COMPREHENSIVE ANALYSIS OF THERA BANK LIABILITY CUSTOMERS

Charles Bryant

CONTEXT

This case is about a bank (Thera Bank) whose management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

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, and the customer response to the last campaign

DATASET ATTRIBUTES

1. ID: Customer ID
2. Age: Customer's age in completed years
3. Experience: Years of professional experience
4. Income: Annual income of the customer ($000)
5. ZIP Code: Home Address ZIP code
6. Family: Family size of the customer
7. CCAvg: Average spending on credit cards per month ($000)
8. Education Level
   1: Undergrad
   2: Graduate
   3: Advanced/Professional
9. Mortgage: Value of house mortgage if any ($000)
10. Personal Loan: Did this customer accept the personal loan offered in the last campaign?
11. Securities Account: Does the customer have a securities account with the bank?
12. CD Account: Does the customer have a certificate of deposit (CD) account with the bank?
13. Online: Does the customer use internet banking facilities?
14. Credit card: Does the customer use a credit card issued by the bank?

DATA ANALYSIS METHODOLOGIES

Feature Selection

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

Threshold Adjustment

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

Evaluation Metrics

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

Ensemble Learning

Leveraged ensemble learning methods to boost overall model performance.

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	CreditCard
0	1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
1	2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
2	3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
3	4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
4	5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1

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

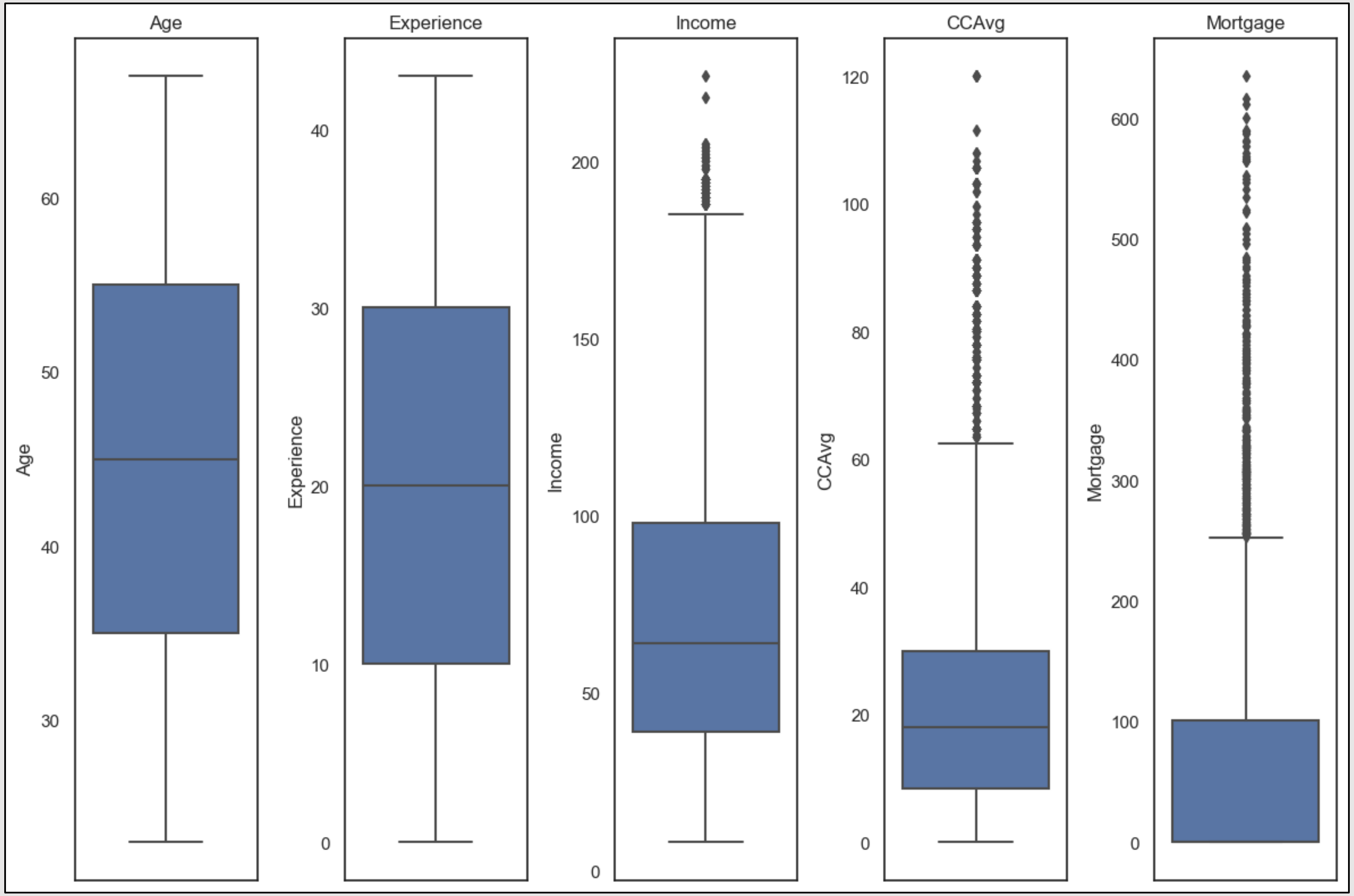
- Converted CCAvg to an annual scale to ensure consistency with income data.

