Mortgage Climate Risk Report

Corporate Sponsor: Freddie Mac
Faculty Advisor: Professor Clifford Rossi

Mar 2023

Robert H. Smith School of Business

Meet the Team

Kanglong Li

Jiaxin Liu

Jiahao Zhang

Lin Zheng

Mengxin Xie

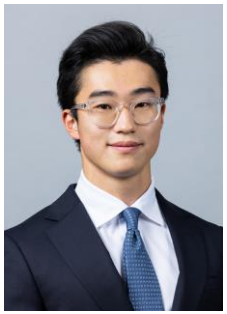
Fengyu Yuan

Yongkang Zheng

Haoruo Yang

Ning Xu

Siddhesh Gore



Qiuyu Xue

Project Agenda

1

Overview and Roadmap

Why Climate Risk?

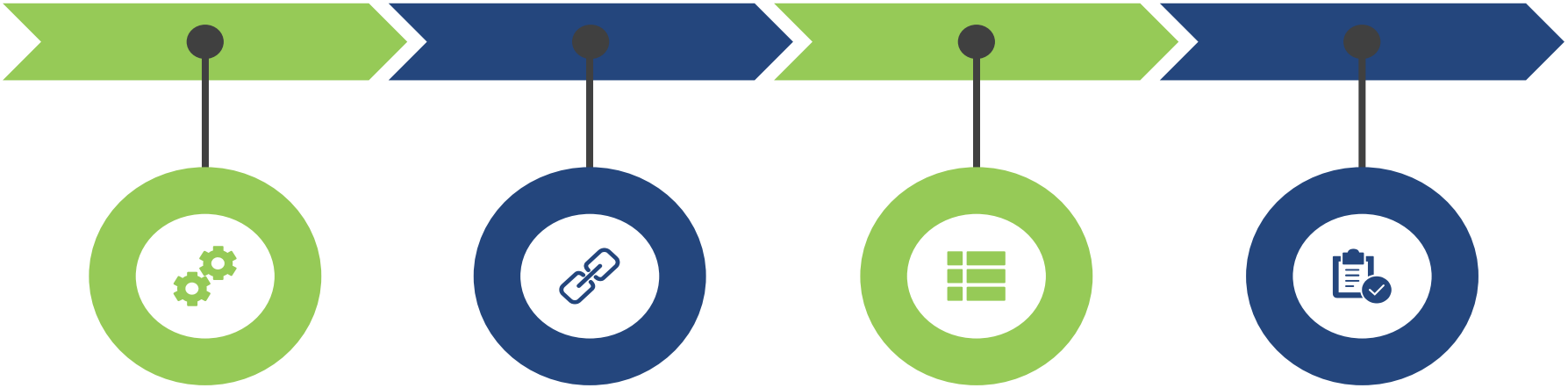
- **Rising Topic:** Climate Change continues to be at the top of the global agenda, and the risks associated with natural hazards should not be overlooked.

- **Regulatory Changes:** Regulators are reacting to climate risks by increasing their expectations of financial institutions through more information disclosure and impact assessments.

- **Lack of Studies:** Fewer studies have been done on climate risks compared to other risks

Our Goals

With the effects of climate change becoming increasingly pronounced, it is important not to overlook the effects of climate risk poses on the mortgage market. More importantly, are there predictability characteristics in loan applications that can differentiate between low and high climate risk loans.



Data
Obtain useful data related to climate risk on the mortgage market

Visualization
Provide visualization of climate risk characteristics across the U.S.

Prediction
Accurately predict climate risk exposure based on loan attributes

Inference
Interpret relationship of loan attributes on climate risk, can be used for business decisions

Results

Data	Finalized over **13 million** useable mortgage data on climate risk, and used **1 million** observations for data analysis
Adverse Selection	Freddie Mac is **not subject** to any **adverse selection** exposed to climate risk while purchasing mortgages
Racial Inequality	**No particular race** or **ethnicity** is exposed to more climate risk
Protected Veteran	**VA Loans** have a statistically significant **lower odds** of being in a high climate risk area
Pinpoint Predictions	The loan **TRACT-Level** variables have strong predictive power on climate risks, can predict with over **99%** accuracy

