Evaluating and Mitigating Credit Default

An analysis of default probability to improve outcomes

CreditOne, LLC.

Understanding Credit

A recent increase in the number of defaults among our customers is prompting a reevaluation of our lending policies. This analysis will help CreditOne better understand the credit worthiness of our current and future customers and make recommendations for policy adjustments in order to ensure timely repayment of loans.

KEY QUESTIONS:

Which factors are driving customer defaults?

- What is the threshold for determining basic credit worthiness?
- Are the factors that are most important regarding default status gathered pre-approval or post-approval?
- How does CreditOne determine balance limits?

Our Process

PHASE I: Pull historical data and assess the complete data set

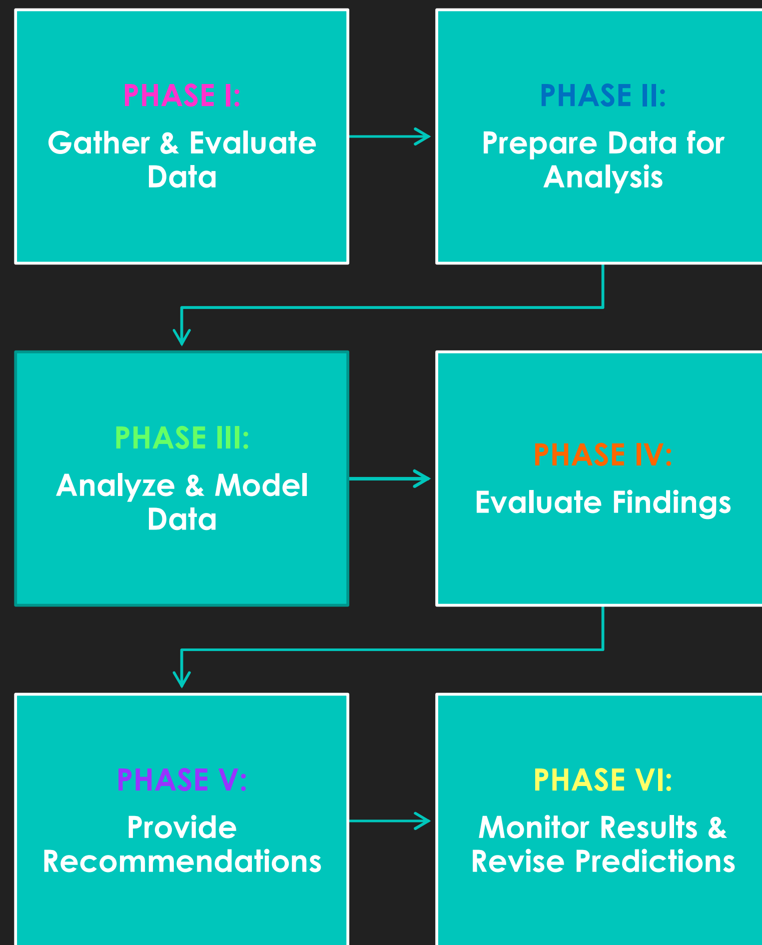
PHASE II: Remove incomplete data and create factor-based data groups

PHASE III: Identify patterns using visualizations, formulate hypotheses based upon patterns, and test hypotheses using predictive models

PHASE IV: Determine reliability of models and suitability of findings for our business

PHASE V: Recommend policy changes to affect the desired outcomes

PHASE VI: Utilize new data to determine if implemented policies have desired outcomes



ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	...	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4
1	20000	female	university	1	24	2	2	-1	-1	...	0	0	0	0	689	0	0
2	120000	female	university	2	26	-1	2	0	0	...	3272	3455	3261	0	1000	1000	1000
3	90000	female	university	2	34	0	0	0	0	...	14331	14948	15549	1518	1500	1000	1000
4	50000	female	university	1	37	0	0	0	0	...	28314	28959	29547	2000	2019	1200	1100
5	50000	male	university	1	57	-1	0	-1	0	...	20940	19146	19131	2000	36681	10000	9000
6	50000	male	graduate school	2	37	0	0	0	0	...	19394	19619	20024	2500	1815	657	1000
7	500000	male	graduate school	2	29	0	0	0	0	...	542653	483003	473944	55000	40000	38000	20239

Data Harvesting & Initial Evaluation

Where does our data set come from?

