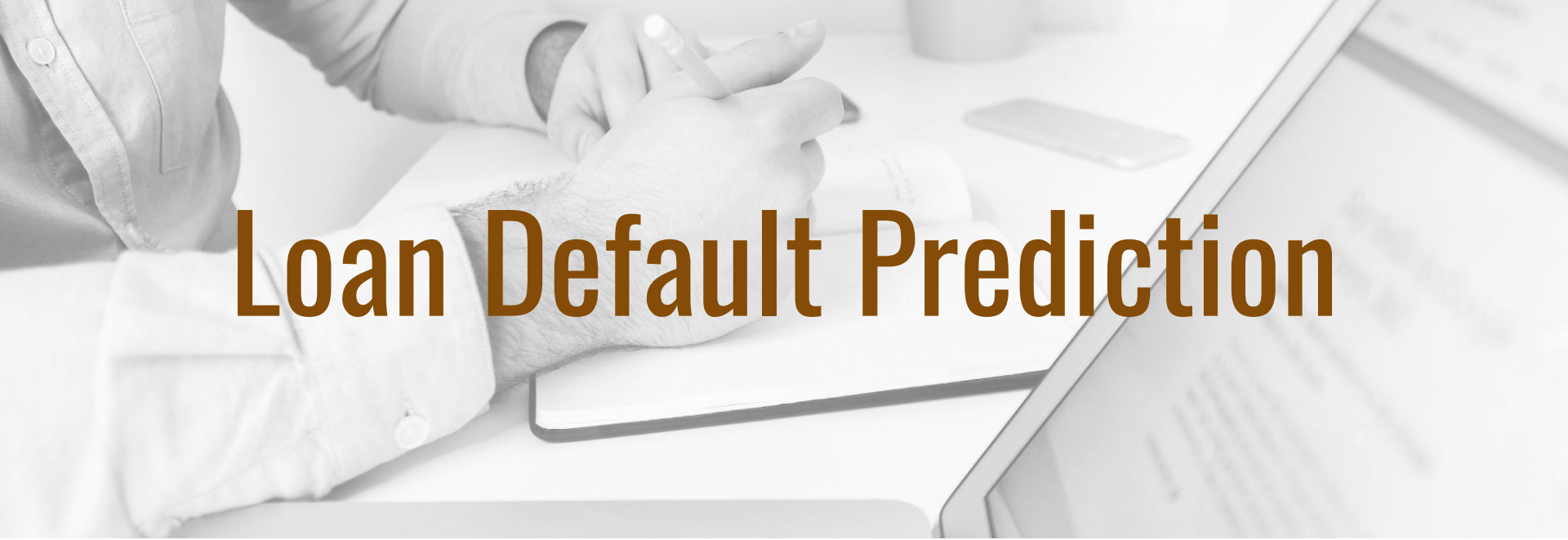


Loan Default Prediction

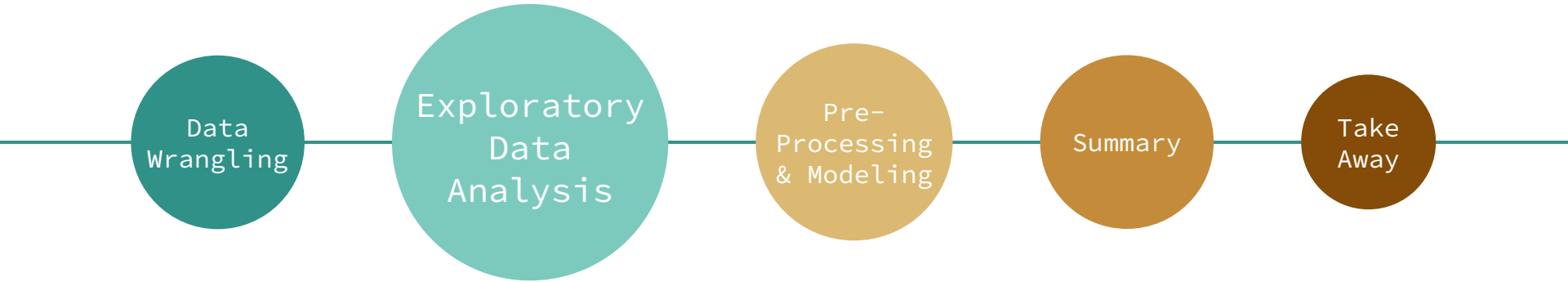
Machine Learning Case Study

Data Source: University of Wisconsin

15%

of loans are resulting in a default in our dataset. Can we use machine learning to identify problematic applications that will likely result in a default?