2

Data Gathering and Data Manipulation

Data is the most important foundation of any analysis, a skewed data set would lead to inaccuracy in predictions and inferences

Data Gathering

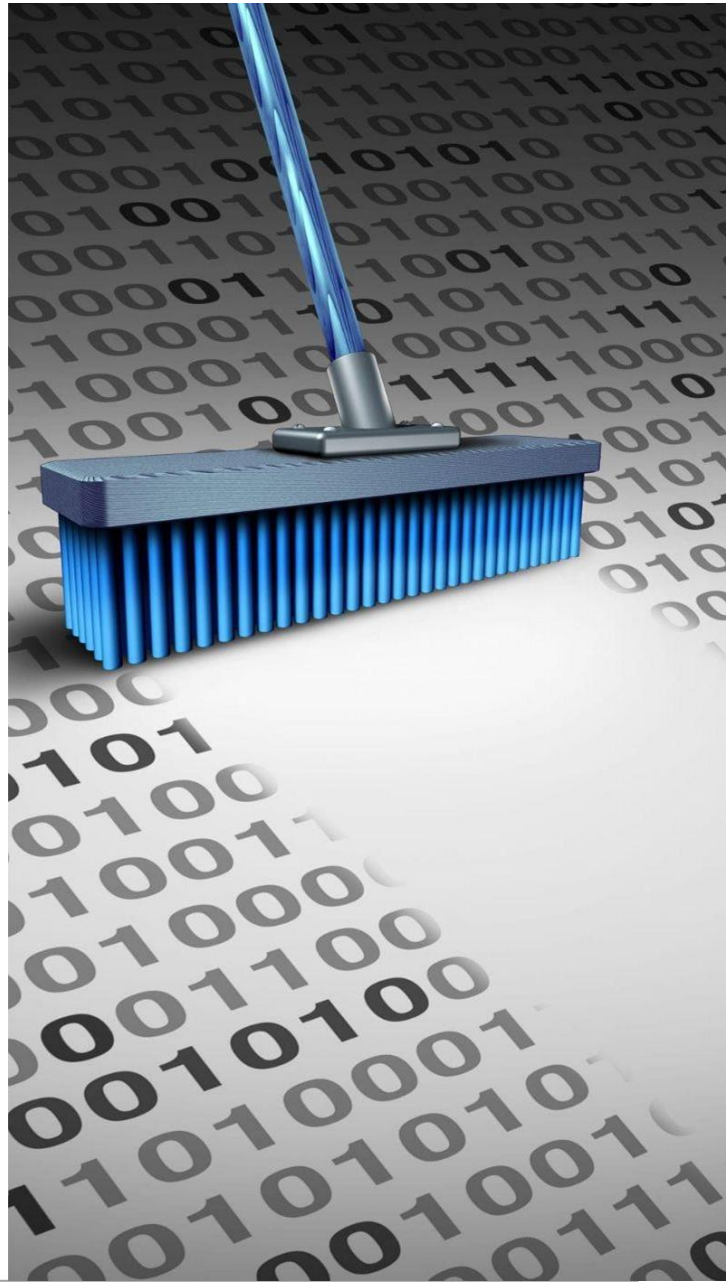
- **Loan Application Data:** Obtained over 26 million 2021 loan applications data from the FFIEC HMDA website.

- **Risk Index:** Obtained over 70 thousand 2021 tract-level risk indices from the FEMA NRI website.

- **Geographical Shapefile:** Obtained over 70 thousand 2021 tract-level Shapefile from the FEMA NRI website.