DATA VIEW

AFTER INITIAL PROCESSING

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

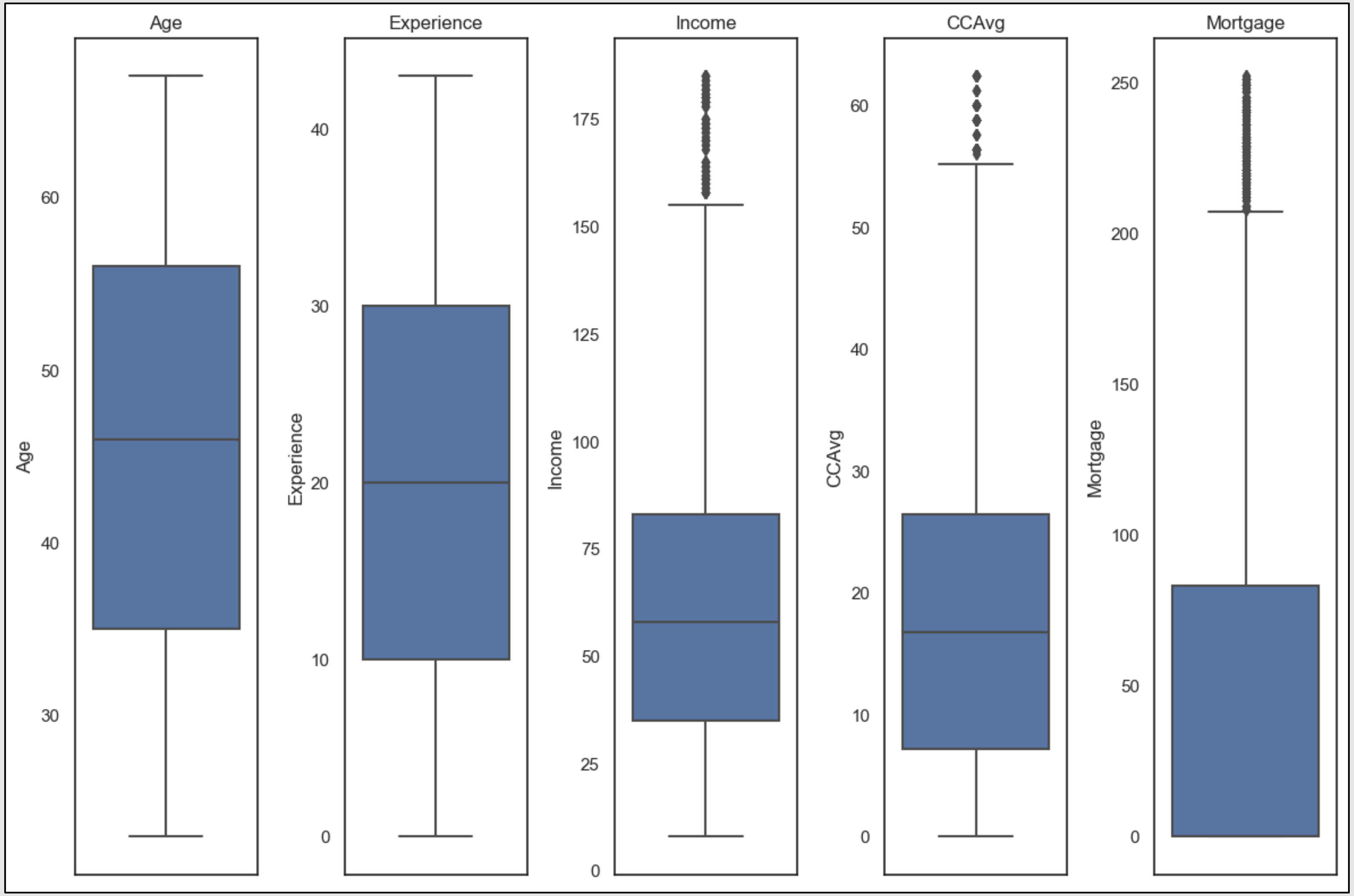
BOXPLOTS
BEFORE OUTLIER REMOVAL



BOXPLOTS

AFTER OUTLIER REMOVAL

Income, CCAvg, and Mortgage exhibit significant outliers. Using the IQR method, we removed these outliers and analyzed the summary statistics to evaluate their impact on the dataset. This approach allowed us to understand how outliers influence key metrics and ensure a more reliable foundation for modeling.



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

	Age	Experience	Income	CCAvg	Mortgage
count	5000.000000	5000.000000	5000.000000	5000.000000	5000.000000
mean	45.338400	20.134600	73.774200	23.255256	56.498800
std	11.463166	11.415189	46.033729	20.971908	101.713802
min	23.000000	0.000000	8.000000	0.000000	0.000000
25%	35.000000	10.000000	39.000000	8.400000	0.000000
50%	45.000000	20.000000	64.000000	18.000000	0.000000
75%	55.000000	30.000000	98.000000	30.000000	101.000000
max	67.000000	43.000000	224.000000	120.000000	635.000000

Summary without outliers

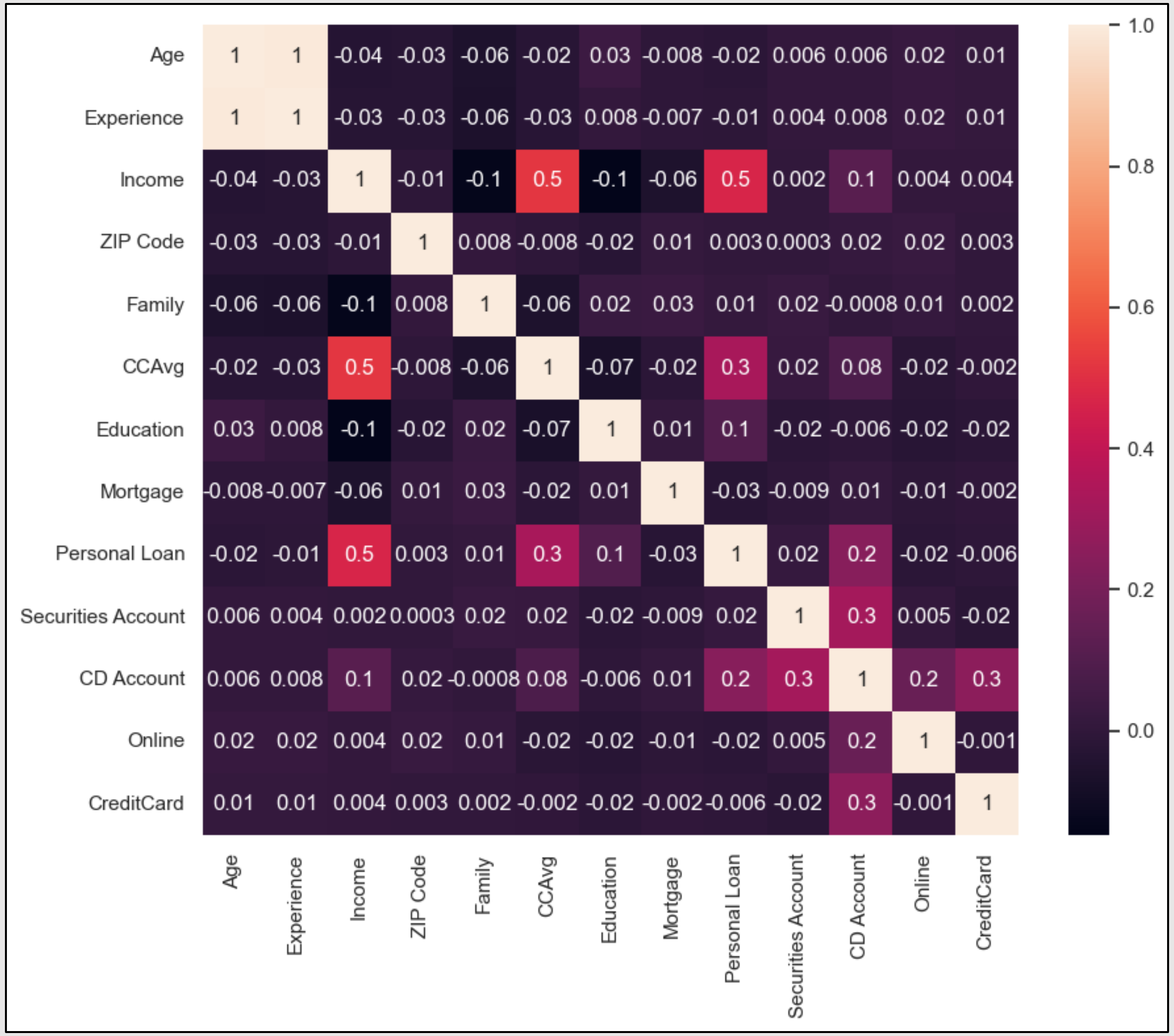
	Age	Experience	Income	CCAvg	Mortgage
count	4398.000000	4398.000000	4398.000000	4398.000000	4398.000000
mean	45.536608	20.309004	64.084584	18.613752	38.490678
std	11.490289	11.458770	38.024646	13.890412	68.108115
min	23.000000	0.000000	8.000000	0.000000	0.000000
25%	35.000000	10.000000	35.000000	7.200000	0.000000
50%	46.000000	20.000000	58.000000	16.800000	0.000000
75%	56.000000	30.000000	83.000000	26.400000	83.000000
max	67.000000	43.000000	185.000000	62.400000	252.000000