Project Pipeline

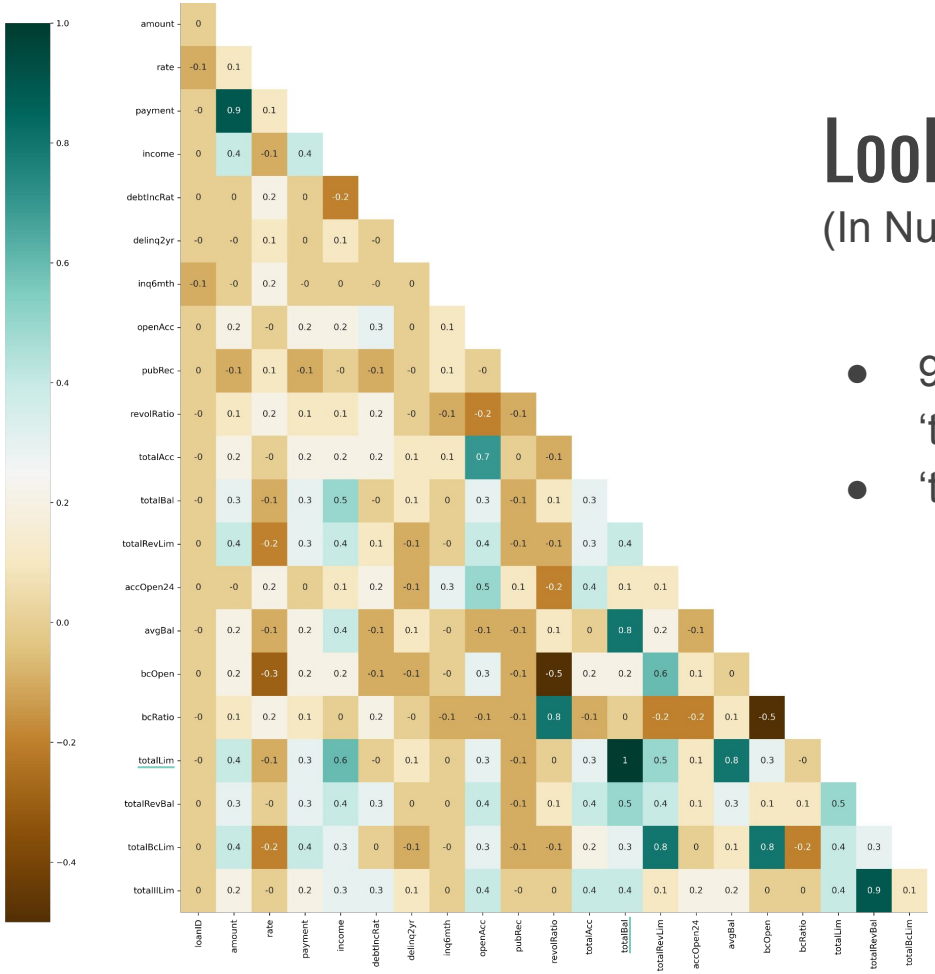


Data Wrangling

- Understanding missing values & how to address them.
- Consolidation of target feature: Loan Status
- Late status categories dropped (1.6% of the data)

Missing Value Summary		
Feature	**Count**	**%**
Employment	2784	5.57
Length (Employment)	2680	5.36
Ratio Total Credit Card Balance to Total Credit Card Limits	520	1.04
Open Credit on Credit Cards	488	0.98
Revolving Ratio	18	0.04

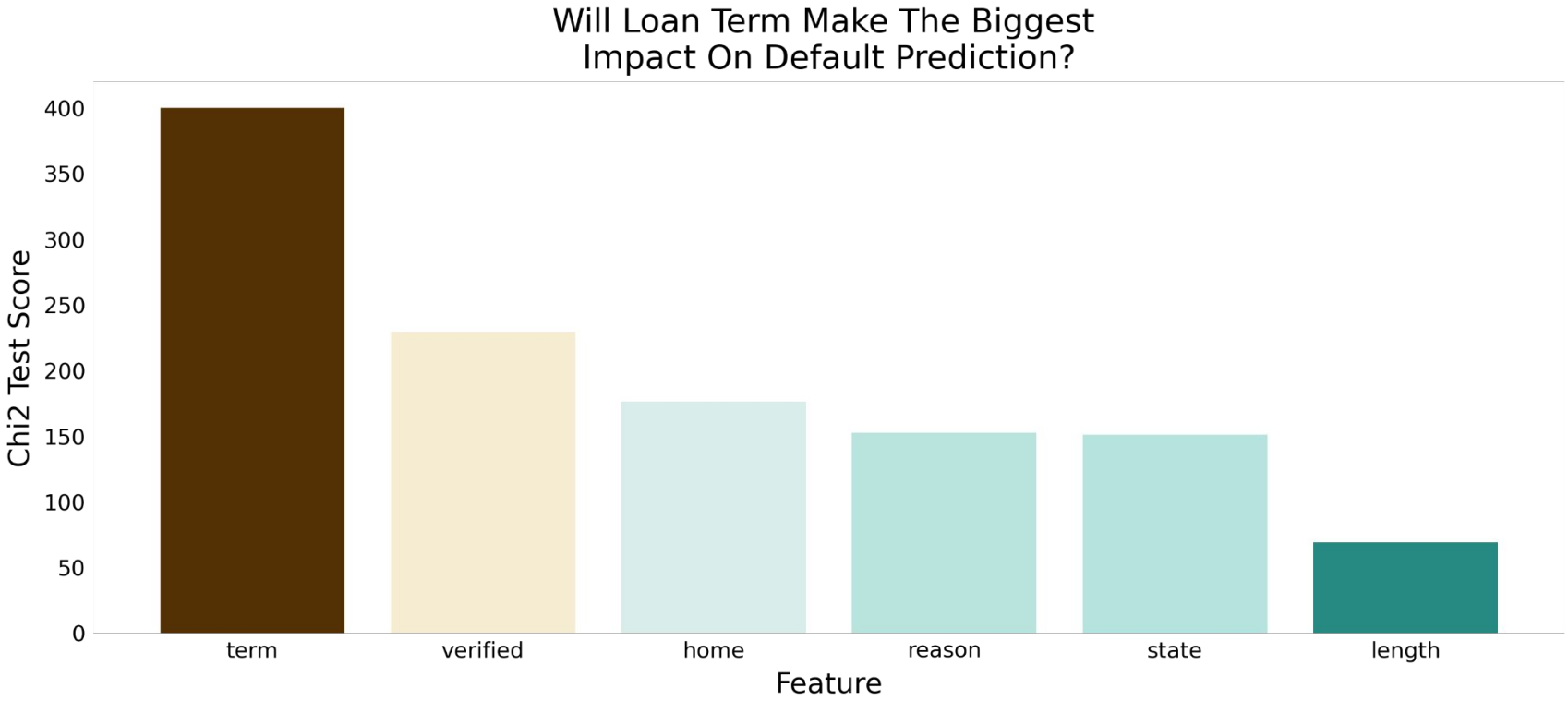
(Pre-Consolidation) Loan Status Summary		
Loan Status	**Count**	**%**
Fully Paid	27,073	54.15
Current	14,531	29.06
Charged Off	7,578	15.16
~~Late (31-120 days)~~	~~449~~	~~0.90~~
~~In Grace Period~~	~~261~~	~~0.52~~
~~Late (16-30 days)~~	~~102~~	~~0.20~~
Default	2	0.004



