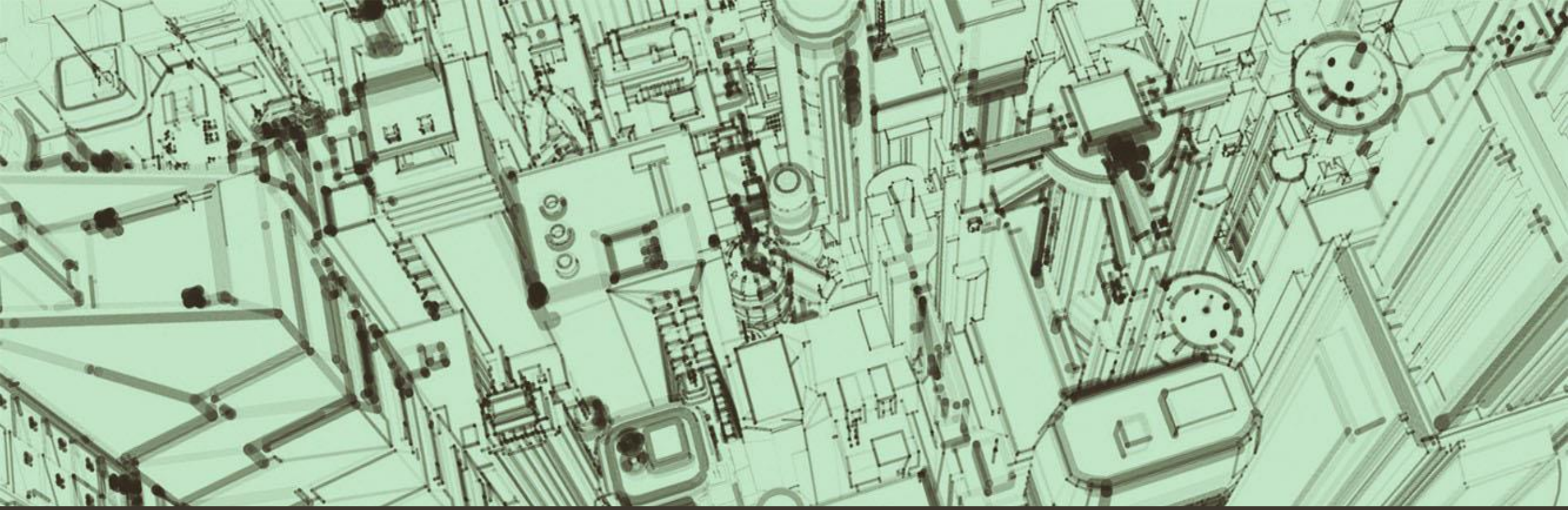


Applying Advanced Modeling Techniques to Credit Risk Modeling

Alexander Bechler

Executive Summary

Background	• Credit risk modeling has become a growing discipline within financial services following the crisis of 2008. Banks have established teams full of model developers that produce models to estimate credit losses on financial instruments. The output of these models then are deployed for financial planning by senior leaders. • Simple 'baseline' methods have been traditionally employed, but this analysis will explore more advanced modeling techniques to see if they perform better than baseline methods.
Dataset	• To test our hypothesis, a dataset with home mortgages will be used. • The dataset contains ~50k observations with 16 different variables. The variables are both customer level attributes (data specific to a customer) as well as macroeconomic attributes (attributes that impact the entire economy). • Expect some from both categories to be important in the model
Methodology	• Utilizing the traditional credit risk modeling framework, we will build a probability of default (PD) model and a loss at default (LaD) model to calculate expected losses. • We will use the baseline methods of simple linear and logistic regression as well as more advanced modeling such as a random forest or support vector machine. 6 total models will be used, 3 for PD and 3 for LaD. • We'll calculate expected losses based on all possible combinations of the PD and LaD models. The best combination will be what is used to compare to the baseline modeling results
Key Insights	• The **advanced** modeling efforts generally performed better than the baseline modeling efforts. The combination of a KNN model and a random forest produced a result that reduced the overestimate by the baseline method by 38% and freeing up $1.2M that would have otherwise been held for losses. • The actual credit losses of the dataset were ~$42.6M, the baseline methods estimated a $45.8M loss while the best combination of the advanced models produced an estimated loss of $43.6M • More accurate model results ensure a better estimate, reducing an over/underestimate and freeing up capital for better productive use

But first … a little background for the uninitiated ..

- Credit Risk modeling is a process banks use to project the losses they will incur on their financial instruments (loans, stocks, bonds, etc.). Loss forecasts are run on a **quarterly basis** and are released publicly to investors[1]
  - **Credit loss forecasting** informs how much money banks need to keep on their balance sheet to **hold against expected credit losses**. This capital that is reserved for losses **cannot be used for investment purposes or dividend purposes**, so it represents a significant line-item for banks on a balance sheet.
  - Depending on the size of a banks assets, this could become billions of dollars. In 2023 alone, Bank of America had about $20 billion in reserves held for losses[2]
  - Accuracy in these results is important, as an over/underestimate for actuals can result in insufficient capital (requiring a reduced dividend) or excess capital that would sit on the balance sheet that could otherwise be deployed representing an opportunity cost.
- Methodology for credit loss modeling is tightly regulated and guided by the US Federal Reserve System since the crisis of 2008.[3]
  - CCAR (Comprehensive Capital Analysis and Review) and CECL (Current Expected Credit Loss) are two of the many different regulatory frameworks for establishing a fair market estimate of credit losses for a bank.
  - IFRS-9, DFAST, etc are the formal regulatory guidelines that establish the rules for credit risk modeling
  - Credit risk modeling takes place in three stages: model development, model validation, and model governance. Development builds models, validation challenges models, and governance monitors the models as they are put into production.
  - Banks have countless teams that are devoted to due diligence in each of these areas. Each separate area must remain **strictly independent** of each other to ensure that the most accurate models are constructed, free of bias and influence from pressure.