Percentage of outliers

8.8 %

CORRELATION HEAT MAP

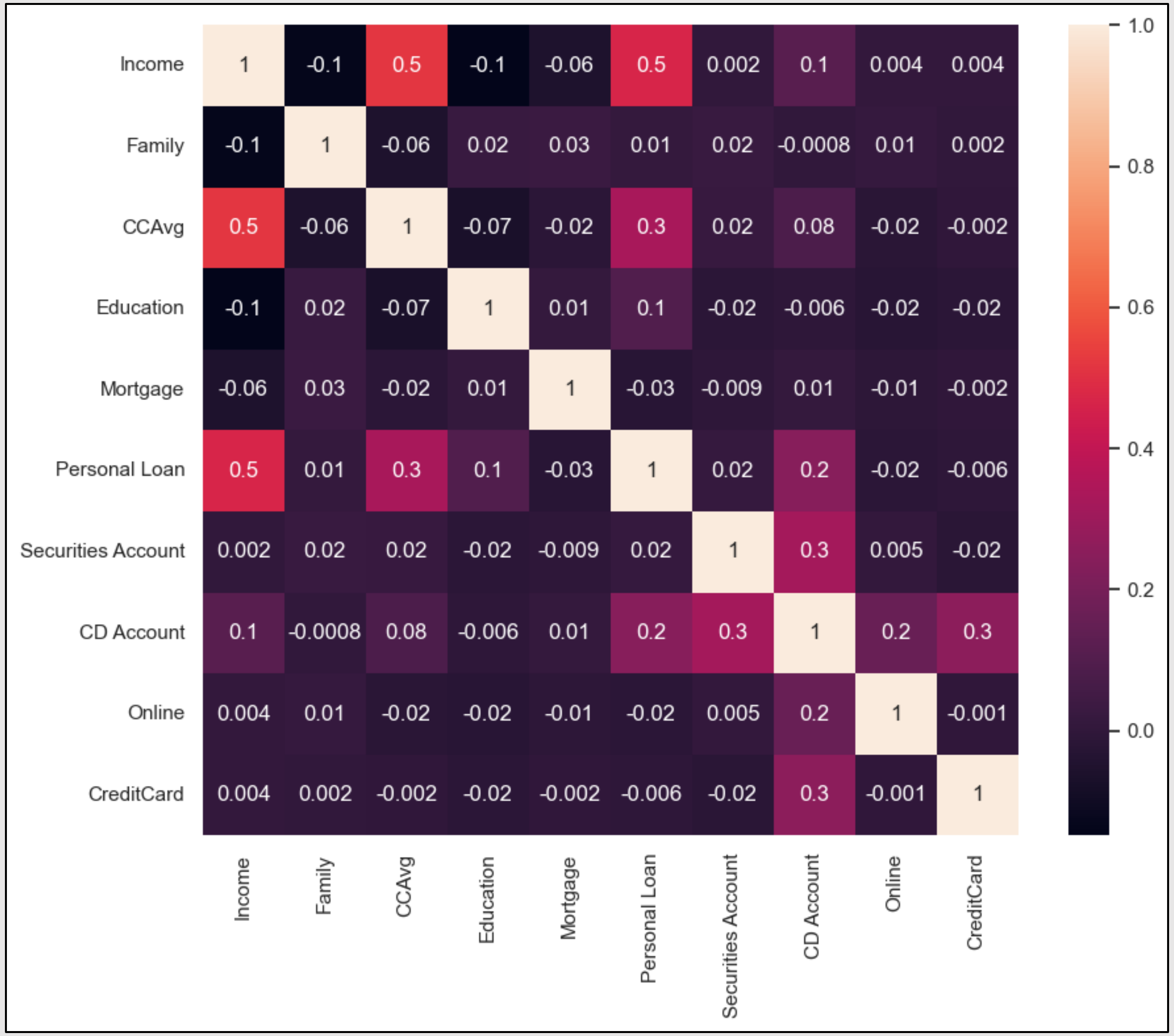
BEFORE REMOVAL OF INSIGNIFICANT FEATURES



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.