Looking for Correlation

(In Numerical Features)

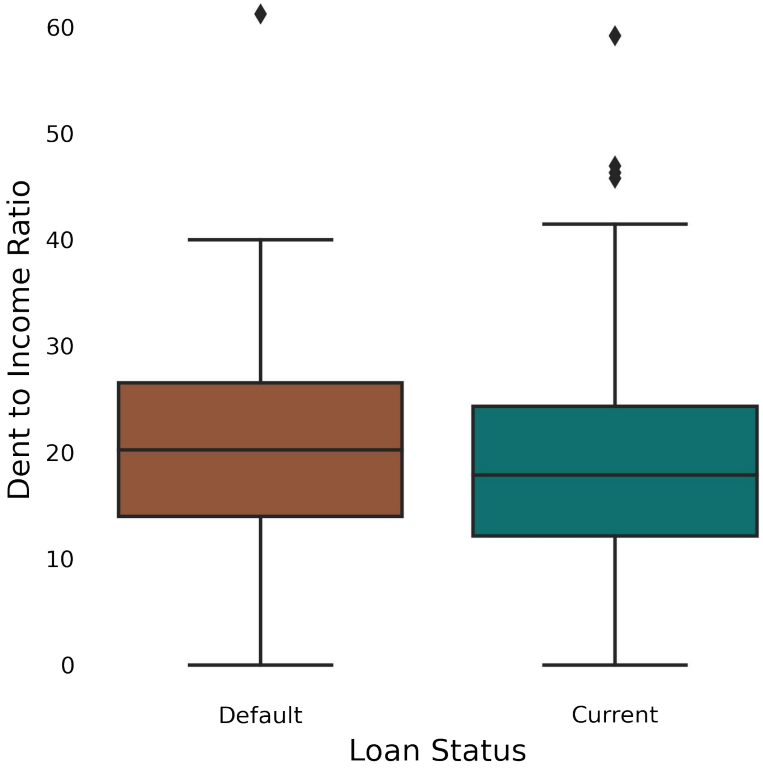
- 99% correlation between ‘totalBal’ & ‘totalLim’,
- ‘totalLim’ feature removed



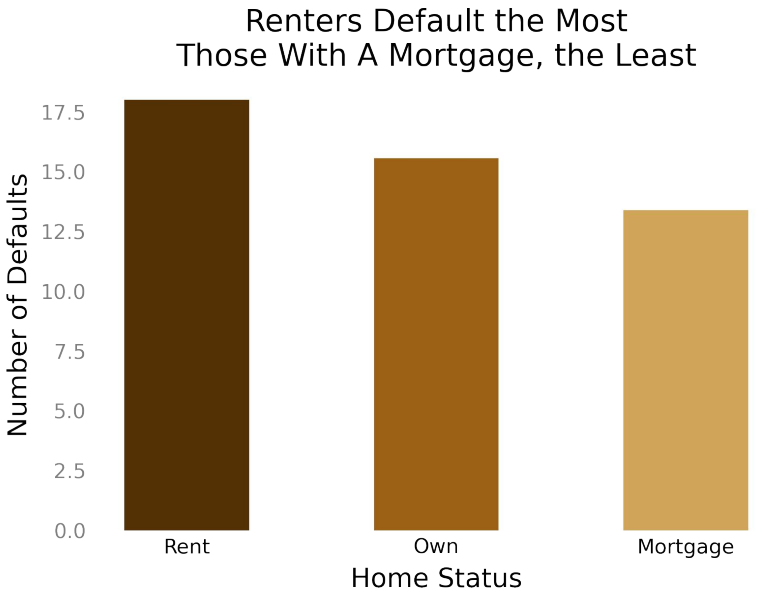
Chi Squared Testing

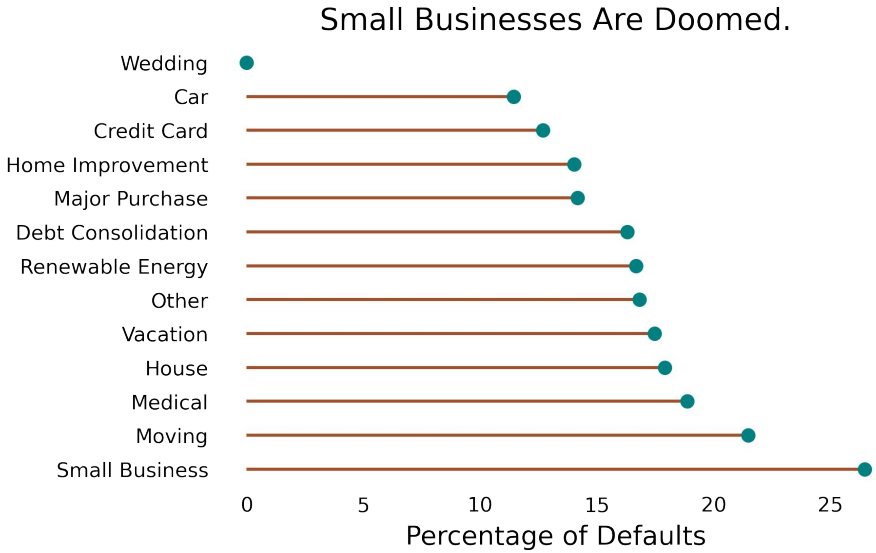
(Between Categorical Features)