1. https://www.nasdaq.com/CCaR_model_guidelines20240812
2. https://investor.bankofamerica.com/regulatory-and-other-filings/all-sec-filings/xbrl_doc_only/6549
3. https://www.federalreserve.gov/publications/comprehensive-capital-analysis-and-review-questions-and-anwers.htm

More accurate loss forecasting models will improve financial planning, allowing for better capital planning and lower risk of a forecasting miss

Methodology:

- Credit losses are typically modeled as the product of two separate calculations for each financial instrument in a banks portfolio: the **probability of default** (PD) for a loan and the **loss at default** for a loan (LaD).
  - The expected loss (EL) of a portfolio is the **sum of the product of the PD and LaD** for the entire loan portfolio and takes the following functional form: $EL = \Sigma (PD * LaD)$
  - PD models predict a probability of default for each customer and are **classification** problems.
  - LaD models estimate a financial loss, predict a **continuous** response variable, and are standard **regression** style problems.

Business Problem:

- Federal reserve regulation requires banks to have interpretable and explainable results, so this has strictly limited banks to utilizing baseline modeling types such as logistic regression and linear regression.
- Since 2008, many more advanced statistical modeling techniques such as random forests and xGBoosting have becoming more available and practiced in the data science community. These techniques have not been employed in the financial services industry given the requirements by the Fed and have largely not been explored by modelers until recently.
- This analysis will seek to explore if these advanced modeling techniques are superior to the baseline methods.
  - It will test k-nearest neighbors/support vector machine models against logistic regression for PD modeling and random forests/xGBoost models against linear regressions for LaD modeling. It will also test several combination of these models for expected loss calculations and identify which is the best at forecasting expected losses.
  - If the accuracy of the advanced modeling techniques is **greater** than the baseline methods, that would provide a significant benefit in terms of capital planning for a bank.
  - More accurate loss forecasting gives leaders a better picture of their portfolio performance as well as the ability to be more confident in building credit reserves.
  - More accurate loss forecasting results ensures that the reserves actually taken will be much closer to the actual loss predictions, reducing the risk that banks will reserve too much or too little capital – **freeing up capital for more productive uses and reducing the risk of an estimate that is less than what is actually required**

Dataset includes 50,000 observations over 16 variables

- Dataset contains 50,000 observations spread across 16 different variables.
  - Each observation is a mortgage loan that is held by a customer.
  - The customer could be a homeowner, an investor, or an institution.
- Dataset contains two separate types of variables – economic variables and customer level variables
  - Economic variables are variables taken from the economy as a whole
  - Customer variables are attributes ascribed to a customer
- There are two target variables for the two separate models for credit loss calculations: default rates (default_time) and loss (credit loss)
  - Default_time is an indicator variable (0,1) and takes a value of 1 if the customer defaulted
  - Loss represents the financial loss incurred if the customer defaults, calculated by taking the value of the defaulted asset and subtracting the value owed to the bank. The value in the dataset already is this calculation

Variable Name	Type	Response Variable?	Binary?	Description
FICO_orig_time	Customer	N	N	FICO Score at Origination
balance_orig_time	Customer	N	N	Original Loan Size
balance_time	Customer	N	N	Remaining Loan Size
REtype_PU_orig_time	Customer	N	Y	Urban Development Indicator
REtype_CO_orig_time	Customer	N	Y	Condominium Indicator
hpi_orig_time	Economic	N	N	House Price Index at Origination
LTV_time	Customer	N	N	Loan-to-Value Ratio
uer_time	Economic	N	N	Unemployment Rate
investor_orig_time	Customer	N	Y	Indicator if the Owner is an Investor
Interest_Rate_orig_time	Customer	N	N	Interest Rate of the Loan (%)
LTV_orig_time	Customer	N	N	Loan-to-Value Ratio when loan was booked
REtype_SF_orig_time	Customer	N	Y	Indicator for Single Family Home Investor
gdp_time	Economic	N	N	Gross Domestic Product
hpi_time	Economic	N	N	House Price Index
default_time	Customer	Y	Y	Indicator if the Customer Defaulted
Loss	Customer	Y	N	Credit Losses if the Customer Defaulted

Several macroeconomic and consumer-level indicators are correlated with default rates

- Correlation Analysis reveals expected macroeconomic trends and consumer behavior