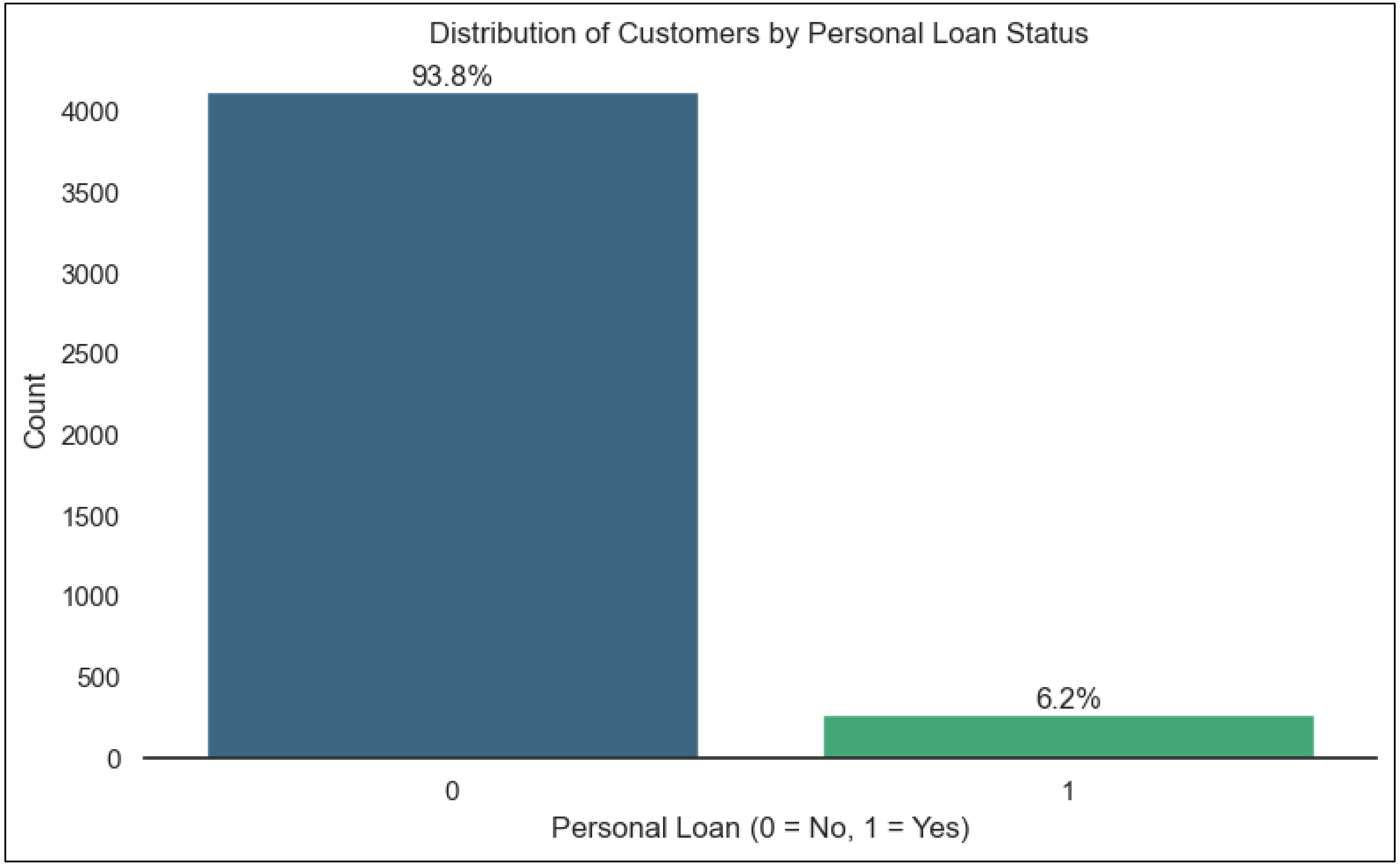
FINALIZE FEATURE SPACE

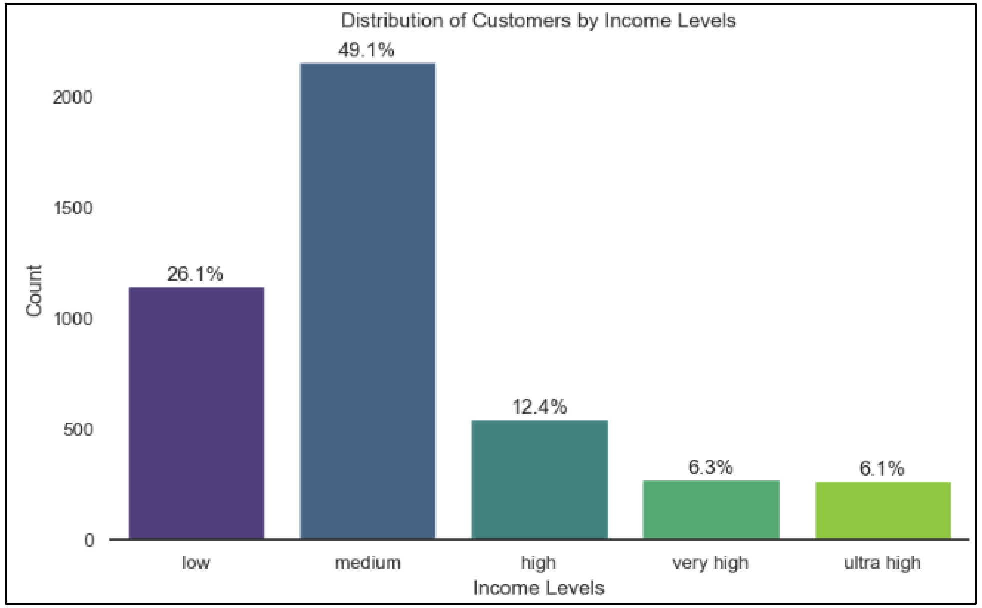
Income, CCAvg, and Education were identified as the most predictive features. Reducing the feature space improved the model's ability to distinguish between classes effectively.

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

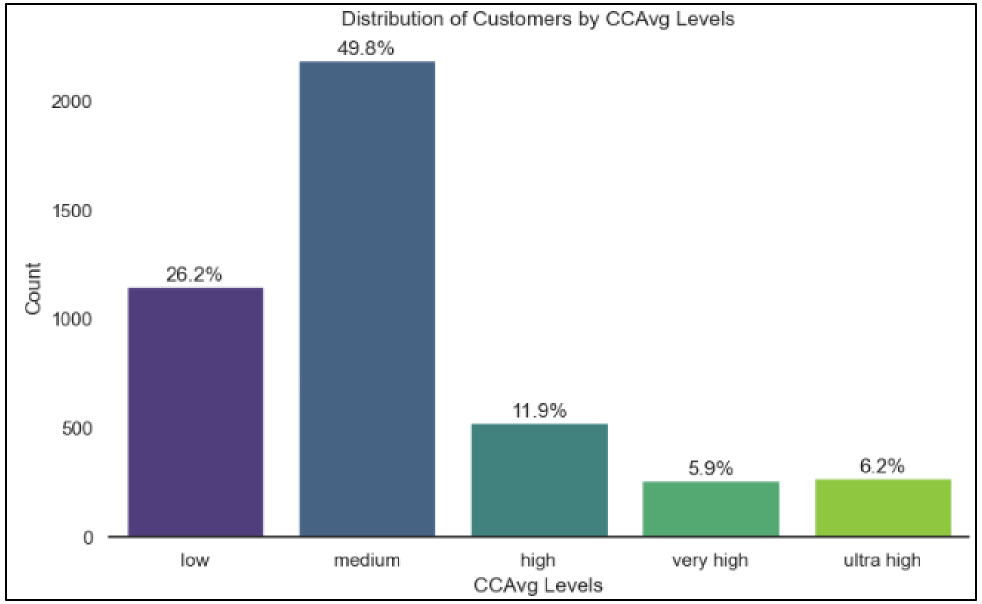
Output from Random Forest Feature Importance