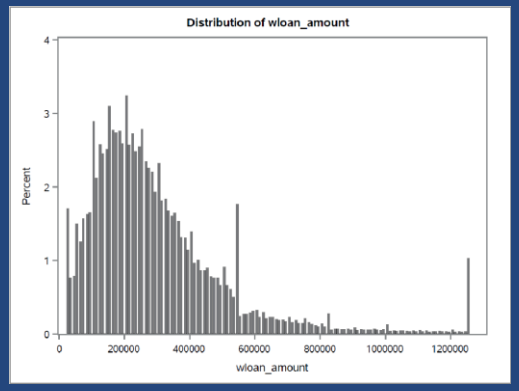
Data Cleaning and Manipulation

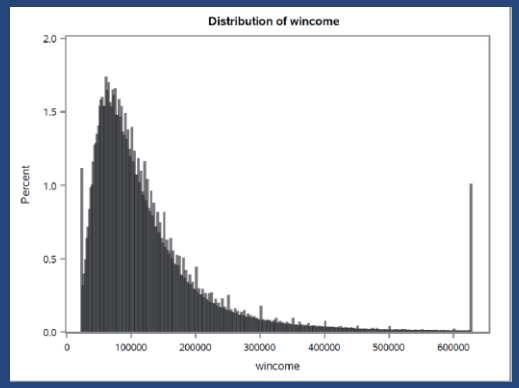
- **Setting Constraints:** Only kept the data that is first lien and single family
- **Missing Data:** Deleting observations where important attributes are missing.
- **Standardize Units/Range:** Manipulate the attributes of different variables to make them consistent and useable for analysis.
- **Outliers:** First Winsorized then trimmed the data that lies above the 99th or below the 1st percentile.
- **Merging Data:** Mapped the NRI data to the cleaned HMDA data, each application is matched to its respective census tract.
- **Dummy Variables:** Created dummy variables for key borrower's attributes
- **Selecting Data:** Randomly selected 1 million observations to use for data analysis.

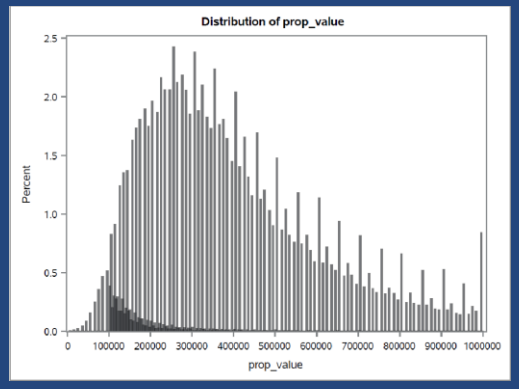


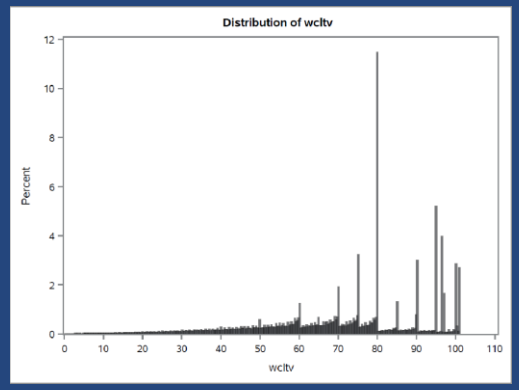
Exhibit

Univariate Plots of Continuous Variables



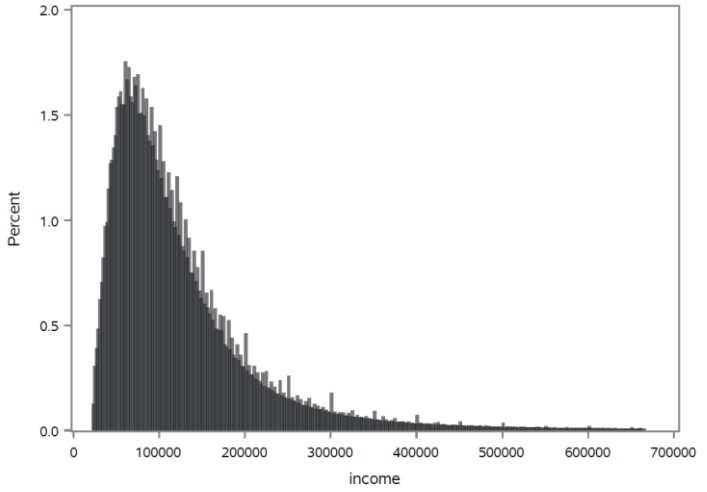






Why Trimming

While Winsorizing may preserve the data to a fuller extent, preliminary univariate analysis shows it may cause skewed data, and since the dataset still has over 13 million observations, the trimming method is chosen.