- Historical customer data

What does the data set "look like"?

- Over 30k customer records with demographic and payment information

Preparing Our Data

- Identify and remove null values
- Identify and remove duplicate values
- Simplify headings
- Reclassify data types

Analysis with Visualizations

Why analyze?

- Calculate basic statistics
- Simplify feature data by grouping:
  - Age groups
  - Balance limit groups
- Identify patterns in the data
- Assist with selecting features for predictive modeling

Customer Demographics

Sex

- Male: 11874
- Female: 18091

Age Range

- 20-29: 9603
- 30-39: 11226
- 40-49: 6456
- 50-59: 2341
- 60-69: 314
- 70-79: 25

Education

- High School: 4915
- University: 14019
- Graduate School: 10563
- Other: 468

Marriage

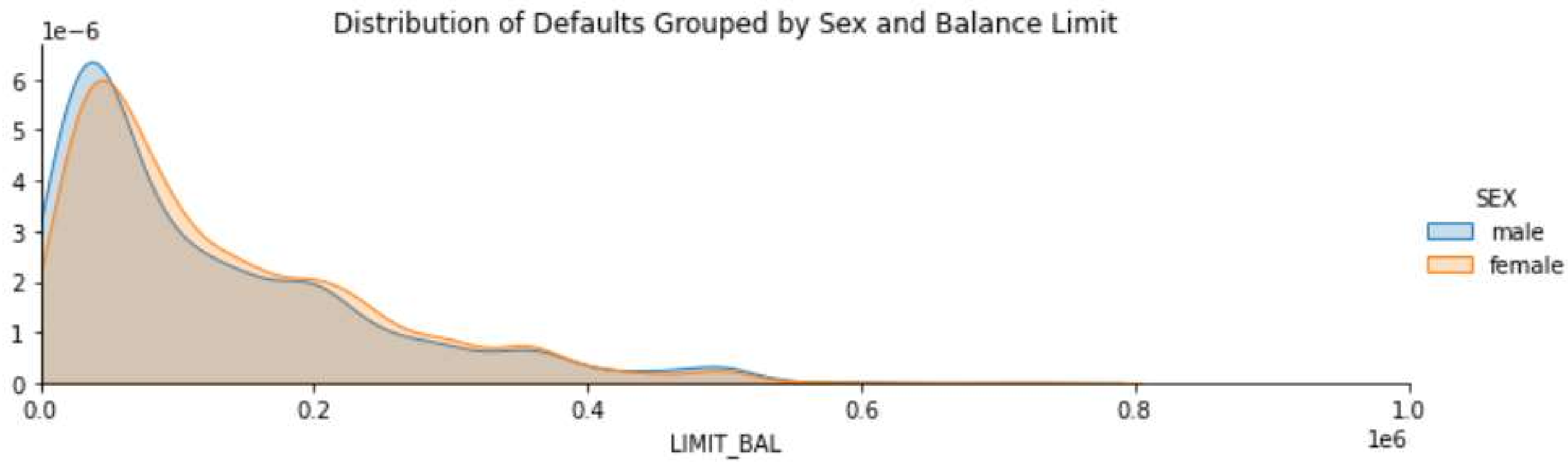
- Single: 15945
- Married: 13643
- Divorced: 323
- Other: 54

Default Status

- Not defaulted: 23335
- Defaulted: 6630

Balance Limit

- Under $100k: 11443
- $100k-$200k: 7390
- $200k-$300k: 6024
- $300k-$400k: 3034
- $400k-$500k: 1147
- Over $500k: 927

