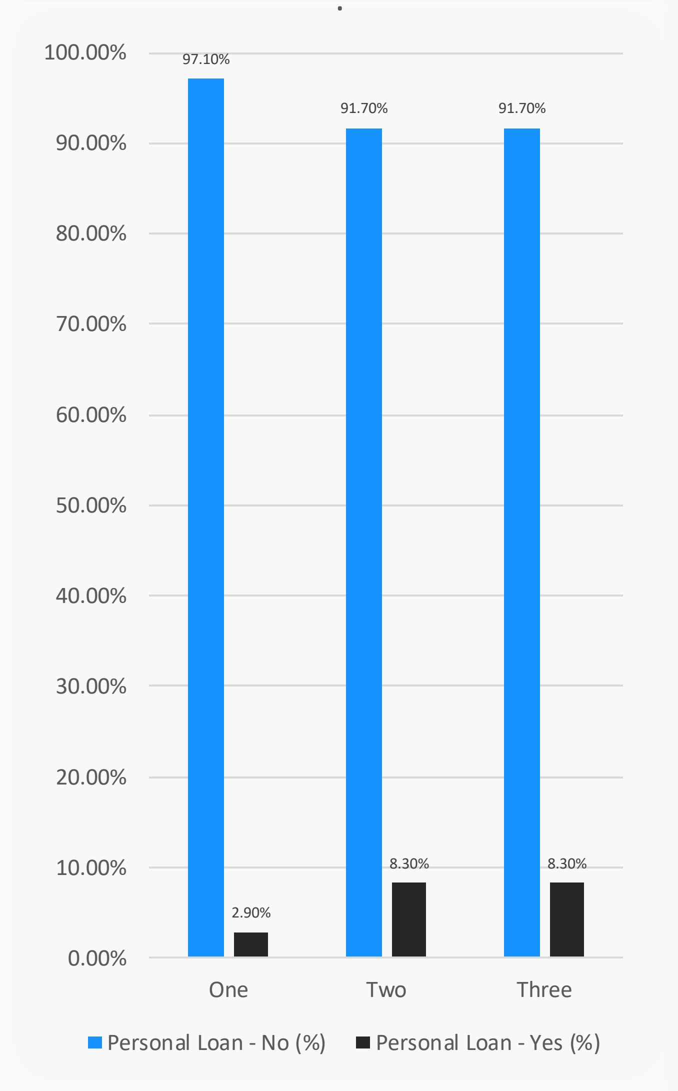
Distribution of CCAvg Levels by Personal Loan Status