Defaulters Have A Higher Debt to Income Ratio



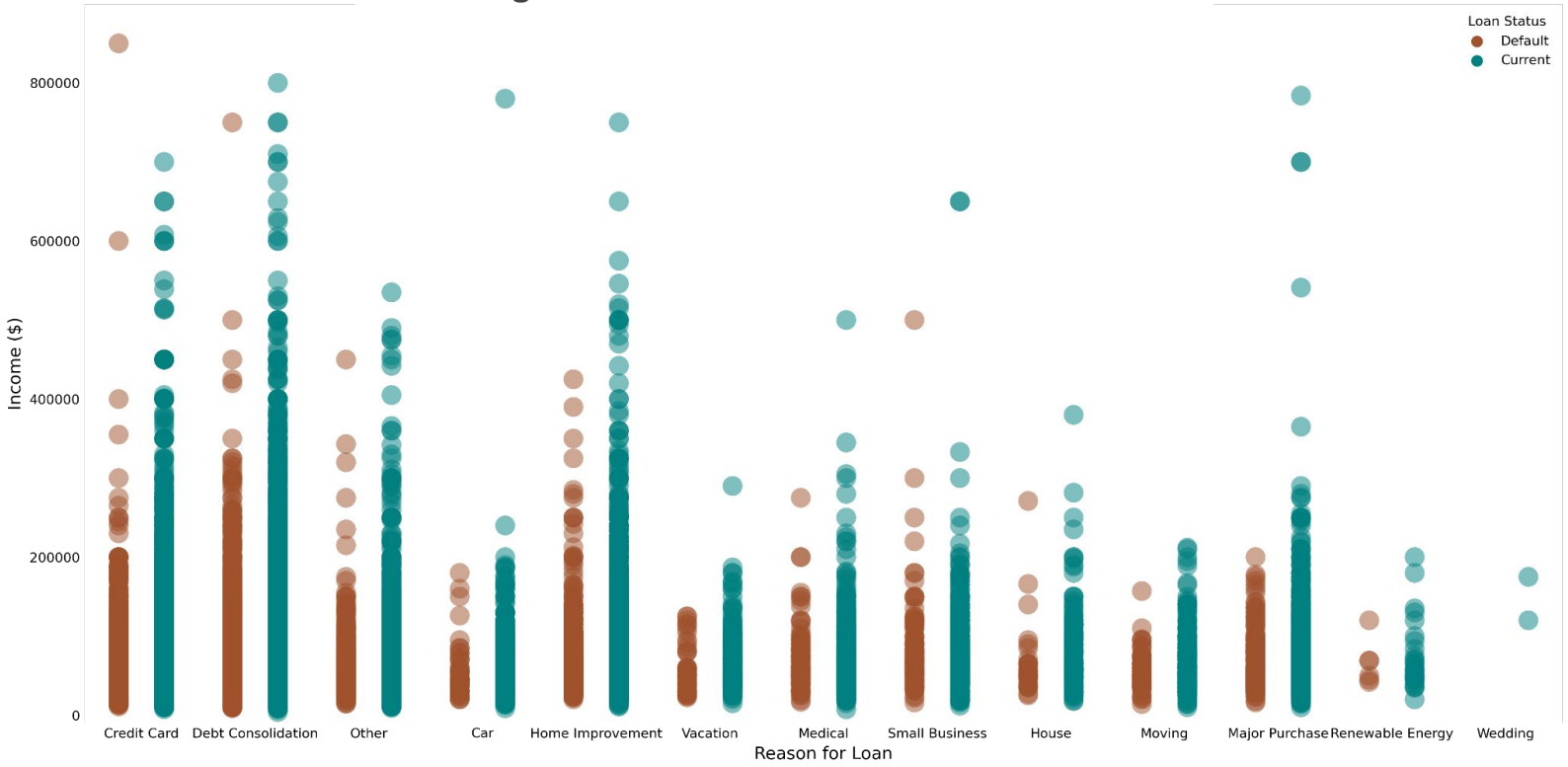
Renters & Small Business Loans Default More