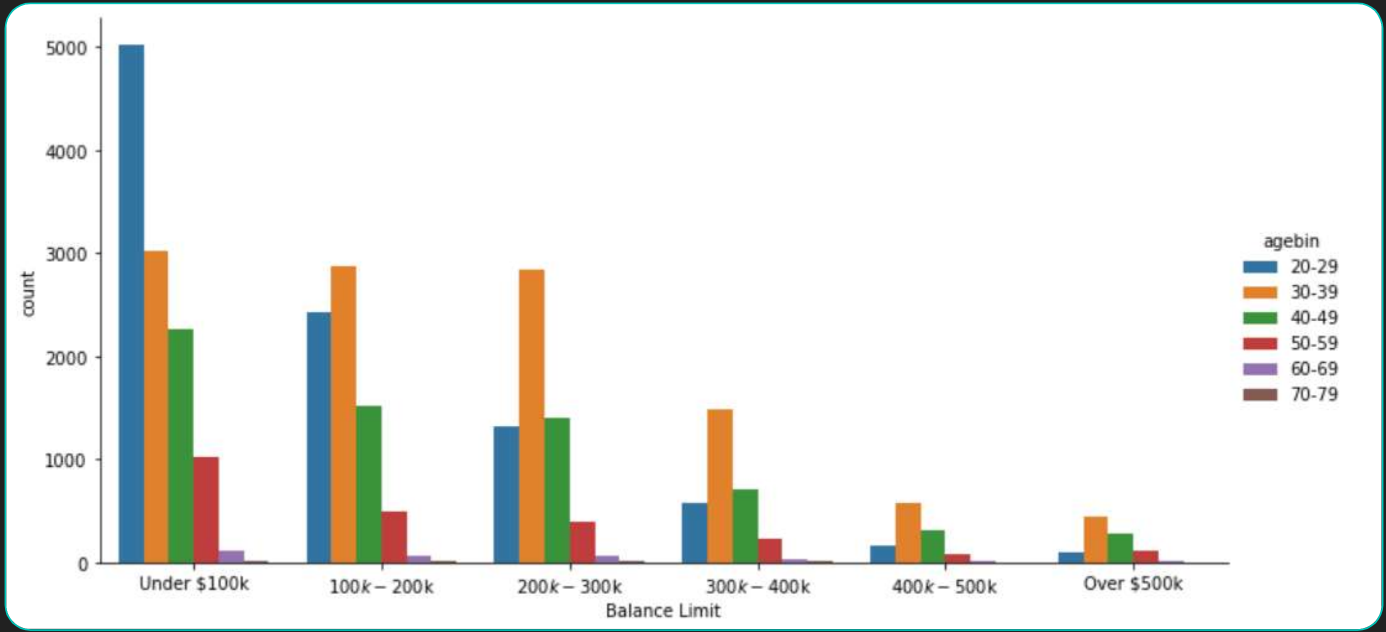


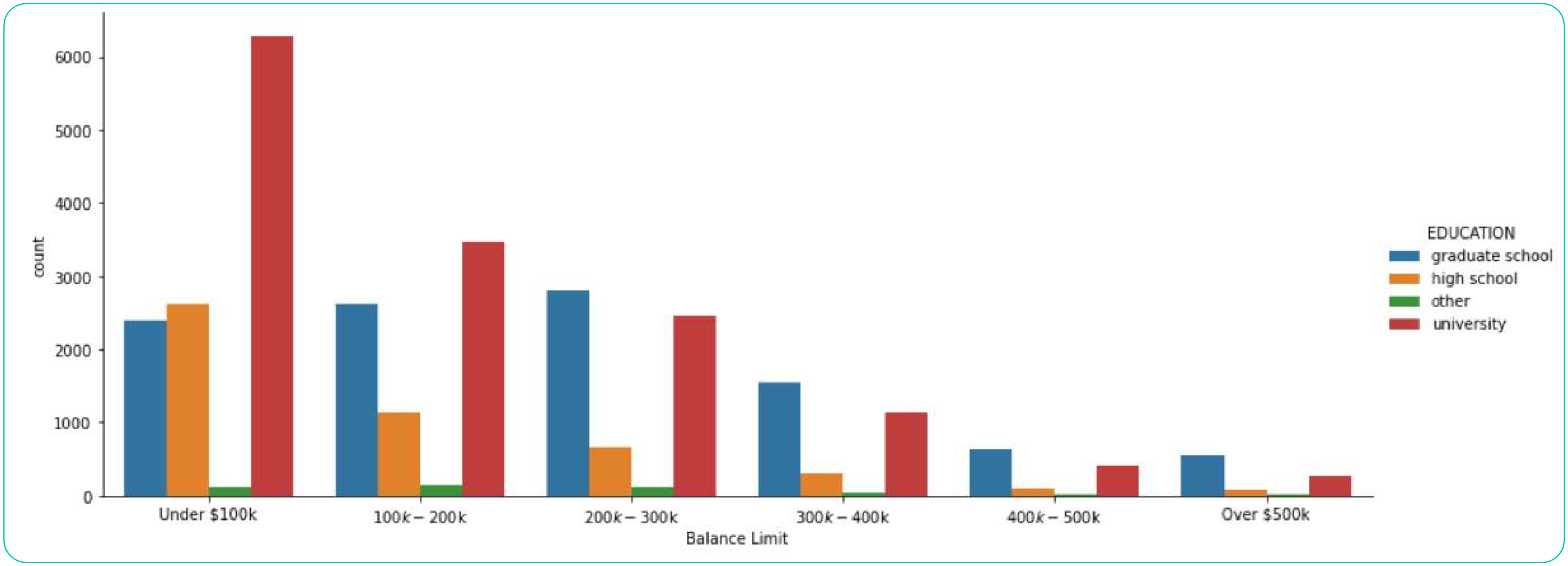
Defaulted Customers by Sex and Balance Limit