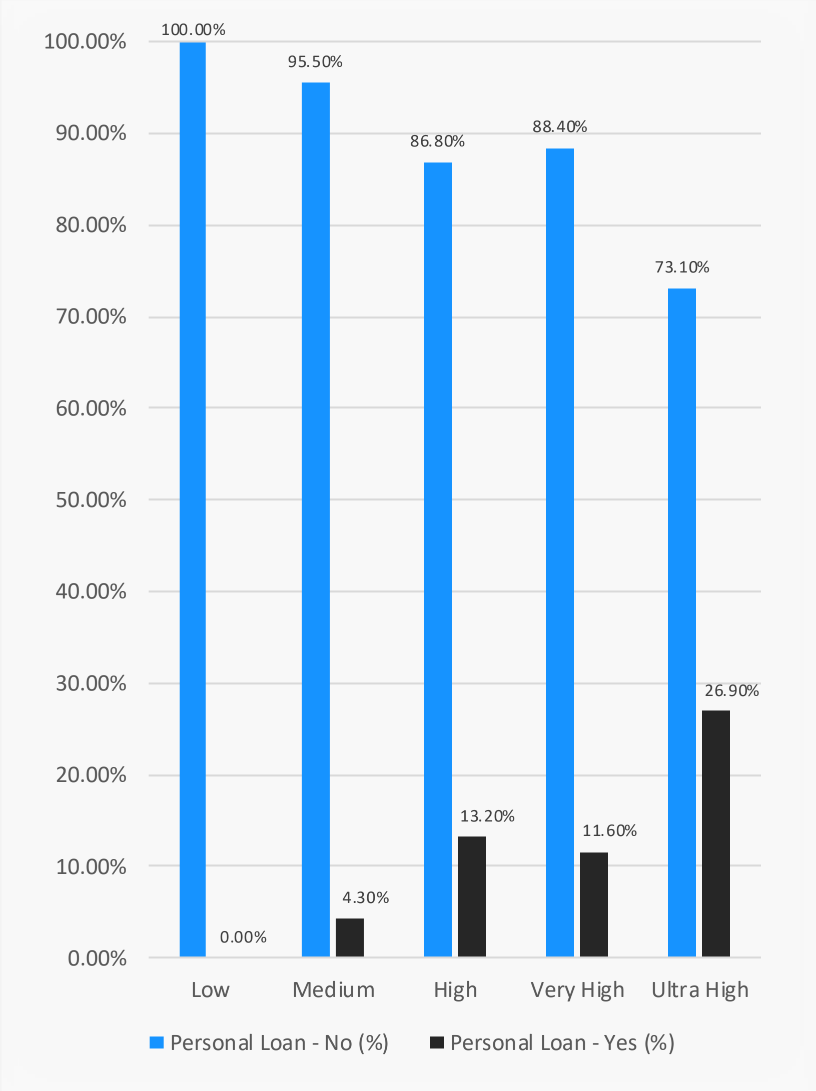
Distribution of Income Levels by Personal Loan Status



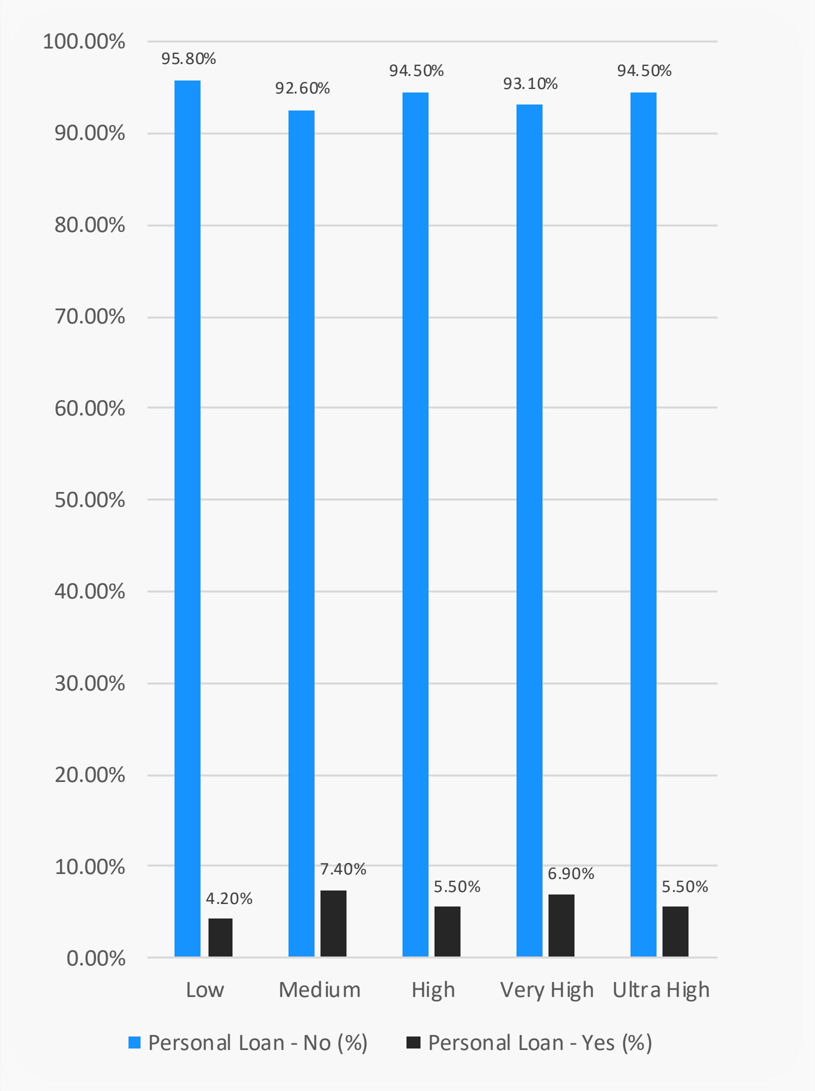
Distribution of Education by Personal Loan Status



