Loan Default/Lending Strategy Project

Alexander Bechler

Portfolio of Personal Projects - Fall 2023

Motivations & Overview of the Problem Statement

Motivations

- Utilizing publicly available data, this project will build an enhancement on an assumed 'existing' lending default model used in an origination strategy for a fictitious lender (named 'Firm A' for this analysis).
  - Project will conduct a short exploratory data analysis, evaluation of model characteristics, and present a recommendation at the end of the process - showcasing the analytics and insights driving the decision along the way.
- Data used in this project is all sourced from **publicly available data** [1] - no proprietary data is used in any of the analysis presented in this document and none of the models presented were sourced from any proprietary sources. Everything in this presentation is originally created by me as purely a learning exercise

Background:

- Firm A currently uses a model-based origination strategy for its installment lending line of business.
  - The existing strategy has been deteriorating, with losses approaching ~30% on this latest data read.
- Firm A has been exploring additional options to improve its lending models and reduce credit losses on the installment lending portfolio.
  - The firm has identified several more additional variables it would like to assess against its data and determine if they capture a reasonable amount of incremental losses to justify the cost.. The cost of implementing these variables is $10,000 and would be a one-time cost.

Data & Model:

- The installment line of business has made available a large dataset of ~1000 tradelines upon which to conduct the analysis.
  - Data was split into training and test data (75%/25%). Results were evaluated on the test data and presented here.

Benefits of New Strategy:

- Lower Credit Losses:
  - Upon implementation of the variables into a new lending PD model, the team was able to save approximately 11% of the existing gross credit losses.
  - The current model approved bad loans in the dataset that had ~$204k of credit losses, while the new model reduced the bad volume by $42k to $162k (or ~20%)
- Fewer 'good loans' rejected:
  - New model also rejected fewer 'good' loans ($112k in good loans rejected vs $138k in the existing model).
  - Utilizing a fair market accounting revenue recognition of ~25%, that would be an increase in revenue of about $7,000
- Cost of Implementation & Associated Savings in Gross Credit Losses more than justify the additional cost of variables
  - New model would reduce losses by >$42k and increase revenues by $7k, more than offsetting the $10k in associated costs to implement the change

1. ***Data used in this analysis:*** *https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data*
2. *Analysis conducted using R and Python programming languages*
3. *Training/Test data assumed to be indicative of the entire population*

Existing Model is not capturing ~48% of credit losses and rejecting ~$138k in ‘good’ installment loan volume

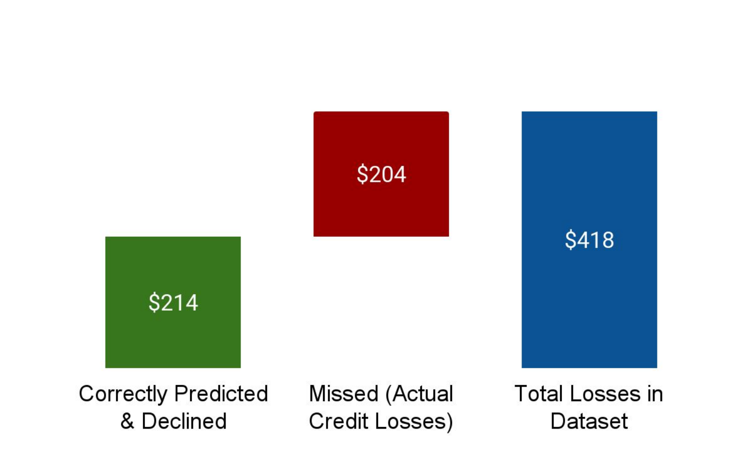
1 Existing model is missing ~48% of losses in the dataset ($204k) and rejecting ~$138k in ‘good’ installment loan data

- Existing model incorporates tradeline size, purpose, duration, credit history, housing, residence length, and other installment loans

2 The overall accuracy of the existing model is 72%

- Existing model is able to capture ‘good’ loans well, represented by a sensitivity value of ~85%
- Model falls short in capturing the ‘bad’ loans, with a specificity value of only ~46%

1 **Dollar Volume of Identified & Missed Defaulted Tradelines for Existing Installment Loan Model ($000’s)**