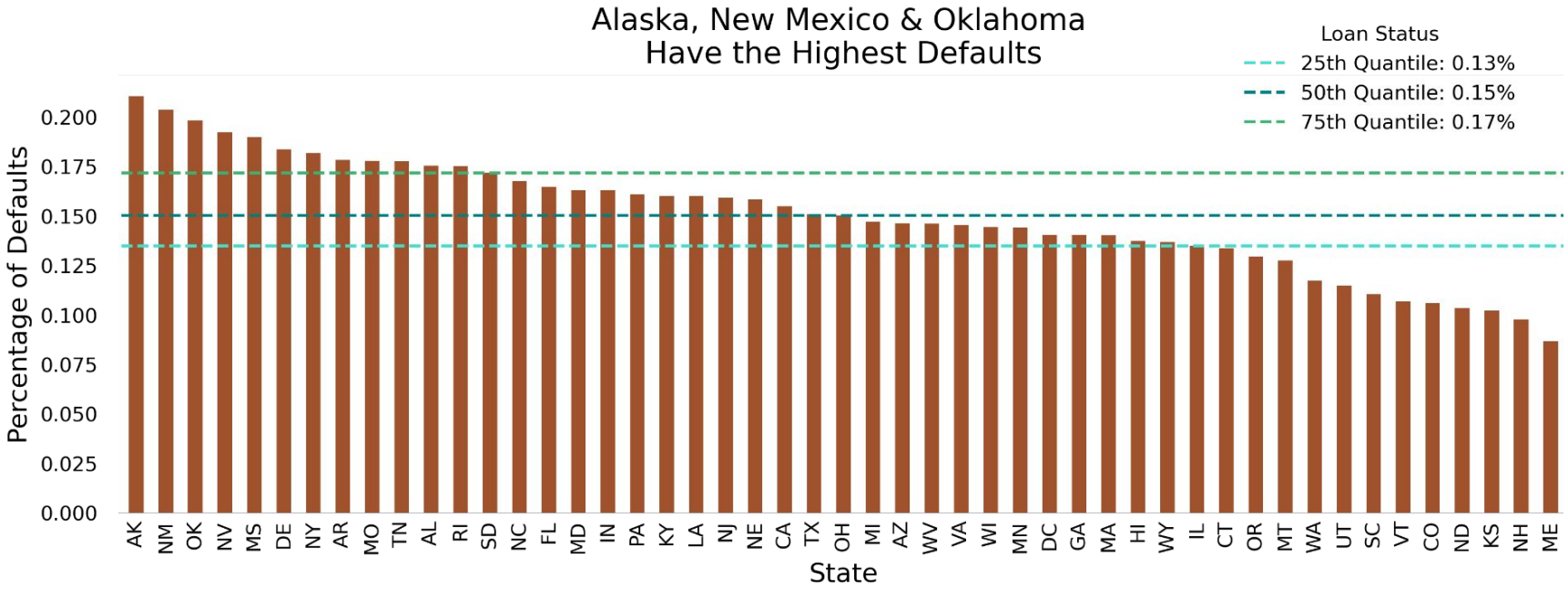


Those With Highest Incomes

Are Getting Loans for Credit Cards & Debt Consolidation

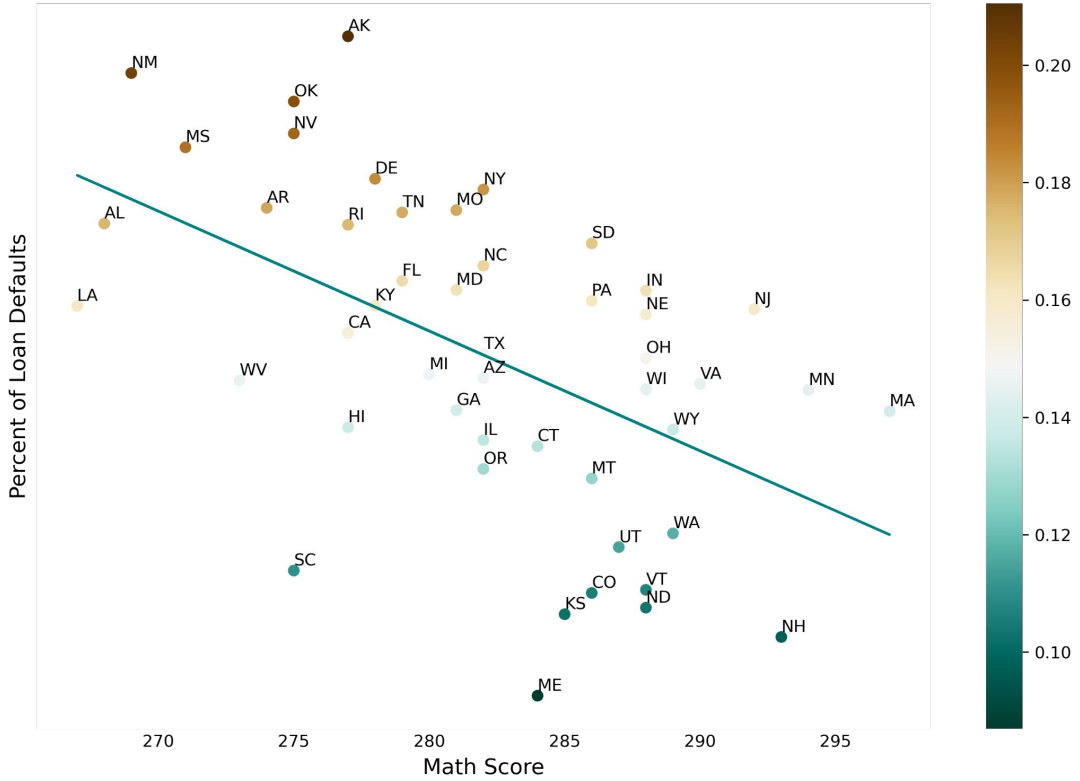


Does it Matter Where The Applicant Lives?