DATA VISUALIZATION

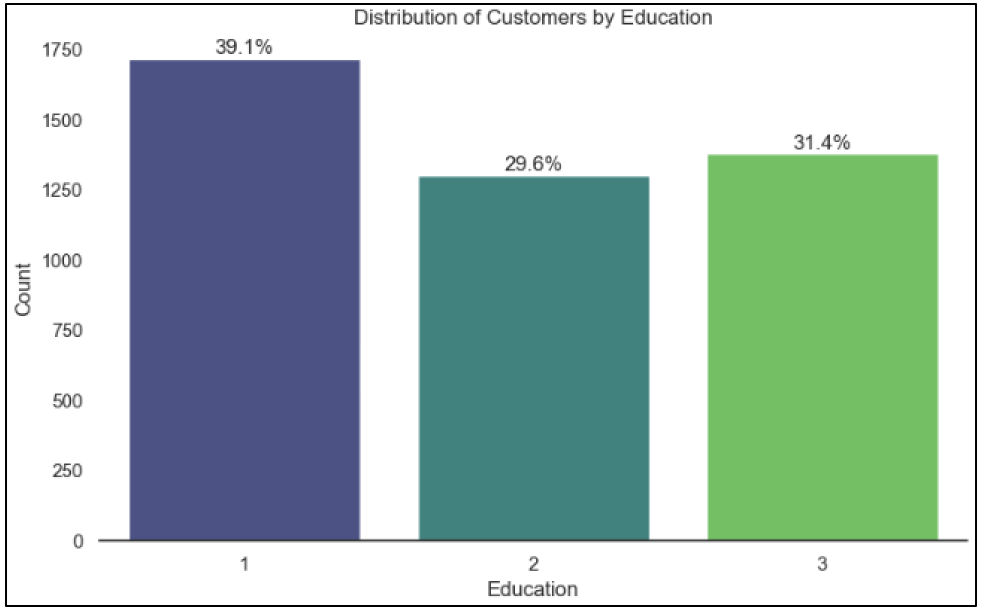

Distribution of Customers by Personal Loan Status
93.8%
6.2%
Count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
Personal Loan (0 = No, 1 = Yes)



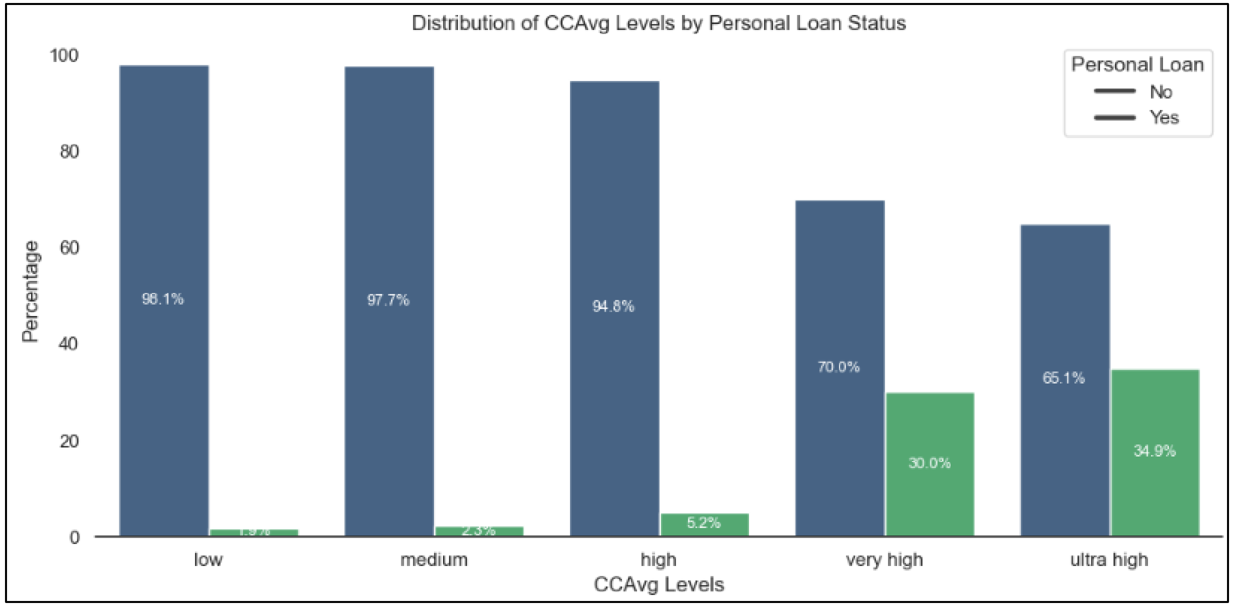
Distribution of Customers by Income Levels
49.1%
26.1%
12.4%
6.3%
6.1%
Count
2000
1500
1000
500
0
low
medium
high
very high
ultra high
Income Levels