2 **Confusion Matrix for Existing Strategy**

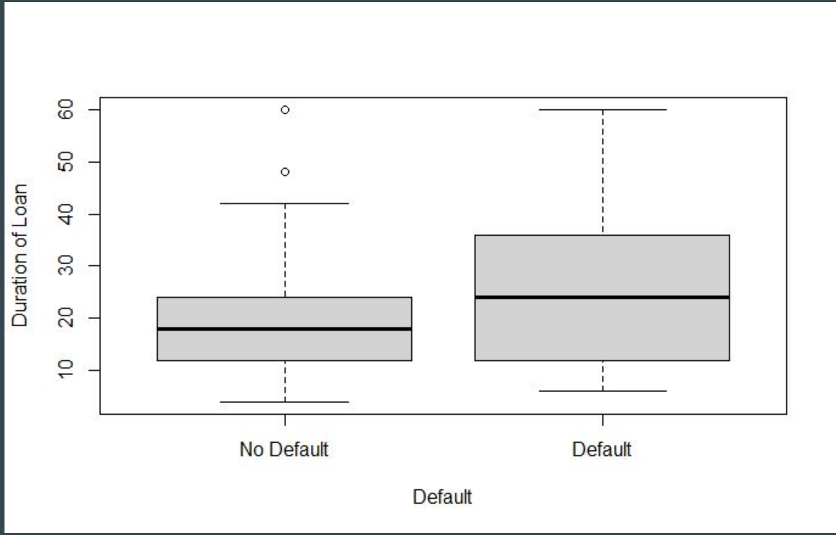
		Actual (reference) Data	
		No Default	Default
Existing Strategy Prediction	No Default	165	50
	Default	28	43

Loans with longer terms and larger origination size tend to default at higher rates for Firm A installment loan customers

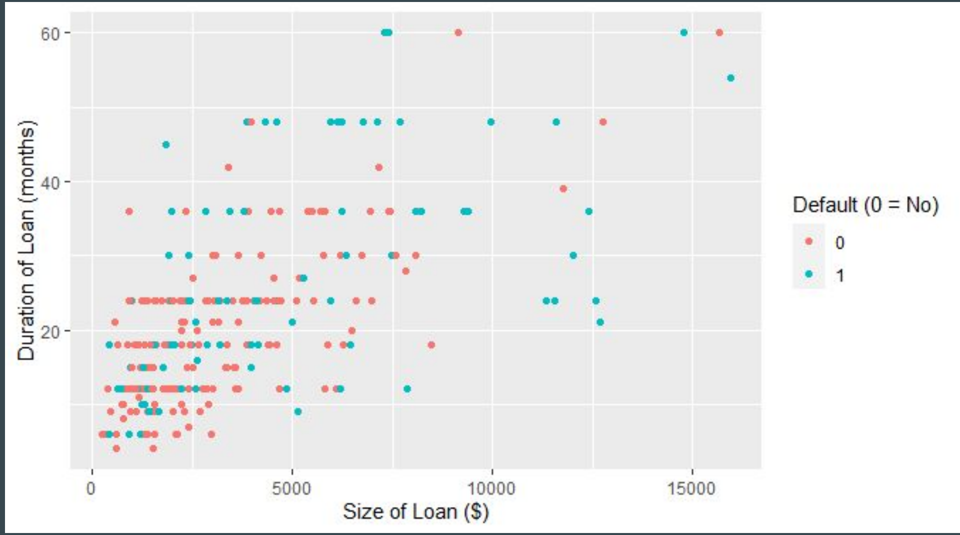
1 Installment loans that default tend to have a longer term (measured in months) than loans that do not default.
- Median duration for defaulted loans is 24 months, compared to 18 months for those that did not default

2 Loans that are longer in duration and larger in origination size also tend to have higher default rates
- For tradelines with a duration of > 30 months, default rates were > 50%.
- For tradelines with duration of less than 12 months, default rates were <15%

1 Comparison of Loan Duration by Default Disposition



2 Duration of Loan vs. Size of Loan, by Default Disposition



Firm A has made three new variables available that all serve as strong risk identifiers