- Roughly similar distribution of defaults between males and females
- Note: women account for 60% of customers, men 40%

Defaulted Customers by Age and Balance Limit

- Highest number of defaults for customers in 20-29 age group with lower balance limits
- Similar number of defaults for 30-39 age group up to $300k
- Number of defaults decrease with age and increased balance limit



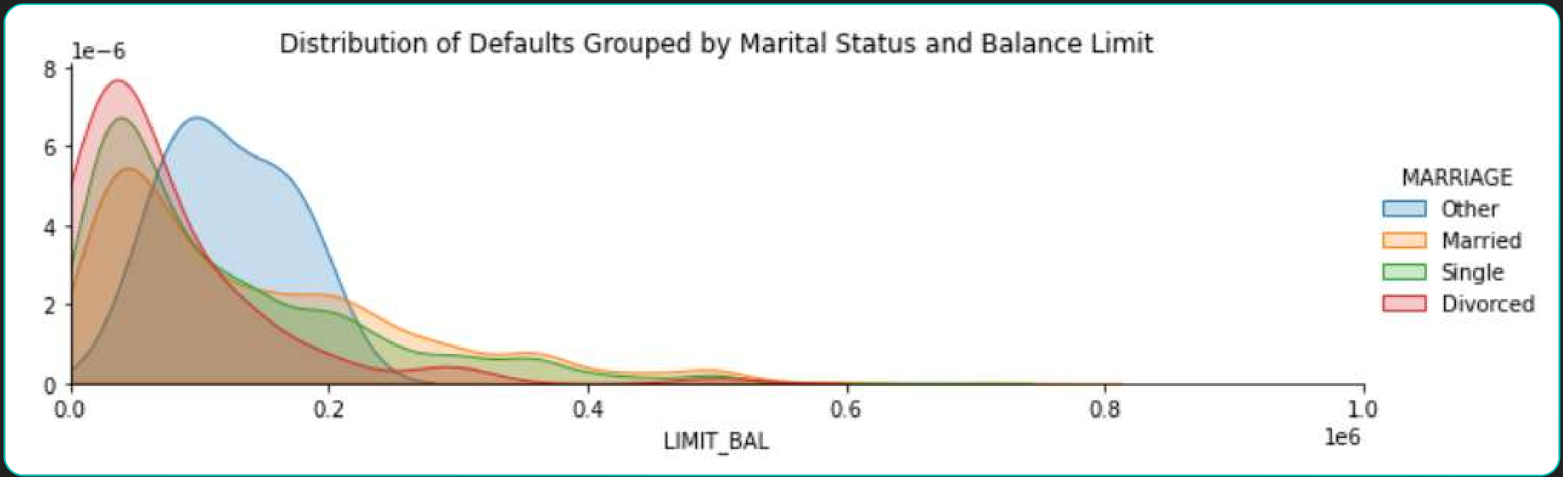


Defaulted Customers by Education and Balance Limit

- University graduates default more than customers with other educational achievement levels, but also account for half of the sample