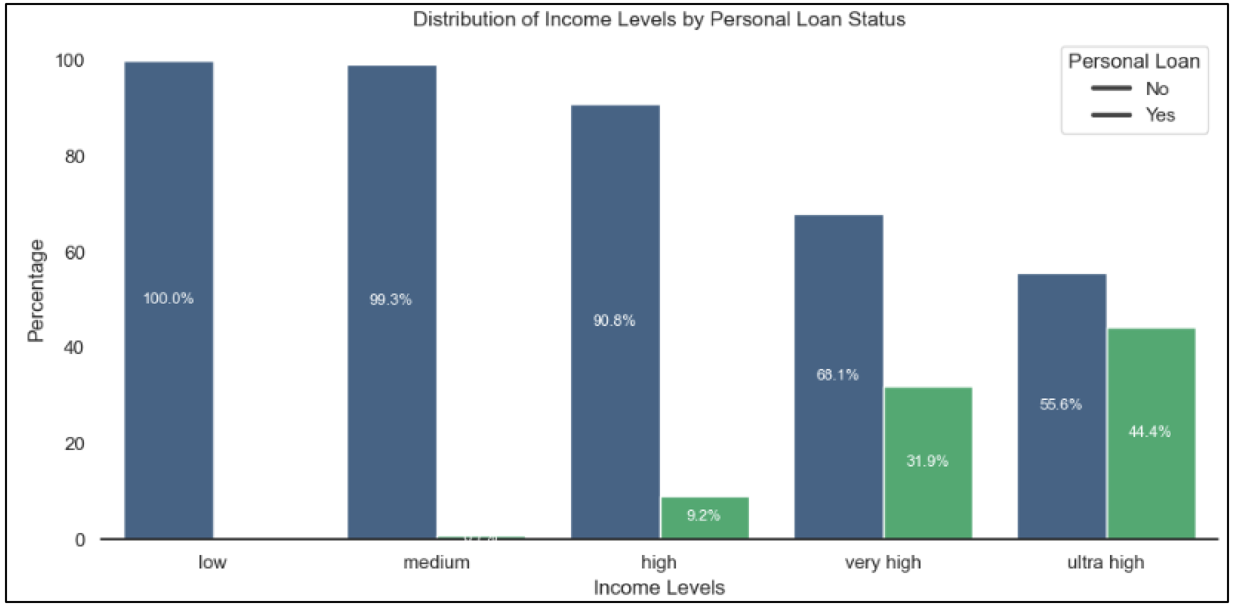


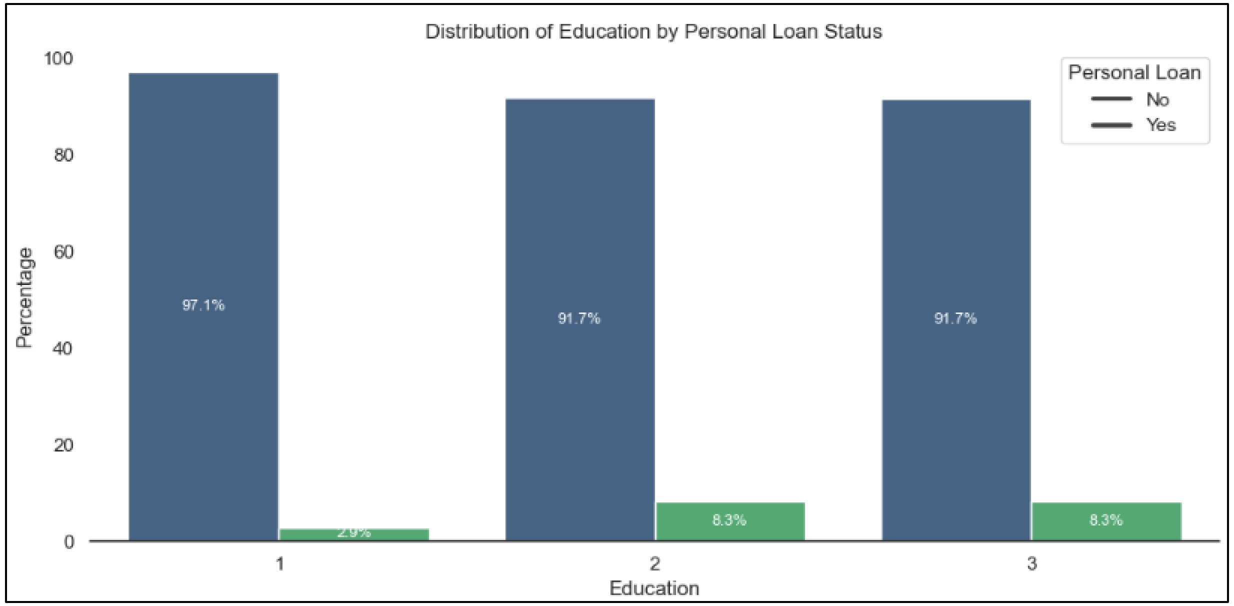
Distribution of Customers by CCAvg Levels
49.8%
26.2%
11.9%
5.9%
6.2%
Count
2000
1500
1000
500
0
low
medium
high
very high
ultra high
CCAvg Levels