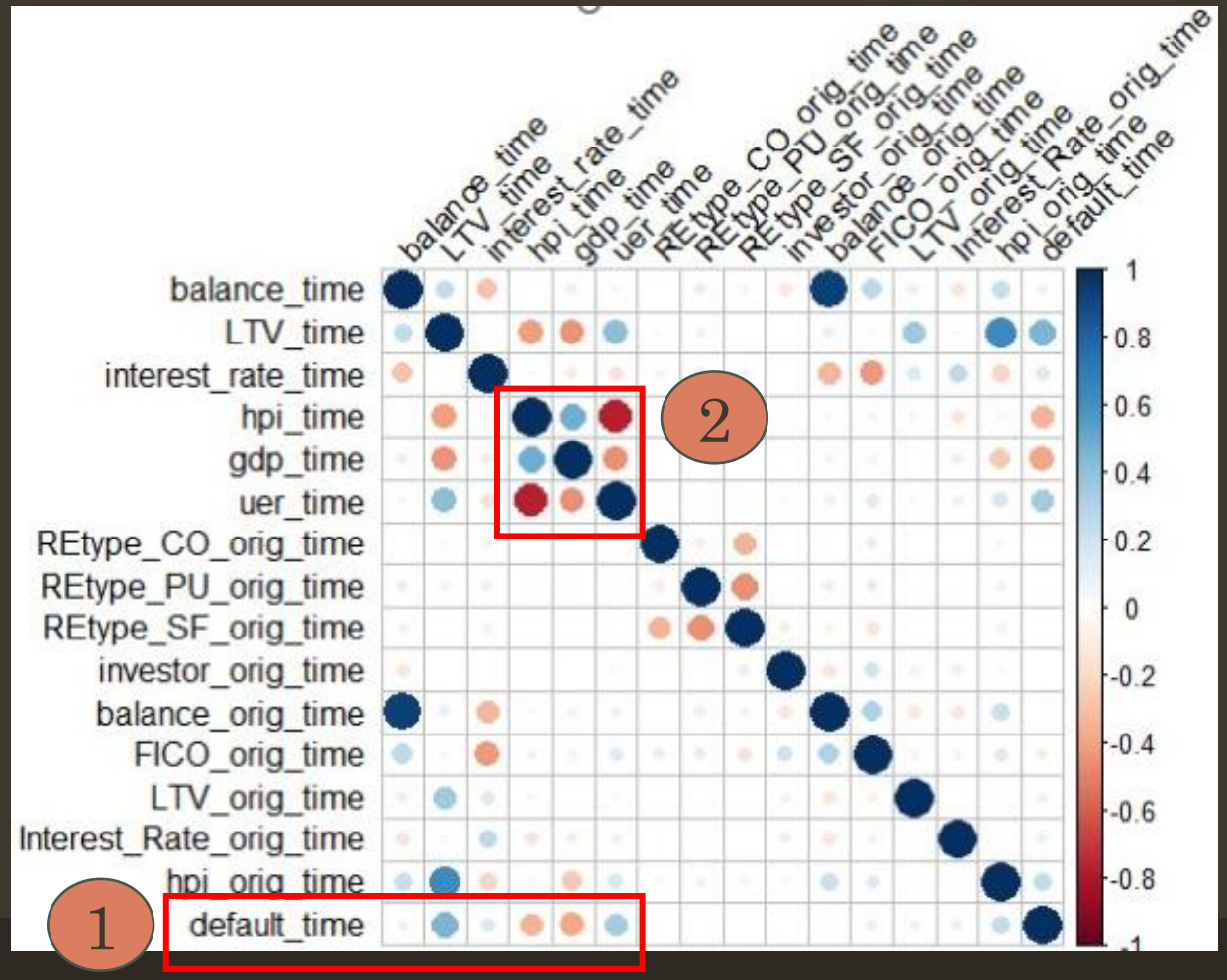
1 Economic stress is a large driver of consumer default rates (default_time)
  - Consumers with a home loan are **more likely to default** when **housing prices decline** (HPI), **the economy enters recession** (GDP), **or unemployment rates rise** (UER).

- As Loan-to-Value (LTV) increases, default rates also increase.
  - Consumers with a high LTV will have higher mortgage payments as a percentage of their income, leading to higher sensitivity to economic stress and therefore default rates.

2 Several macroeconomic variables exhibit correlation between each other
  - HPI, GDP, and UER are all moderately or strongly correlated
  - When economic growth increases, unemployment rate decreases. When economic growth increases, housing prices increases.

Correlation Plot

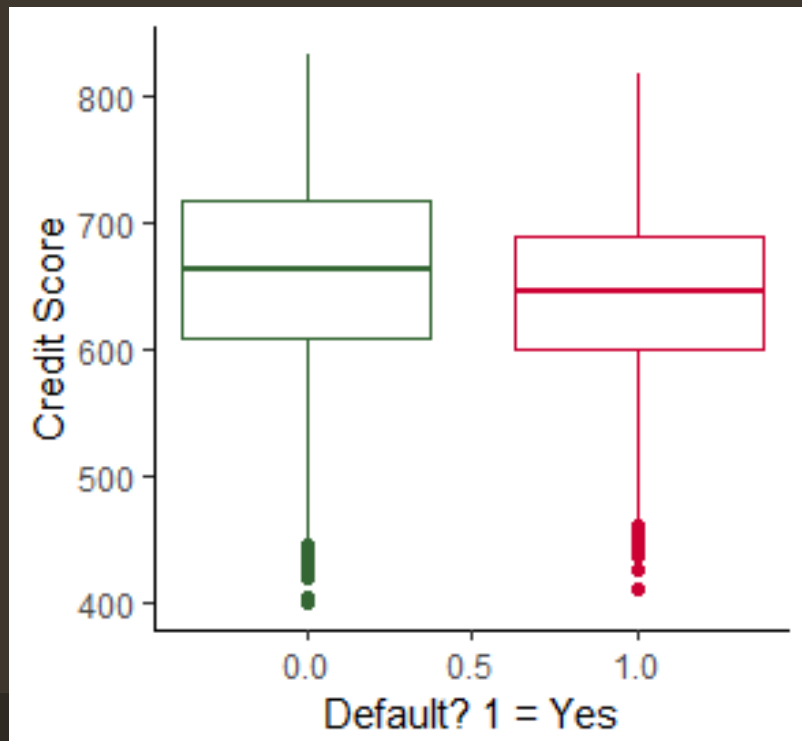
Relationships and Strength of Correlation Between All Variables



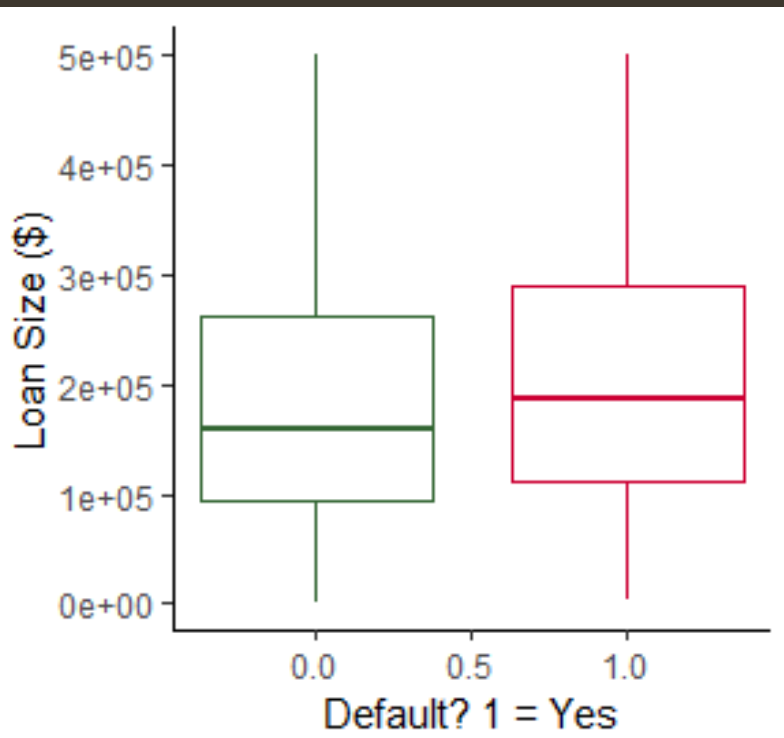
Consumer default behavior appears correlated to lower credit scores, larger loan size, and higher LTV ratios

1. Customers who default on their loan tend to have a slightly lower credit score compared to those who do not, as higher credit scores indicate stronger repayment histories.
2. Customers who default on their loan tend to carry larger loan sizes compared to those who do not.
3. Customers who default on their loan tend to have a higher loan-to-value ratio compared to those who do not, as a higher LTV indicates that the customer is in negative equity on the asset.