Distribution Of Income per Capita by Personal Loan Status



Distribution of Credit Spend Ratio by Personal Loan Status



Insights

From Data Visualization

Income Dominance

Personal loan usage is significantly higher among individuals with higher income levels.

Credit Card Usage

Higher CCAvg levels strongly correlate with personal loan usage.

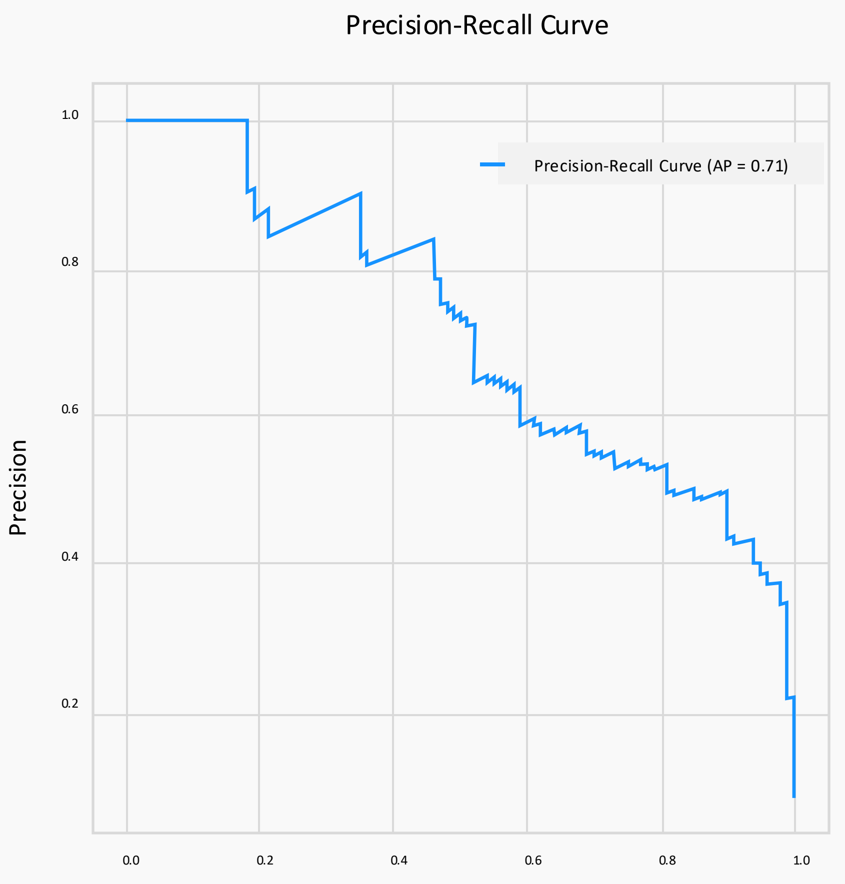
DATA MODELING

Logistic regression

Class	Precision	Recall	F1-Score	Support
0	0.99	0.89	0.94	1220
1	0.41	0.94	0.57	100
Accuracy			0.89	1320
Macro Avg	0.7	0.91	0.75	1320
Weighted Avg	0.95	0.89	0.91	1320

Positive class
High recall (0.94), low precision (0.41) → many false positives.

Negative class
Strong performance with precision (0.99) and recall (0.89).