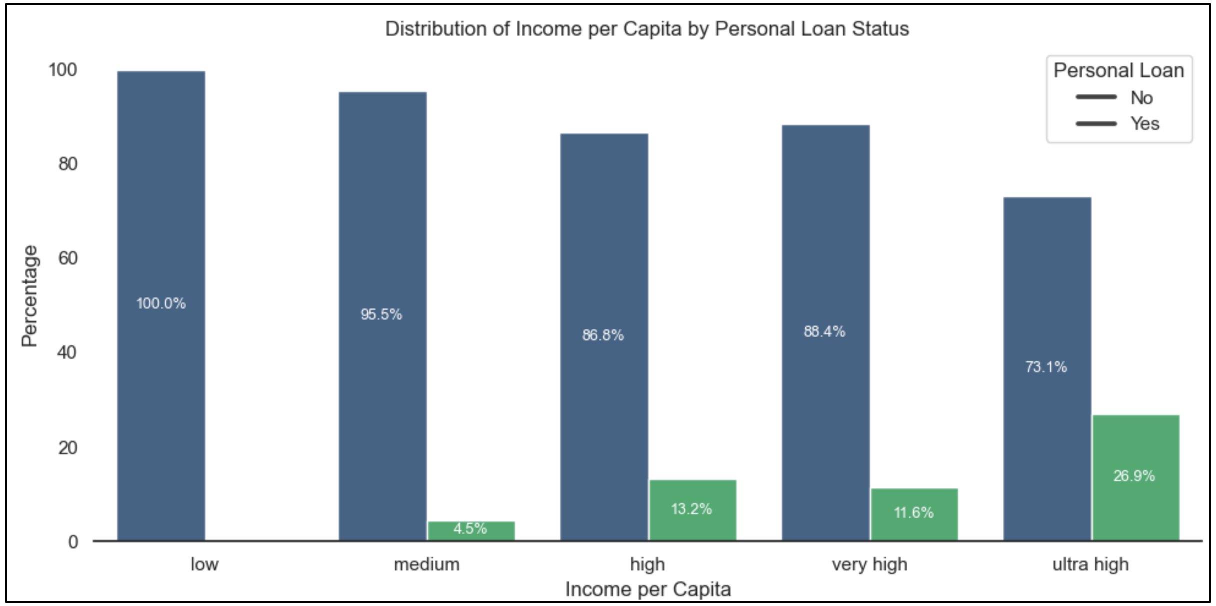


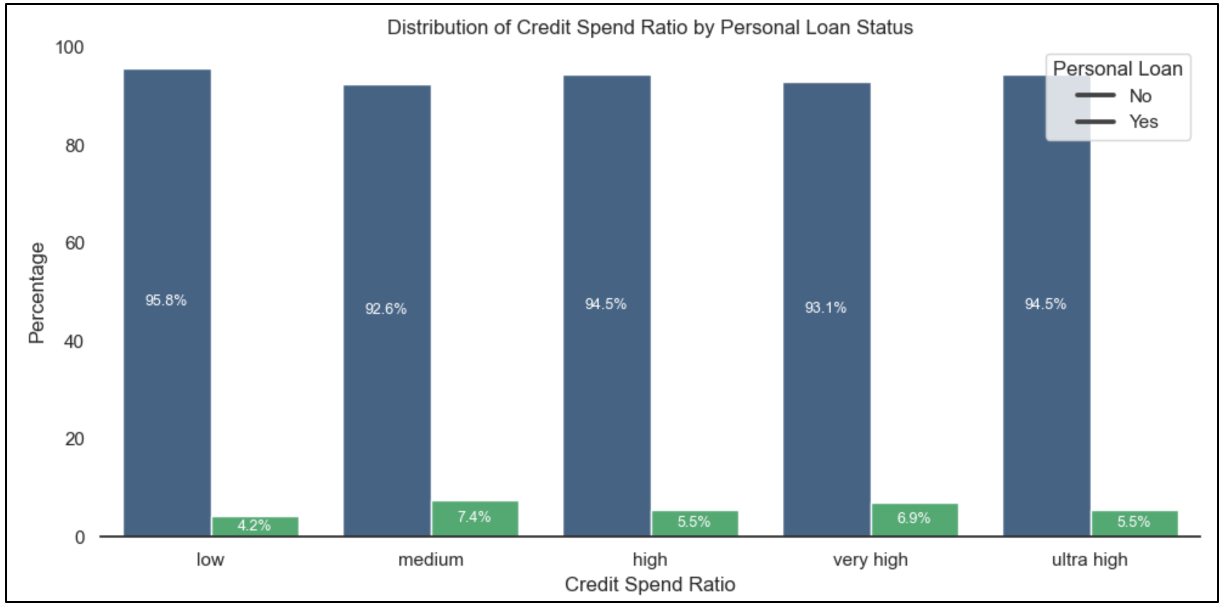
Distribution of Customers by Education
39.1%
29.6%
31.4%
Count
1750
1500
1250
1000
750
500
250
0
1
2
3
Education


Distribution of CCAvg Levels by Personal Loan Status
Personal Loan
No
Yes
Percentage
100
80
60
40
20
0
98.1%
97.7%
94.8%
5.2%
70.0%
30.0%
65.1%
34.9%
low
medium
high
very high
ultra high
CCAvg Levels


Distribution of Income Levels by Personal Loan Status
Personal Loan
No
Yes
Percentage
100
80
60
40
20
0
100.0%
99.3%
90.8%
9.2%
68.1%
31.9%
55.6%
44.4%
low
medium
high
very high
ultra high
Income Levels


Distribution of Education by Personal Loan Status
Personal Loan
No
Yes
Percentage
100
80
60
40
20
0
97.1%
91.7%
8.3%
91.7%
8.3%
1
2
3
Education


Distribution of Income per Capita by Personal Loan Status
Personal Loan
No
Yes
Percentage
100
80
60
40
20
0
100.0%
95.5%
4.5%
86.8%
13.2%
88.4%
11.6%
73.1%
26.9%
low
medium
high
very high
ultra high
Income per Capita


Distribution of Credit Spend Ratio by Personal Loan Status
Personal Loan
No
Yes
Percentage
100
80
60
40
20
0
95.8%
4.2%
92.6%
7.4%
94.5%
5.5%
93.1%
6.9%
94.5%
5.5%
low
medium
high
very high
ultra high
Credit Spend Ratio

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.

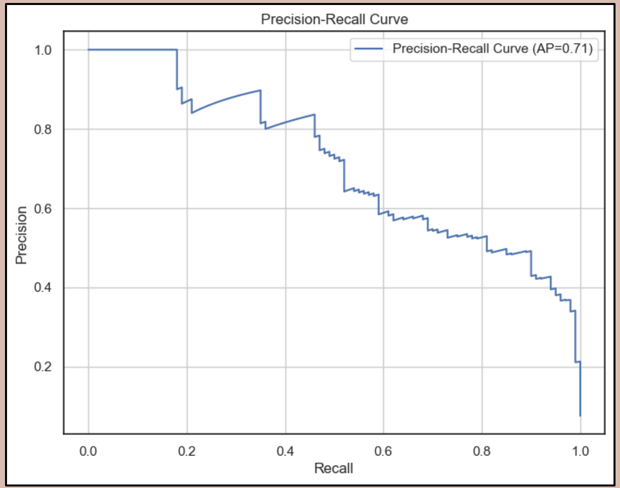
DATA MODELING

LOGISTIC REGRESSION

	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	0.70	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: .793

Adjusted Threshold

	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	0.80	1320
weighted avg	0.95	0.93	0.94	1320

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

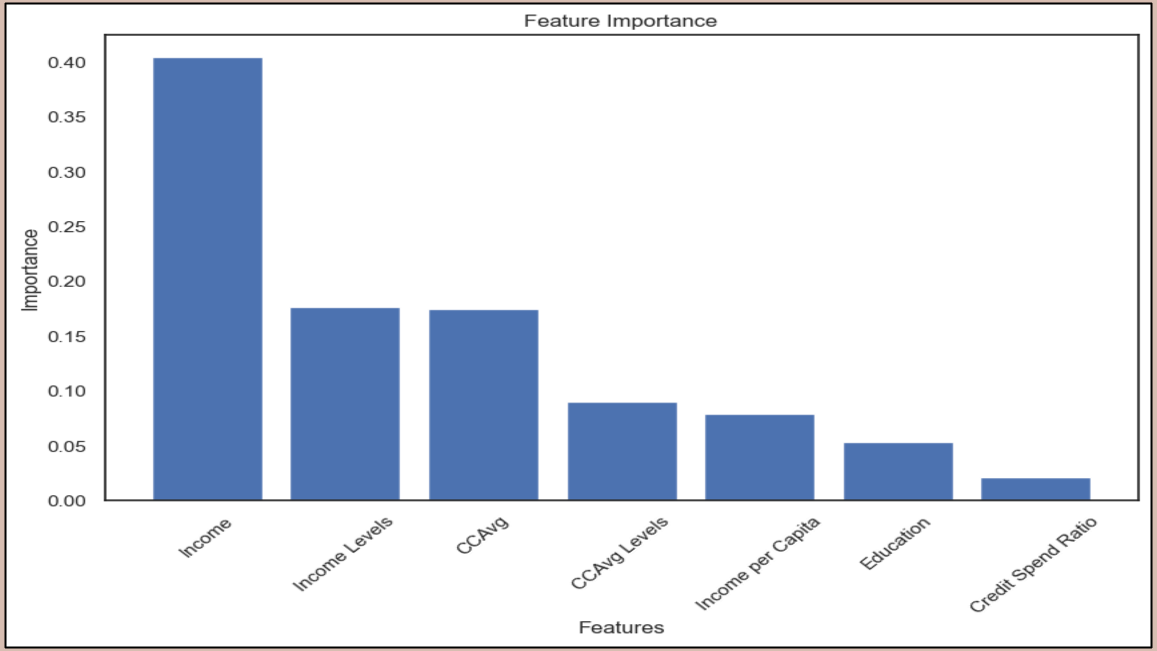
Negative class: Strong performance with precision (0.98) and recall (0.94).