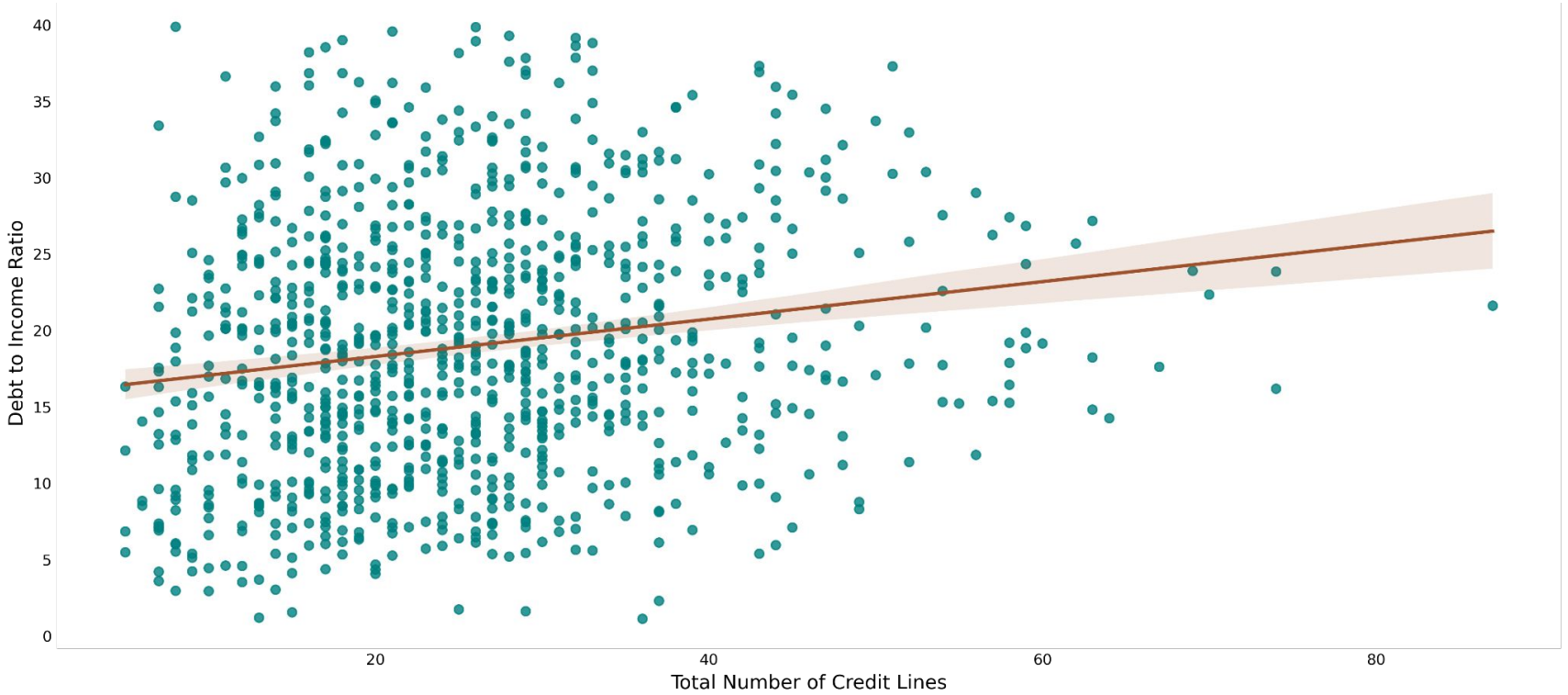
Side Track: Math Scores

Is there a correlation between math scores & loan defaults?



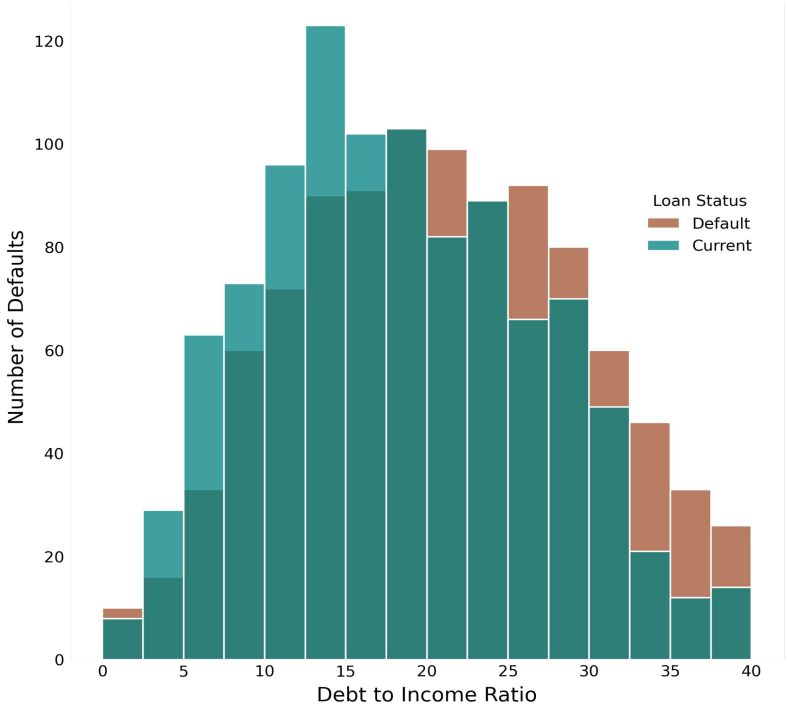
[Link to Education Dataset]

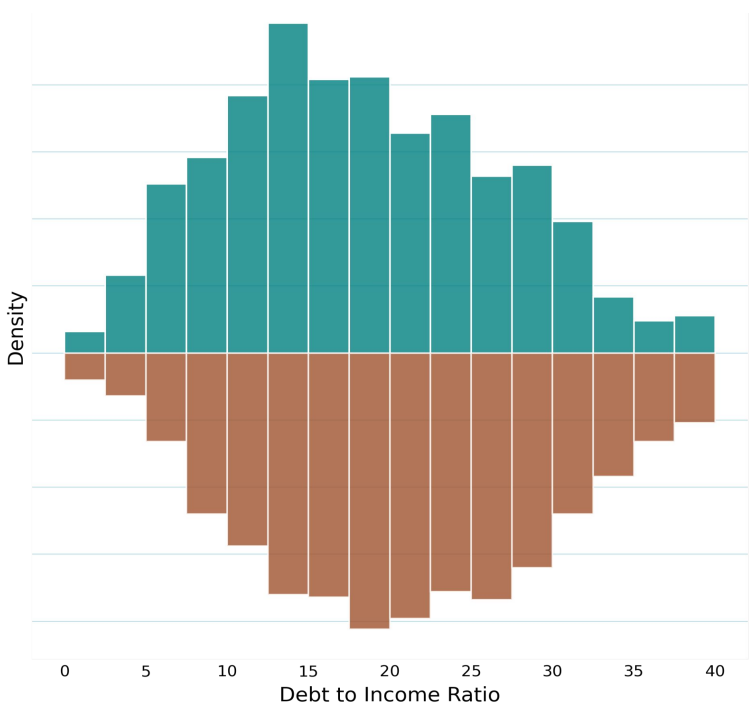
More Credit Lines, Higher Debt to Income Ratio



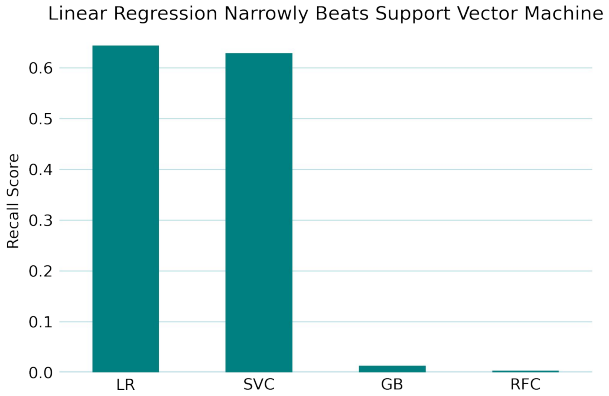
Defaults Go Up As Debt to Income Ratio Increases