Defaulted Customers by Marital Status and Balance Limit

- Married customers have the lowest distribution of defaults



Reflections

- Is there an acceptable rate of default? Current rate at 22%
- Men default at higher rate than women
- Defaults tend to decrease with age and higher balance limit; however, this is proportionate to our overall customer data

Feature Selection:

Correlation Analysis

	DEFAULT
DEFAULT	1.000000
PAY_0	0.324964
PAY_2	0.263656
PAY_3	0.235230
PAY_4	0.216551
PAY_5	0.204059
PAY_6	0.186740
AGE	0.013619
BILL_AMT6	-0.005469
BILL_AMT4	-0.010259
BILL_AMT2	-0.014302
MARRIAGE	-0.024019
PAY_AMT6	-0.053250
PAY_AMT5	-0.055194
PAY_AMT3	-0.056319
PAY_AMT4	-0.056898
PAY_AMT2	-0.058643
PAY_AMT1	-0.073015
LIMIT_BAL	-0.153871

- Values at left represent the correlation of each feature to default status
- Repayment status features (PAY_0…) have highest correlation to defaulted status
- PAY_0 is the most recent billing cycle and PAY_6 is six months prior to the most recent billing cycle
- Some categorical values excluded from analysis

Modeling

How do we select the best model?

- Cross-valuation scores

```
Decision Tree accuracy is 0.7238908912917724
Random Forest accuracy is 0.8132425577359498
Gradient Boosting accuracy is 0.8209406843768078
Support Vector accuracy is 0.7798246785030926
```

Gradient Boosting Accuracy

	precision	recall	f1-score	support
0	0.84	0.95	0.89	6996
1	0.66	0.37	0.48	1994
accuracy			0.82	8990
macro avg	0.75	0.66	0.68	8990
weighted avg	0.80	0.82	0.80	8990

Feature Importance: Gradient Boosting Model

- Top 10 most important features shown at right
- PAY_0 is the most important feature determining default, followed by PAY_2 and PAY_3
- NO demographic factors appear in the top 10 most important features predicting default

	Feature_Names	Importance
3	PAY_0	0.631345
4	PAY_2	0.090294
5	PAY_3	0.033371
14	PAY_AMT3	0.031700
0	LIMIT_BAL	0.027736
9	BILL_AMT2	0.025781
8	PAY_6	0.021677
13	PAY_AMT2	0.020507
12	PAY_AMT1	0.020248
6	PAY_4	0.018377

Interpretation and Recommendations

- GB model can predict default with 82% accuracy
- Features most important in default prediction are data gathered post-credit approval and within three months of default
- Demographic features, which determine credit eligibility, are not in the top 10 most important features predicting default
- Therefore, the current data and model are insufficient to adjust approval and/or lending policies to decrease the number of defaults
- More transaction records or additional features (e.g., customer income) may help identify other factors that contribute to customer propensity to default

Lessons Learned

- Consider all features in EDA! I initially ran visualizations for only demographic factors as I assumed one of them would likely be the key to understanding default – my assumptions were based on real world experience. Here I violated Guido's first rule, "Let the data tell the story – don't make any assumptions."
- To be more thorough, I should have changed the data type of our categorical values to integer in order to include them in the correlation analysis and modeling.
- Step 3 in the POA threw me off when it described credit limit (LIMIT_BAL) as the dependent variable. I am unclear how we would be able to determine a better balance limit for customers, given our conclusions.
- I need to investigate further the "PAY_X" features more thoroughly to determine if any parts of the data could provide more information about default, such as use of revolving credit, late payments, etc.
- I am still a little unclear about how we'd use the model in the future. Assume that CreditOne changed lending policies and had a new batch of records: how would we, as data scientists, "plug" the new customer data into the model to determine if a customer should be granted a line of credit and how much?