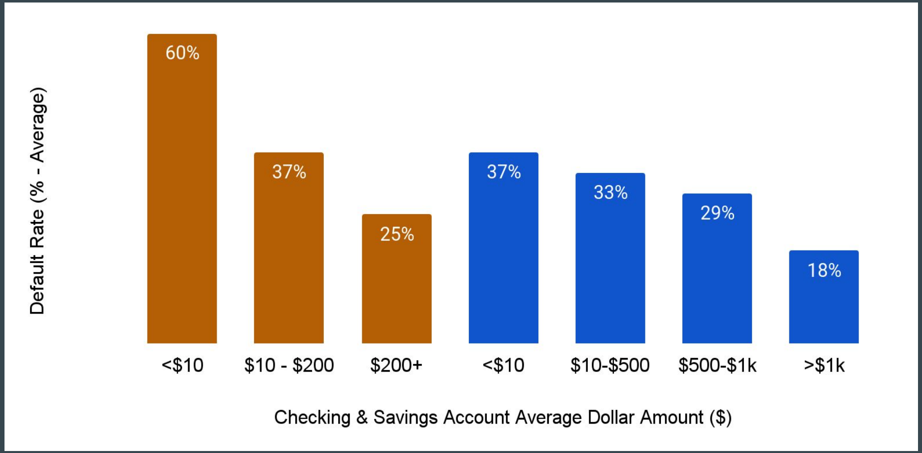
- Analysis has identified three new variables that appear to capture significant incremental risk along a very clear gradient
  - Checking Account Volume ($), Savings Account Volume ($), and Debt to Income Tier all capture a significant amount of incremental risk

1 Customers with high checking account and savings account balances default at significantly lower rates than customers with low balances.
  - Default rates for customers with low checking balances are > 2x higher than customers with high checking account balances

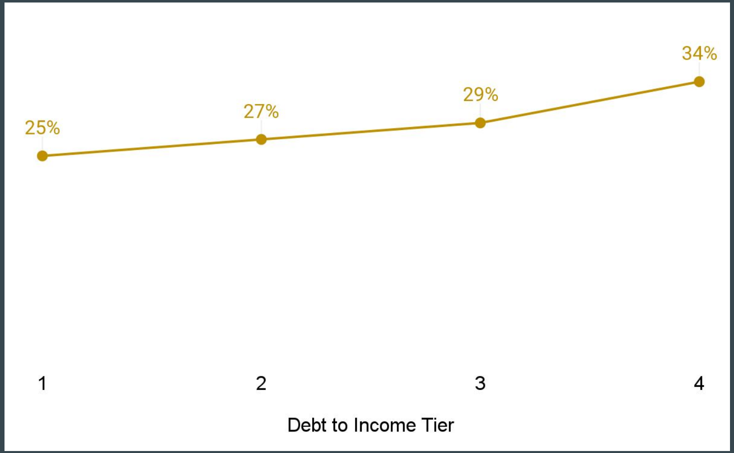
2 Debt to income also serves as a strong incremental risk identifier.
  - Represented in tiers (1 = low, 4 = high), customers with a high debt to income ratio have a default rate ~10% higher than customers with lower debt to income ratios

- Variables will be incorporated into the existing model to see if incremental improvements are worth the financial cost

1 Default Rates by Checking and Savings Account Dollar Volume



2 Debt to Income Tier and Default Rate

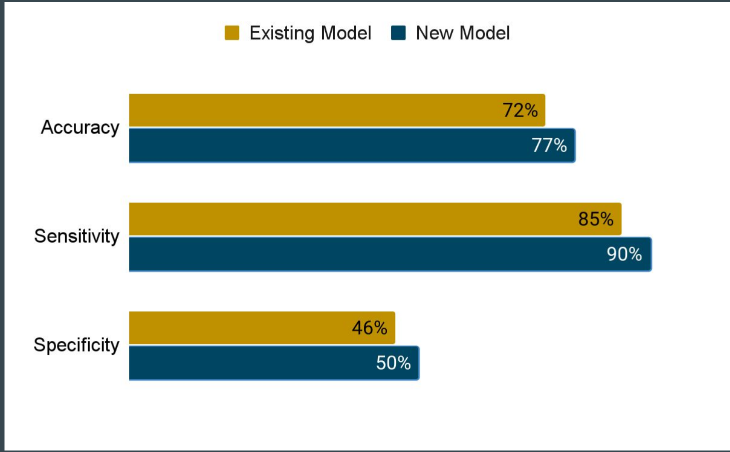


Incorporation of New Variables Improves Model Performance

1 New variables improve model metrics across all three of the major measurements included here

- Model is overall more accurate, with accuracy improves 5%
- Model is able to capture more 'good' loans, with sensitivity improving by 5%
- Model is able to capture 'bad' loans better, with specificity improving by 4%