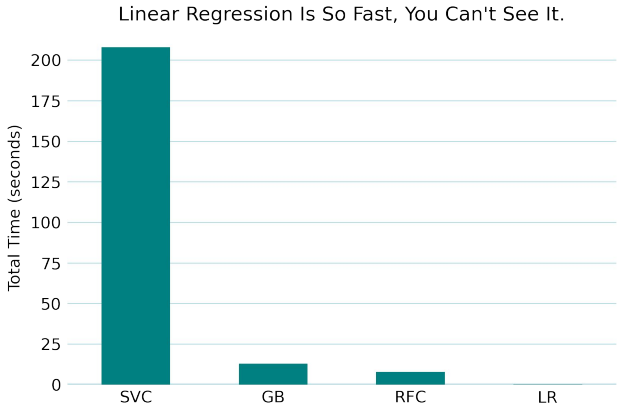
Visualized Two Ways





Model Performance Metrics & Comparison





Model Performance Testing					
Model	**Recall**	**Precision**	**Accuracy**	**F1 Score**	**Run Time (Sec)**
Logistic Regression	64.4%	23.8%	63.9%	34.8%	0.13
Random Forest Classifier	0.3%	55.6%	85.1%	0.7%	8.19
Gradient Boosting	1.4%	55.6%	85.1%	2.7%	13.34
Support Vector Machines	62.9%	23.6%	64.0%	34.3%	208.88

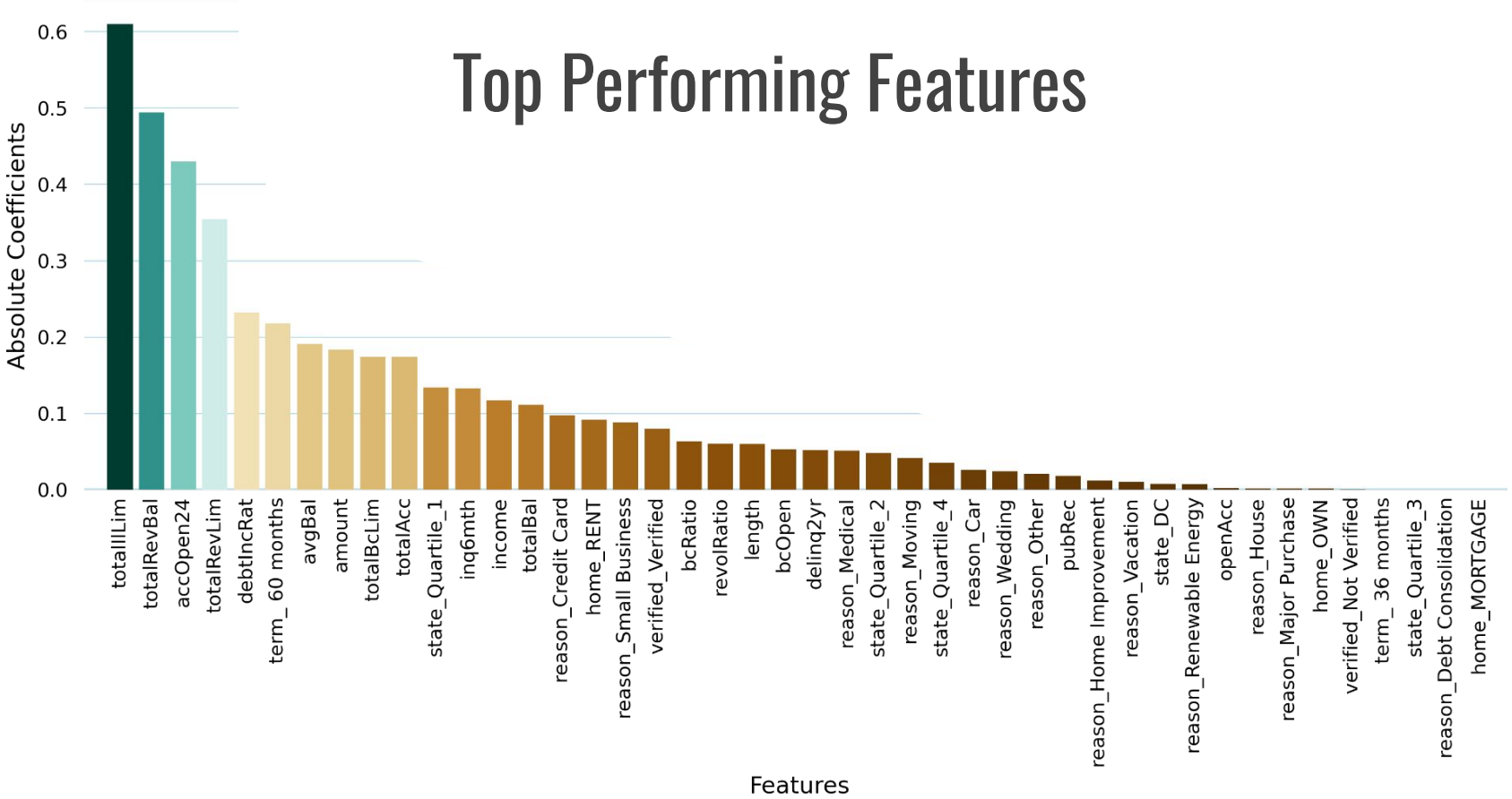
Tuning The Model

Logistic Regression + Sampling Techniques				
Model	**Recall**	**Precision**	**Accuracy**	**F1 Score**
(No Sampling)	64.4%	23.8%	63.9%	34.8%
Smote - Synthetic Oversampling	65.2%	24.0%	63.9%	35.0%
Random Oversampling	64.1%	23.8%	63.9%	34.7%
Random Undersampling	63.6%	23.8%	64.1%	34.6%

Logistic Regression + Sampling + Tuning				
Model	**Recall**	**Precision**	**Accuracy**	**F1 Score**
(No Sampling)	64.4%	23.8%	63.9%	34.8%
Smote - Synthetic Oversampling	65.2%	24.0%	63.9%	35.0%
Smote - Synthetic Oversampling + Hyperparam Tuning	65.3%	24.0%	63.9%	35.1%