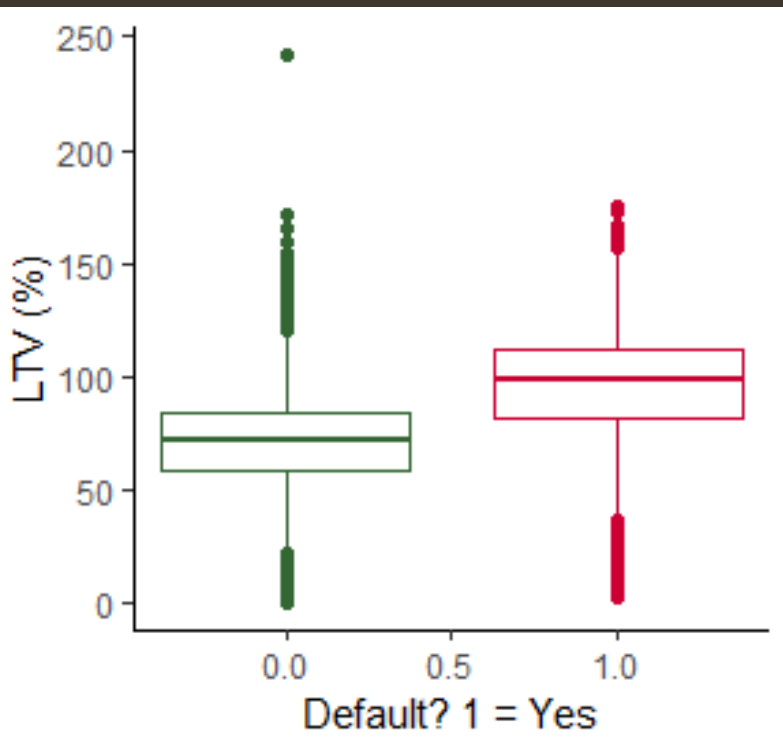
1 Credit Score Distribution
Separated by Default Status



2 Loan Size Distribution
Separated by Default Status



3 Loan-to-Value Distribution
Separated by Default Status



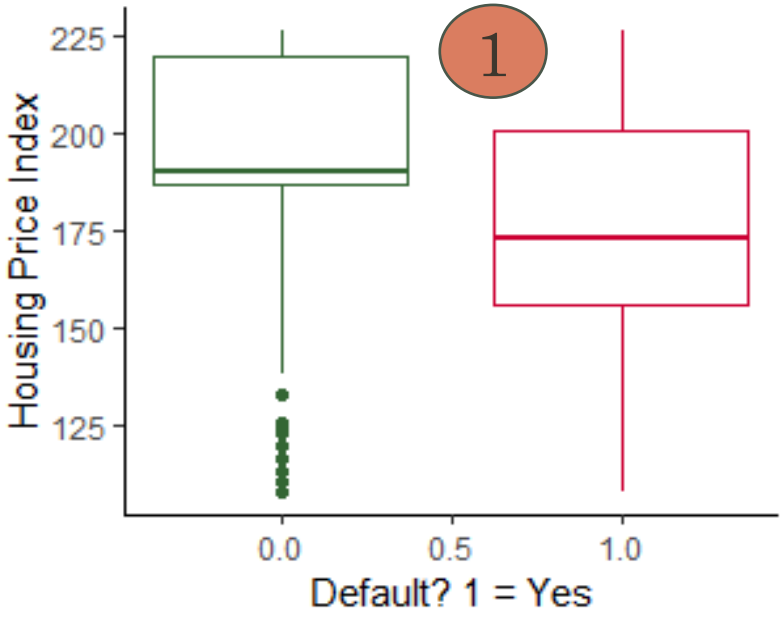
Consumer default behavior appears correlated to lower housing prices, high unemployment rates, and interest rates

1. Customers who default on their loan tend to do so **when the housing price index is lower,** representing falling housing prices and a weakening economy

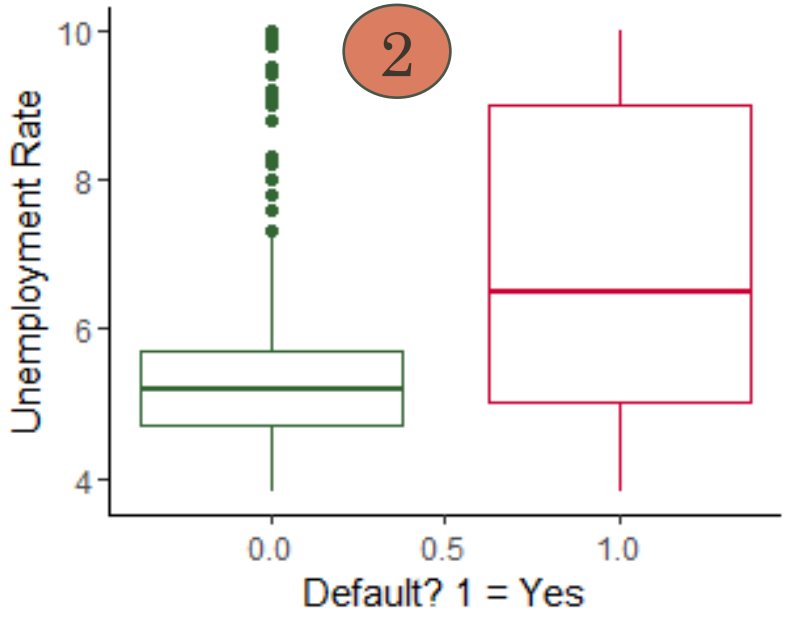
2. Customers who default on their loan tend to do so when the unemployment rate is significantly higher. With higher joblessness, customers are unable to make payments – leading to default

3. Customers who default on their loan tend to have a slightly higher median interest rate, as they typically represent riskier customers and will have higher payments.