Optimal decision threshold: 0.793

Class	Precision	Recall	F1-Score	Support
0	0.98	0.94	0.96	1220
1	0.53	0.81	0.64	100
Accuracy			0.93	1320
Macro Avg	0.76	0.88	0.8	1320
Weighted Avg	0.95	0.93	0.94	1320

Positive class
Greater balance with lower recall (0.81) and higher precision (0.53)

Negative class
Strong performance with precision (0.98) and recall (0.94).

Random forest

From Data Visualization

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1220
1	0.83	0.82	0.82	100
Accuracy			0.97	1320
Macro Avg	0.91	0.90	0.90	1320
Weighted Avg	0.97	0.97	0.97	1320

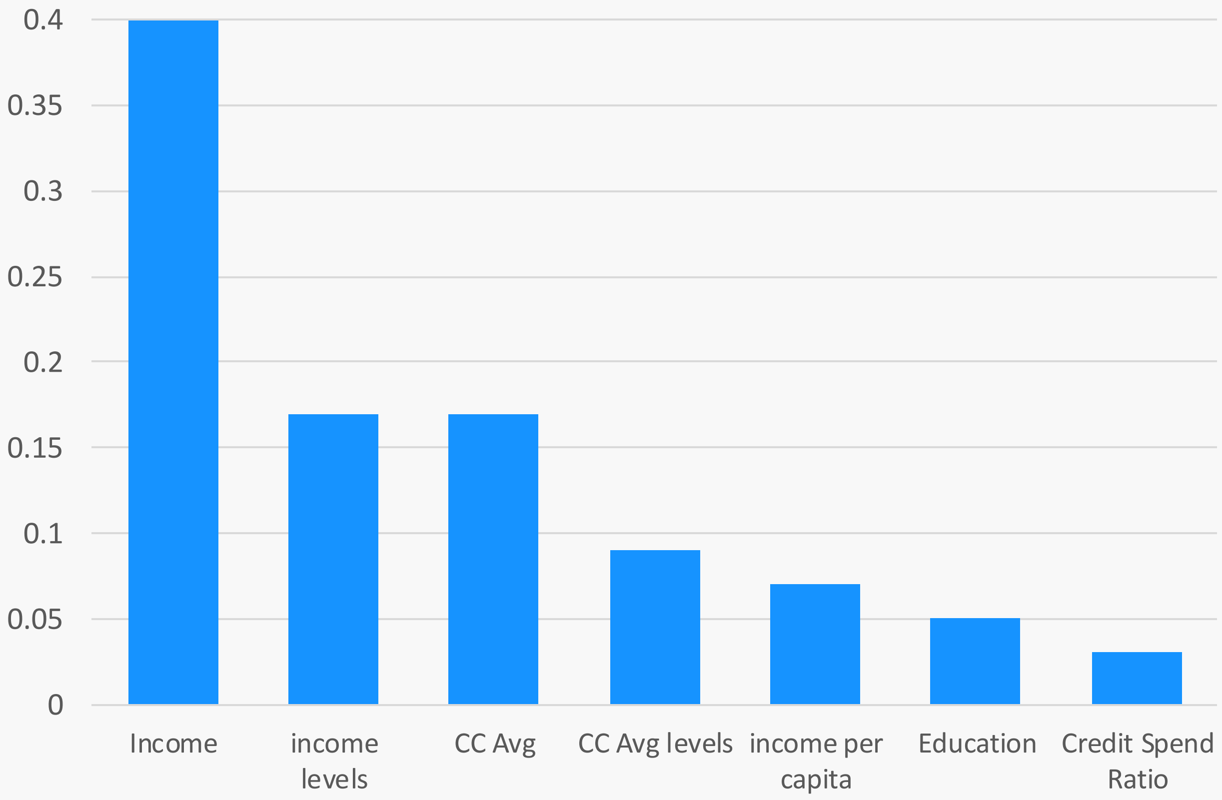
Positive class

Optimal balance with high recall (0.82) and high precision (0.83).

Negative class

Strong performance with precision (0.99) and recall (0.99).