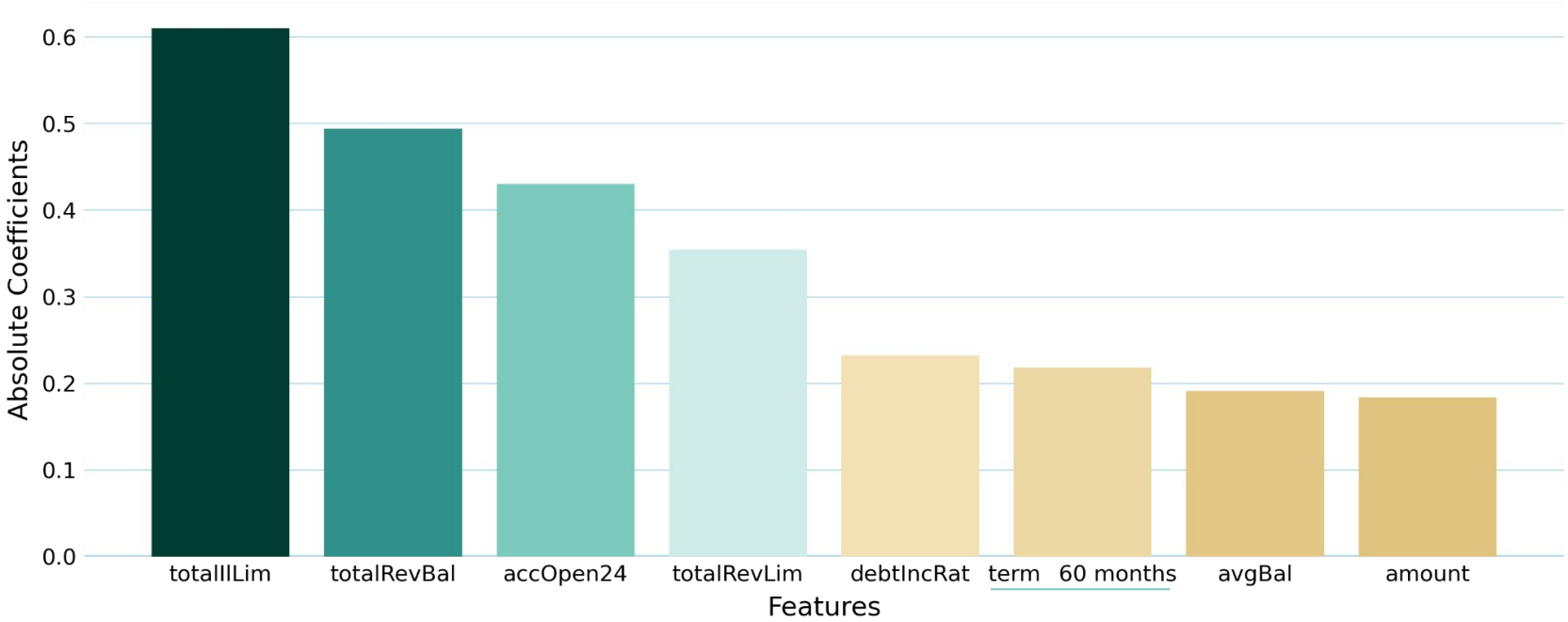
- 3 Different Sampling Methods Explored.
- The winner - Smote Synthetic Oversampling - is hyperparameter tuned using randomized search.

Top Performing Features



Absolute Coefficients: Top 8

Our Chi Test Frontrunner (Term) Makes the Cut



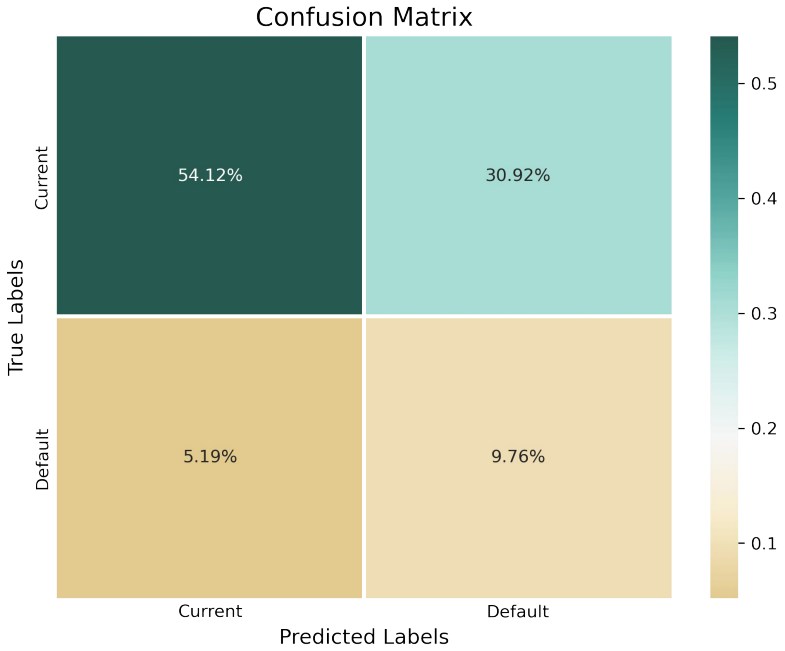
How Can We Achieve Better Results?

Better. Features. More samples wouldn't hurt either.

Final Model Performance Summary				
Model	**Recall**	**Precision**	**Accuracy**	**F1 Score**
Smote - Synthetic Oversampling + Hyperparam Tuning	65.3%	24.0%	63.9%	35.1%

Possible Additional Features

- Borrower Information:
  - Credit score, Age, Education level, Marital status, Number of dependents, Geography
- Loan Characteristics:
  - Interest rate type (fixed or variable), Type of loan (secured or unsecured), Loan origination date, payment frequency
- Economic Factors:
  - Interest rates, Inflation rate, Unemployment rate, GDP growth rate



Contact

Jill Berry
jmarleyberry@gmail.com
LinkedIn Profile
More Details: Project Report
GitHub: Loan Default Prediction

(Datasets & Jupyter Notebooks)