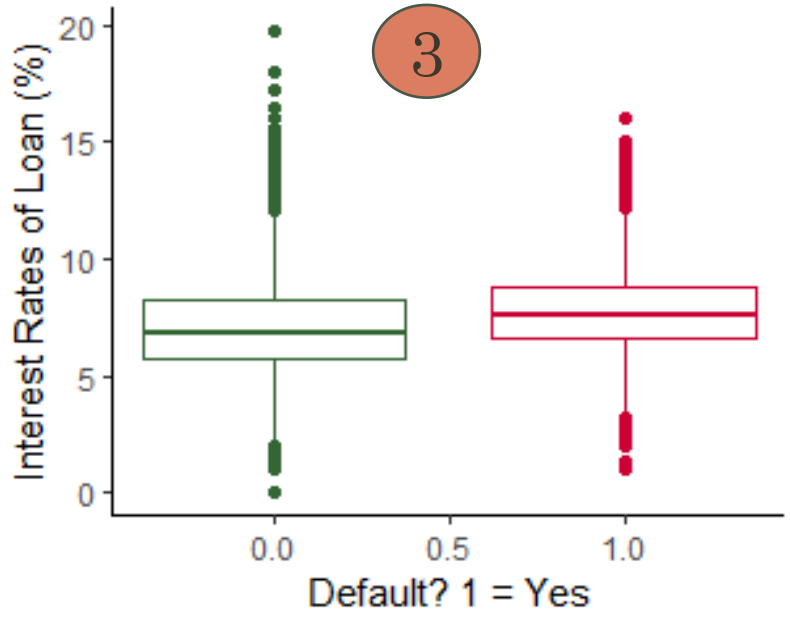
Housing Price Index Distribution
Separated by Default Status



Unemployment Rate Distribution
Separated by Default Status



Interest Rate Distribution
Separated by Default Status

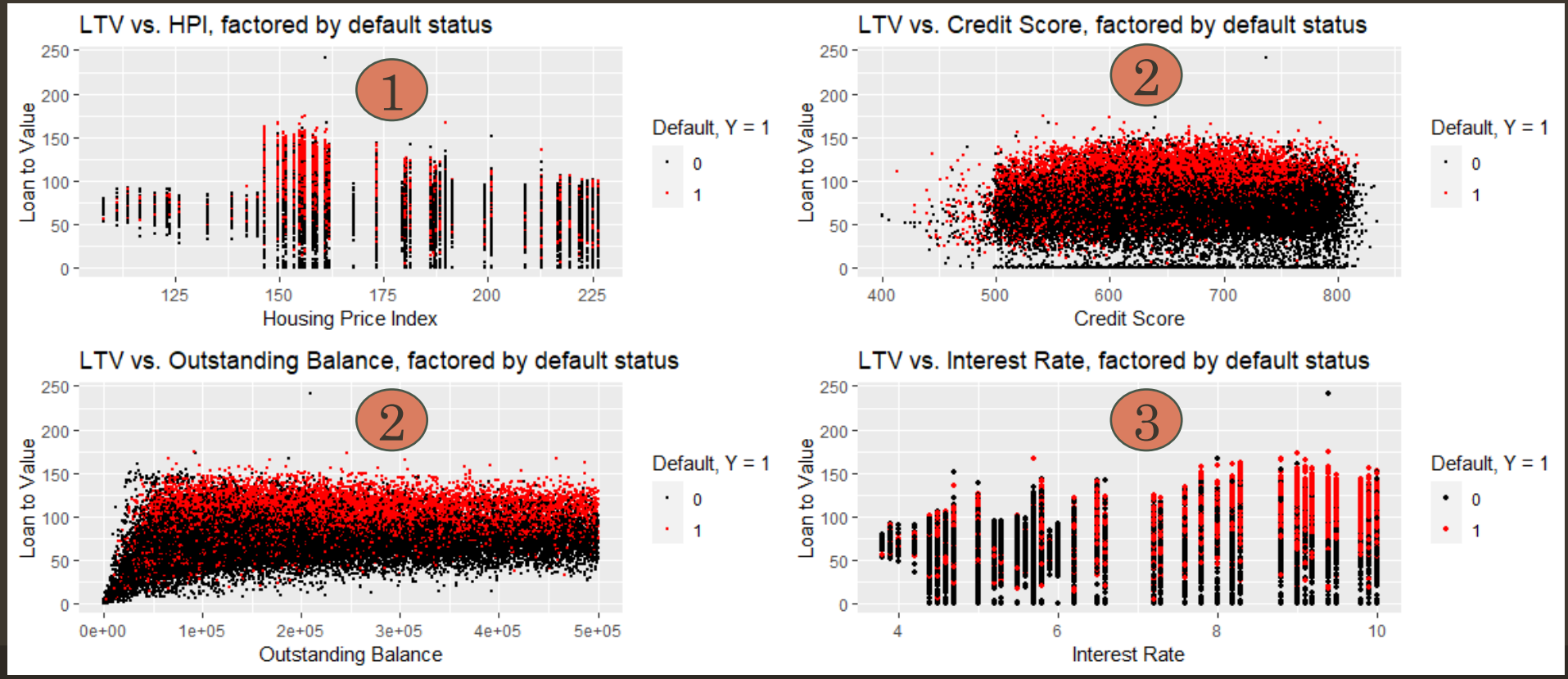


Loan-to-Value ratio exhibits strong default signals at a level of >100 across several different segmentations

1. Plotted against Housing Price Index, there is a large cluster of default observations around housing price index between 150-175 with an LTV above 100.