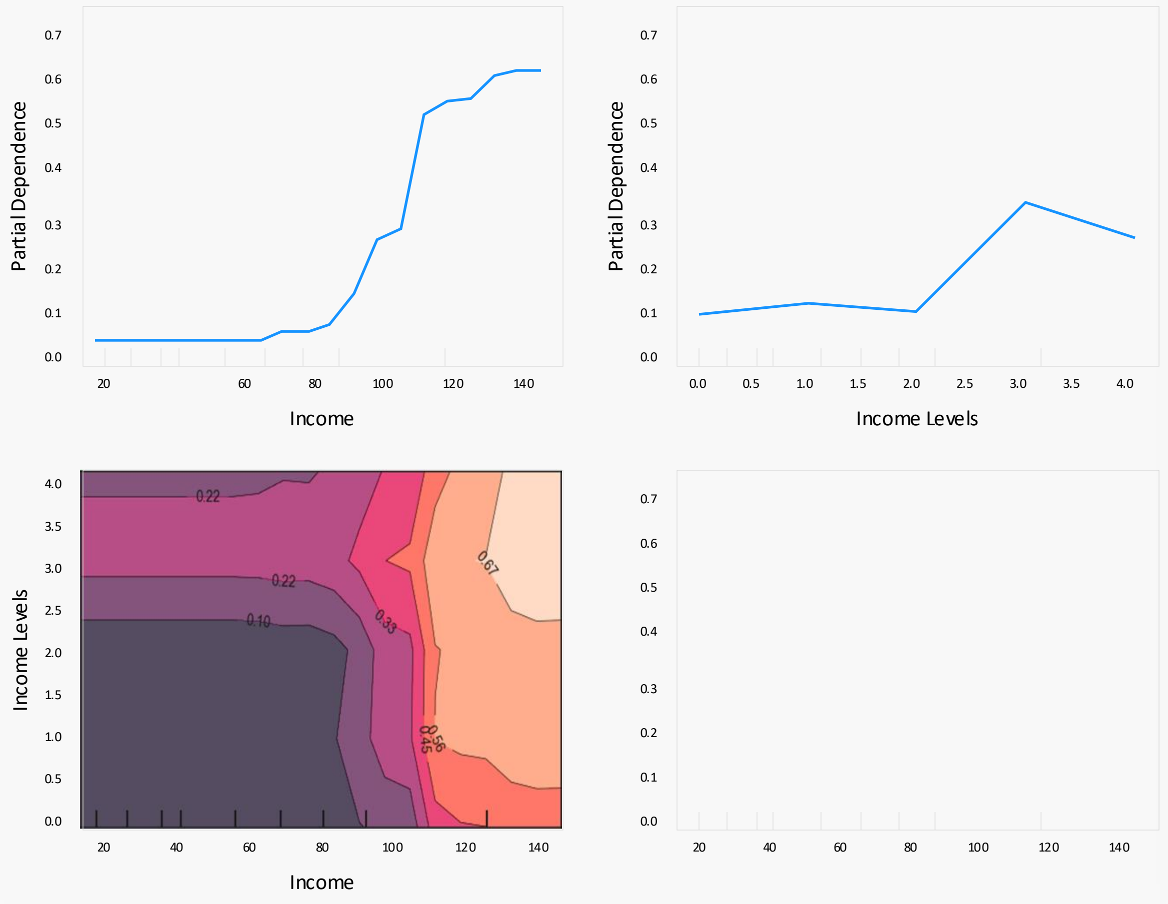
Feature Importance



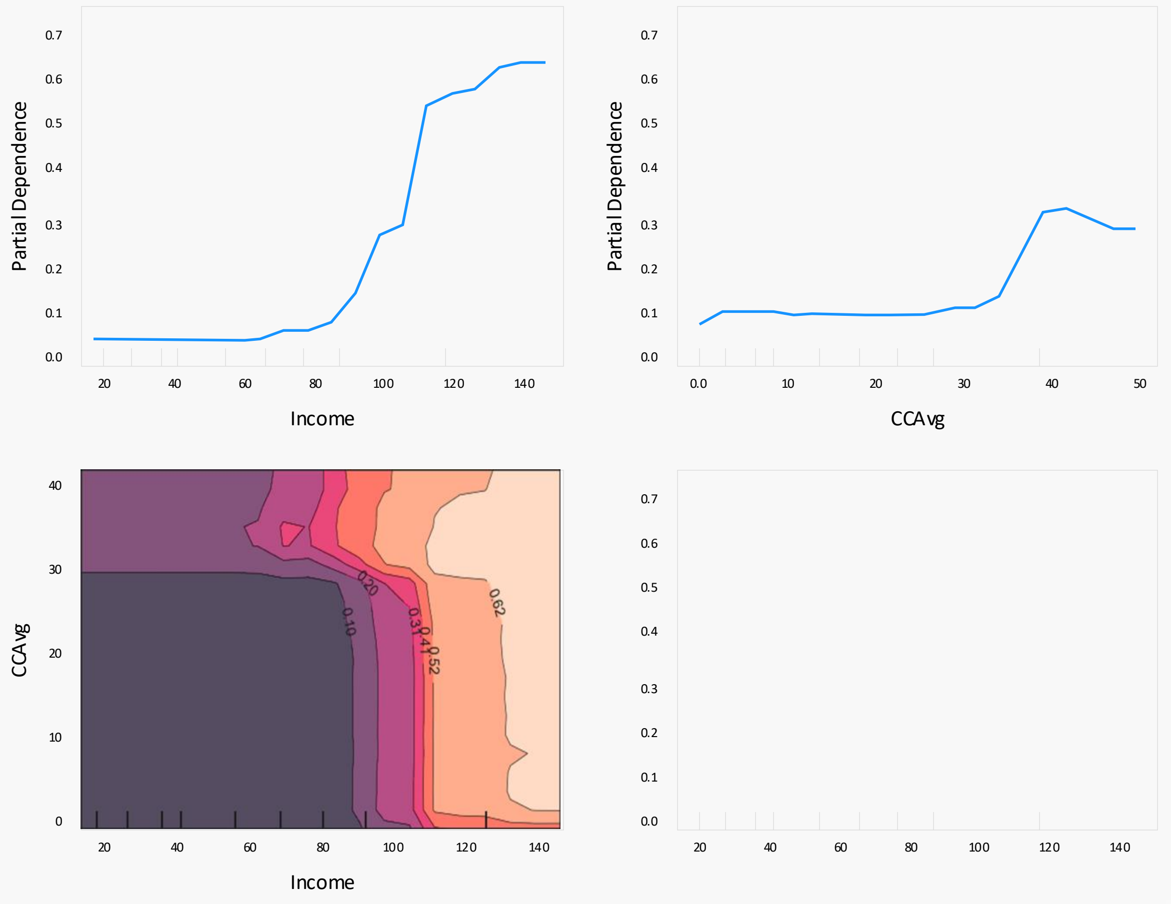
Income and credit card spending are the most influential predictors

Partial dependence plots

1-way and 2-way PDP of Income and Income Levels using Random Forest Classifier



1-way and 2-way PDP of Income and CCAvg using Random Forest Classifier



Partial dependence plots

Income

Loan acceptance probability increases significantly for higher-income individuals, particularly in income levels three and four.

High-income earners are far more likely to accept personal loans, highlighting a strong correlation between income, higher spending habits, and loan acceptance.

CCAvg (Credit Card Usage)

Loan acceptance probability remains stable at lower usage levels but rises sharply for individuals with high credit card usage (levels three and four).

This suggests a link between spending patterns and the likelihood of accepting a loan.

Education

Loan acceptance probability shows minor increases for individuals in higher education categories, likely due to the association between education, income, and financial engagement.

Conclusion

Income is the most significant driver of loan acceptance, with high-income individuals more likely to spend more and accept personal loans. This relationship underscores the importance of targeting high-income borrowers to increase conversion efficiency and enhance customer satisfaction.

Thank You

Charles Bryant