1 Comparison of Selected Model Measurement Parameters



Confusion Matrix for Existing Strategy

<table>
<tr><td colspan="2"></td><td colspan="2">Actual (reference) Data</td></tr>
<tr><td colspan="2"></td><td>No Default</td><td>Default</td></tr>
<tr><td rowspan="2">Existing Strategy Prediction</td><td>No Default</td><td>174</td><td>48</td></tr>
<tr><td>Default</td><td>19</td><td>45</td></tr>
</table>

New model captures ~10% additional sample gross credit losses and reduces losses missed by ~20%

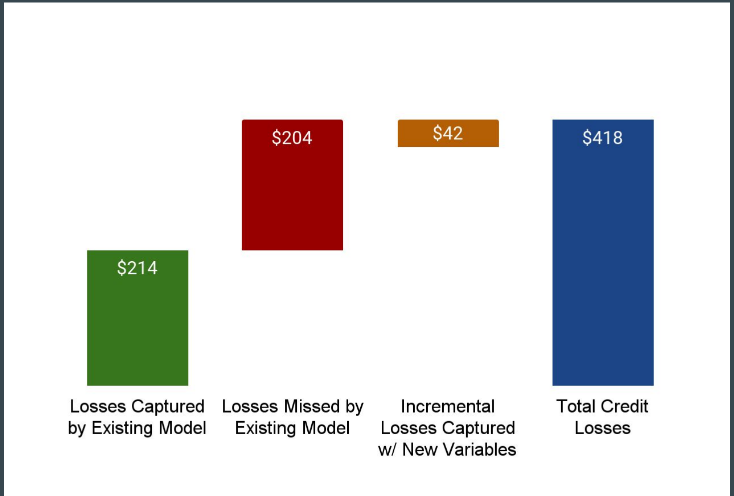
1 Incorporating new variables into the model captures an additional $42k in losses (~20%), or 10% of losses in the entire dataset

- ‘Bad’ loans approved by the model would **drop** from $204k to $162k
- New variables would also approve more ‘good’ loans, representing an increase in revenue on good tradelines that Firm A would experience under this new model

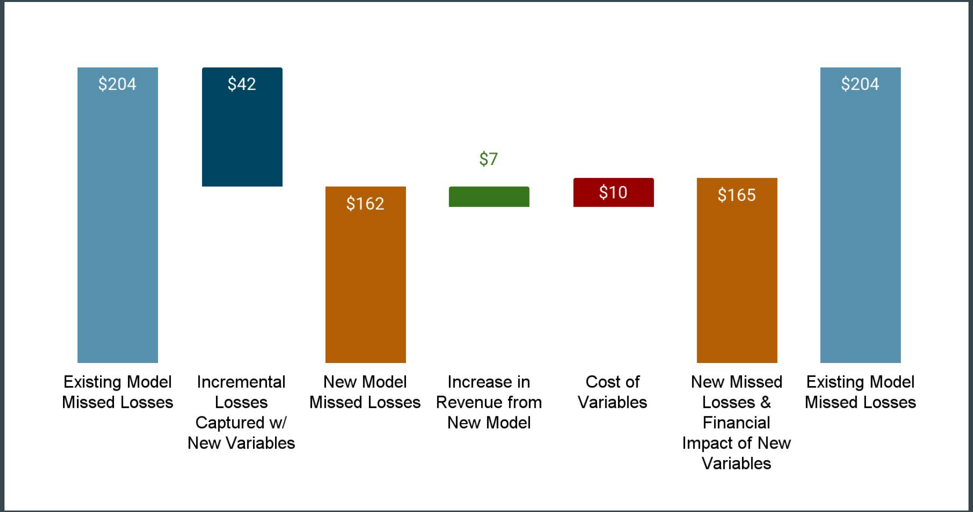
2 Adding the increase in revenue from additional good tradelines ($7k), as well as the cost of implementing these variables ($10k), total credit losses & financial impact of this new strategy would be an overall improvement of ~19% ($204k yo $165k)

- Reduction in Losses more than offset the cost of the new variables

1 Improvement to Credit Loss Outcomes w/ New Model



2 Financial Evaluation of Implementing New Variables



Appendix 1: ROC and AUC Curves

Receiver Operating Characteristic Curve (ROC):

- Generally both models are better than a ‘random guess’ (represented by the black line)
- New Model generally performs better across the entire true positive/false positive rate curve
- Area under the curve (AUC) for the existing model is 0.74, while the AUC for the new model is 0.78.