2. Most defaults occur above an LTV of 100, regardless of credit score or outstanding balance. Different values for credit score and outstanding balance do not appear to differentiate default behavior

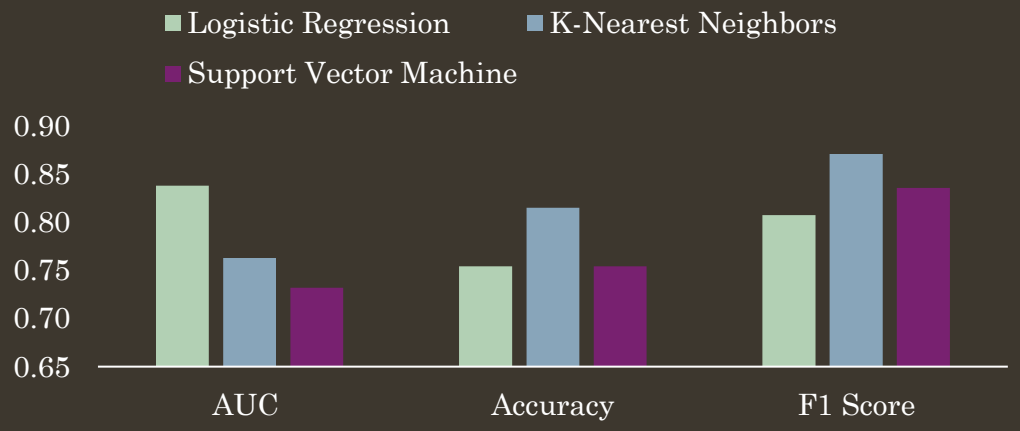
3. Interest rate appears to be a differentiator for default behavior, even against LTV. Most consumer defaults occur above a 7% interest rate and an LTV of 100



Alternative modeling methods outperform logistic regression in two of three performance measures for PD modeling

- For each model type in both PD and LaD modeling, three measures of model performance are provided.
  - For PD modeling, the measures were focused on how well the model distinguished between defaults/non defaults. For LaD modeling, the measures focused on how close the loss estimate was to the actual loss amount. Any tuning carried out is detailed in the appendix
- K-nearest neighbors scored the best in 2 out of 3 areas for PD modeling, while the random forest scored the best in all 3 of the performance measures.
  - K-nearest neighbors liekly outperformed because defaults tend to happen around clusters of observations – high UER, low HPI, and low GDP.
  - Random forests likely overperformed due to its ability to more easily handle complex interactions and fewer tuning inputs. XGBoost *may* outperformed if the parameters were tuned differently, which might serve as an extension of this analysis. Both of these methods did end up significantly overperforming the linear regression model

Probability of Default modeling results
Separated by Accuracy Measure



Model Type	AUC	Accuracy	F1 Score
Logistic Regression	**0.838**	0.754	0.807
K-Nearest Neighbors	0.763	**0.815**	**0.871**
Support Vector Machine	0.752	0.754	0.836

Loss at Default modeling results
Separated by Accuracy Measure

Model Type	MAE	MSE	RMSE
Linear Regression	30,783	579,874,418	50,792
Random Forest	1,587	107,994,523	9,490
xGBoost	2,483	111,837,689	10,575

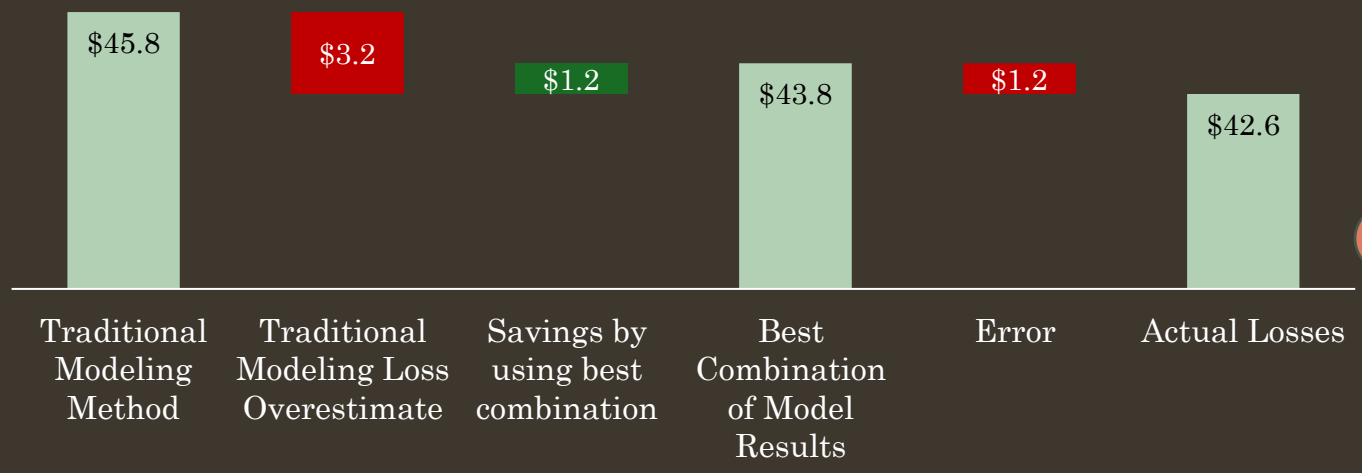
Utilizing the optimal PD/LaD model combination reduces loss overestimate by 38%, freeing up significant capital for alternative uses

1. Taking each PD model and combining it with each LaD model, the combination that produces an expected loss that is the closest to actuals are the k-nearest neighbors (KNN) PD model and the random forest LaD model, with a deviation of only 2.8% over actual losses

2. Using the optimal combination of PD/LaD model reduces the loss overestimate of the baseline logistic/linear regression models by 38%
   - The baseline logistic/linear combination estimated losses of $45.8M, which is $3.2M over the actual losses in the dataset.
   - The optimal combination of KNN/RF estimated $43.8M in losses, which his only $1.2M over the actual losses in the dataset
   - Compared to the baseline linear/logistic regression results, this is a 38% reduction in a loss overestimated – freeing up $1.2M in this dataset to be deployed to alternative uses.