RANDOM FOREST

	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

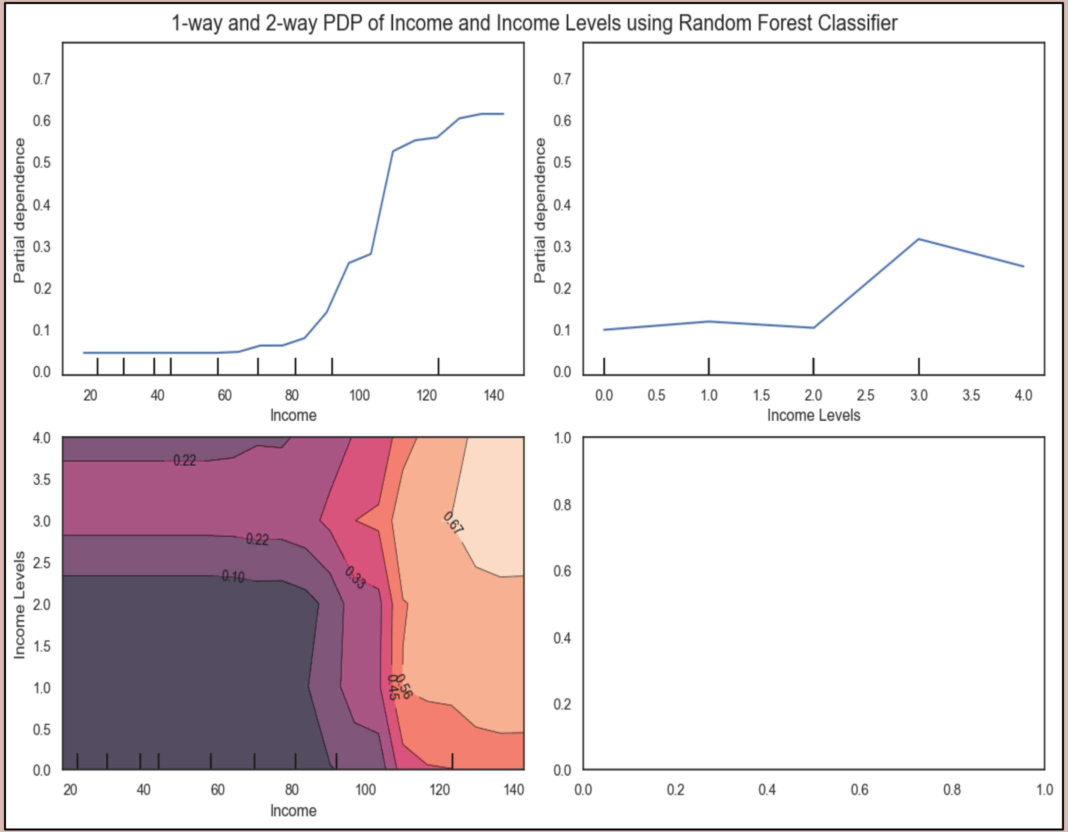
Positive class: Optimal balance with high recall (0.82) and high precision (0.83)

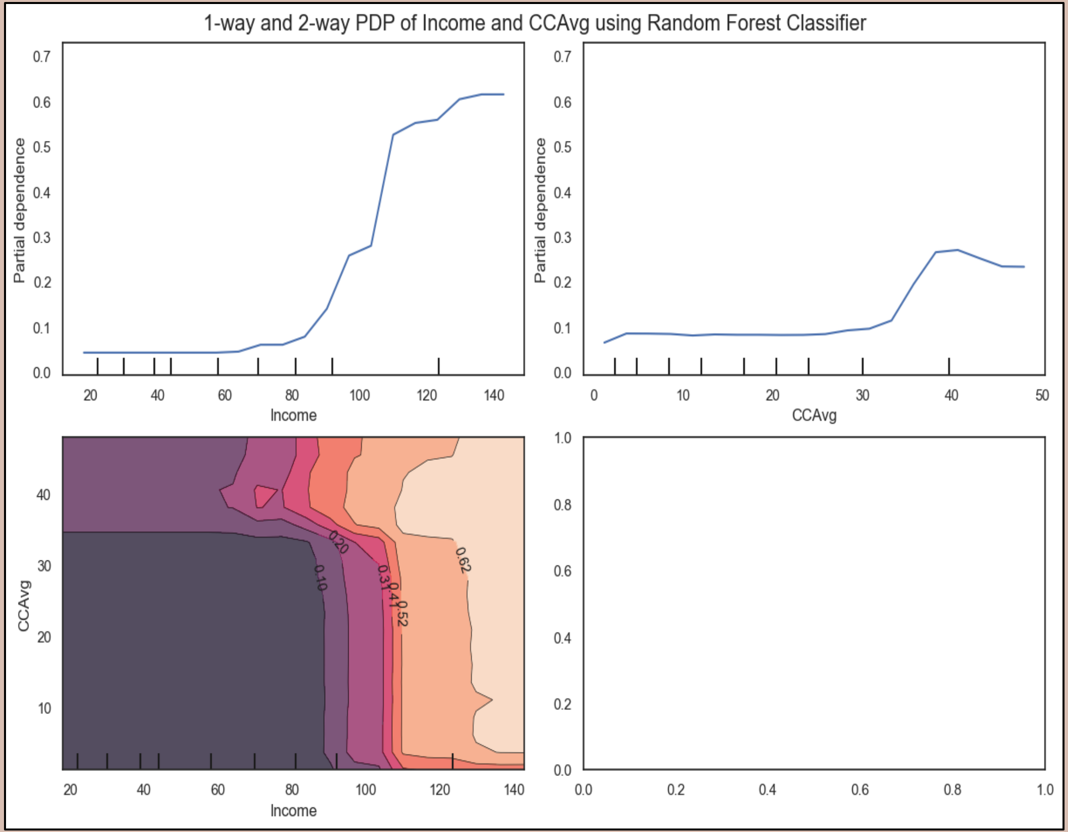
Negative class: Strong performance with precision (0.99) and recall (0.99).



Income and credit card spending are the most influential predictors

PARTIAL DEPENDENCE PLOTS





PARTIAL DEPENDENCE PLOTS

INSIGHTS

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