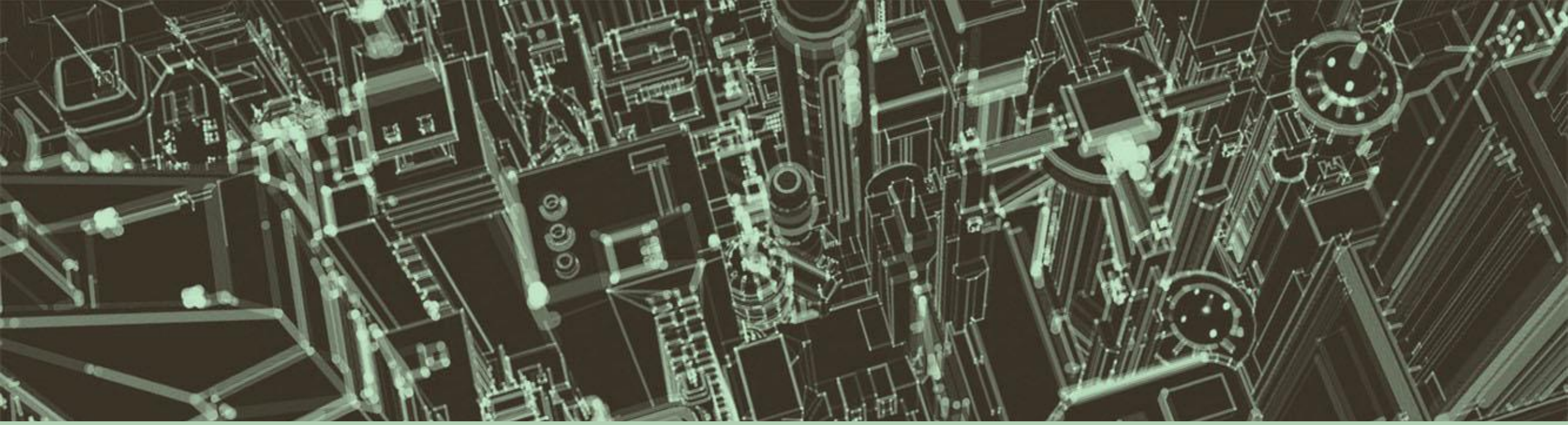
2 Expected Loss Forecast Comparison vs. Actual Losses in the dataset

Walk from Traditional modeling to best combination to actuals.



Exhaustive combinations of all model types and their expected loss calculations vs. Actual losses

Probability of Default Model	Loss at Default Model	Expected Loss	% Error
Logistic Regression	Linear Regression	$45,839,575	7.6%
Logistic Regression	Random Forest	$45,283,538	6.3%
Logistic Regression	xGBoost	$45,064,573	5.8%
K-Nearest Neighbors	Linear Regression	$46,940,201	10.2%
1 K-Nearest Neighbors	Random Forest	$43,783,246	2.8%
K-Nearest Neighbors	xGBoost	$45,987,601	8.0%
Support Vector Machine	Linear Regression	$40,277,781	-5.4%
Support Vector Machine	Random Forest	$44,278,220	4.0%
Support Vector Machine	xGBoost	$45,444,917	6.7%
Actual		**$42,588,595**	**0.0%**

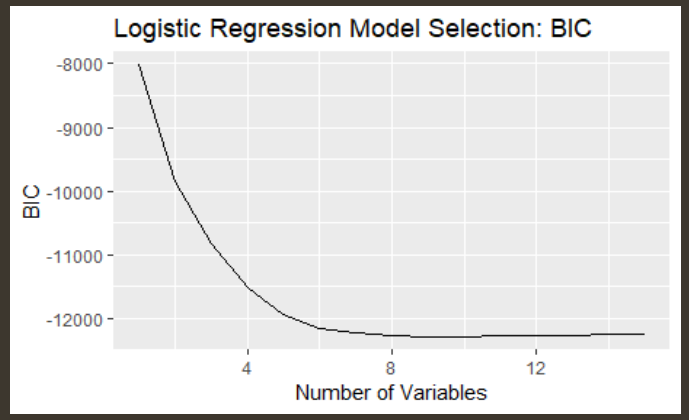


Appendix

Appendix I: PD Model Development

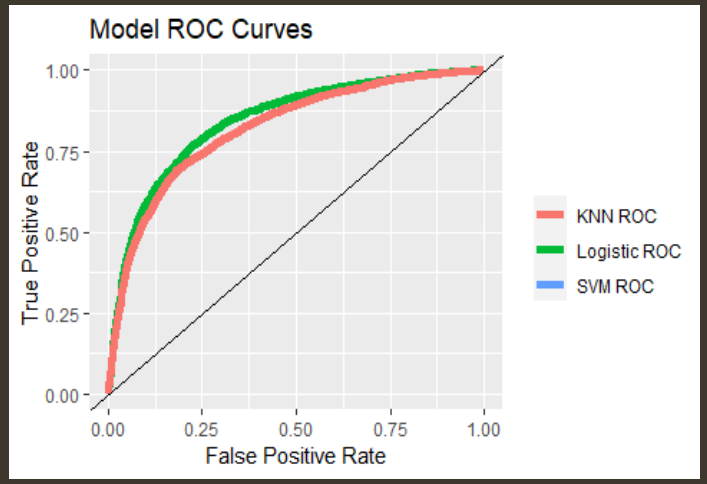
Logistic Regression Variable Selection

- Utilized best subset selection to identify the best logistic regression. Final model included 9 variables with LTV, interest rate, GDP, UER, Investor, Balance, FICO, HPI, and HPI time



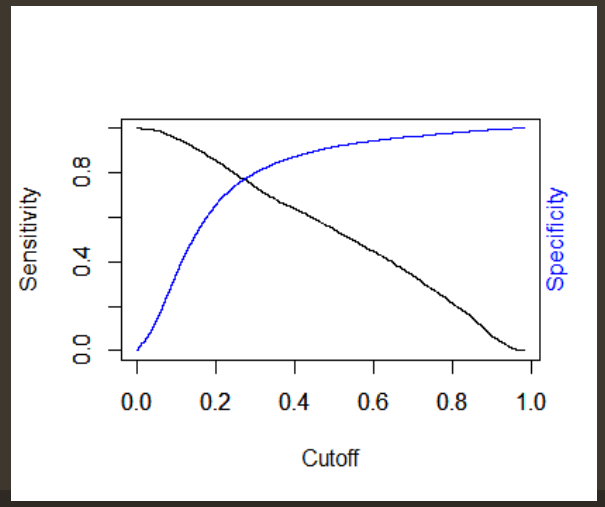
ROC Curves for each model type

- ROC curves measure the true positive rate against the false positive rate. The further skewed left the curve, the better the model is at classifying defaults over a random guess



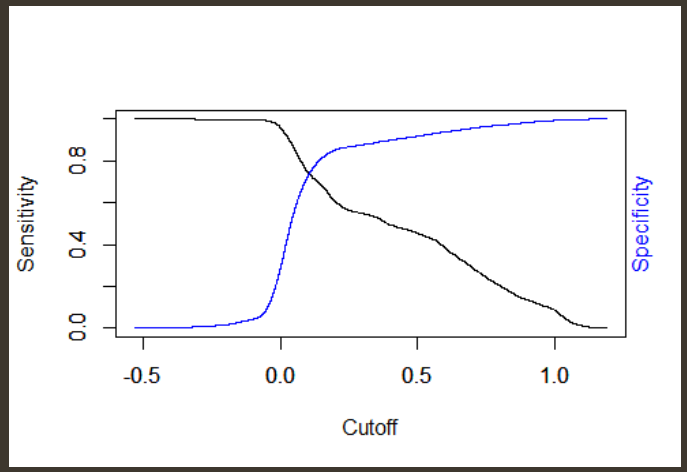
Logistic Regression Cutoff Value

- The optimal cutoff value for a default prediction is where sensitivity = specificity, for the logistic regression that value is 0.25.



SVM Cutoff Value

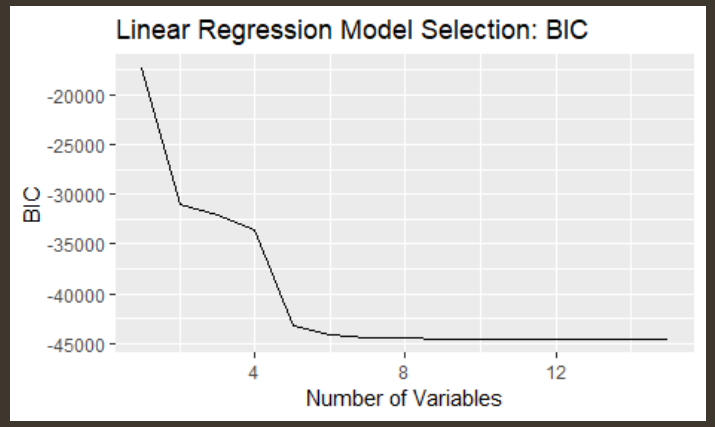
- The optimal cutoff value for a default prediction is where sensitivity = specificity, for the SVM that value is 0.1.



Appendix II: LaD Model Development

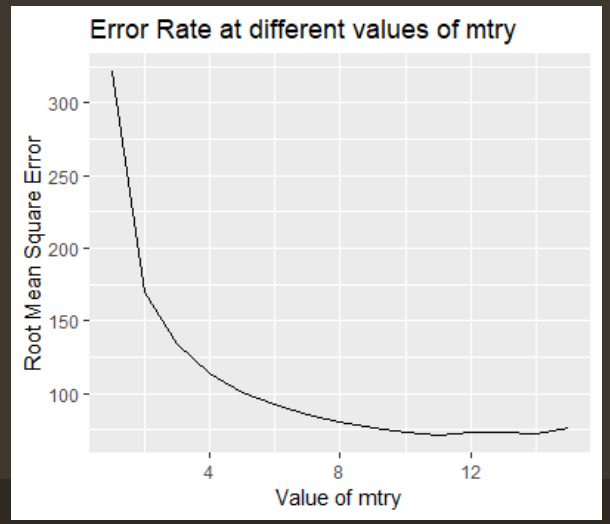
Linear Regression Variable Selection

- Utilized best subset selection to identify the best linear regression (the smallest BIC value).
- Final model included 11 variables with LTV, interest rate, GDP, UER, Investor, Balance, FICO, HPI, and HPI time, remaining loan, and LTV_time



Random Forest Node Size (mtry)

- Mtry is the number of variables per split in a random forest. The optimal value is the lowest error (Root mean square error).
- The optimal value for mtry for this dataset is 11
- We kept the number of trees constant at 500, since computational time was not a problem



xGBoost Tuning Parameters

- XGBoost model requires several different parameters to be tuned. We used random selection to try several different parameters
- The optimal parameters were found in the shaded row, with an eta of 0.259, depth of 5, and a gamma of 6.45.

eta	max_depth	gamma	colsample_bytree	min_child_weight	subsample	nrounds	RMSE	MAE
0.464	7	1.783	0.416	19	0.339	70	25488	12376
0.403	8	2.733	0.541	19	0.793	139	19417	8286
0.160	8	5.653	0.339	5	0.412	297	17800	7503
0.288	5	9.922	0.500	17	0.337	317	17322	8033
0.131	4	0.365	0.342	17	0.829	374	14894	6423
0.332	2	4.667	0.685	2	0.380	395	17545	9265
0.583	3	0.320	0.620	15	0.682	475	15102	6835
0.259	5	6.454	0.595	12	0.529	559	10575	2483
0.578	2	4.195	0.571	15	0.311	601	22766	12776
0.063	7	9.423	0.380	12	0.476	604	15754	5593
0.428	9	5.811	0.629	8	0.741	627	19607	7922