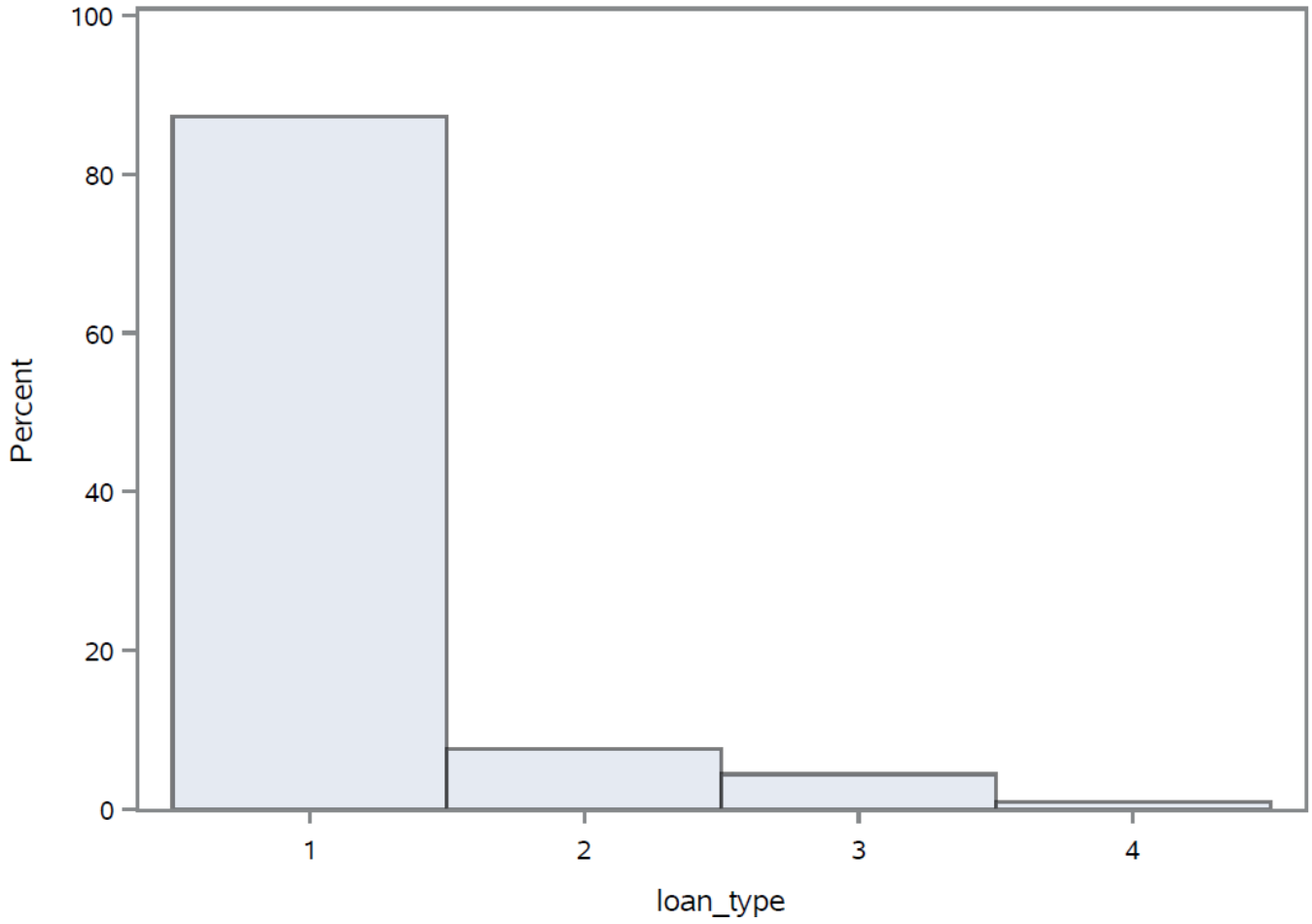
3

Data Visualization

- Univariate Analysis
-
- Bivariate Analysis

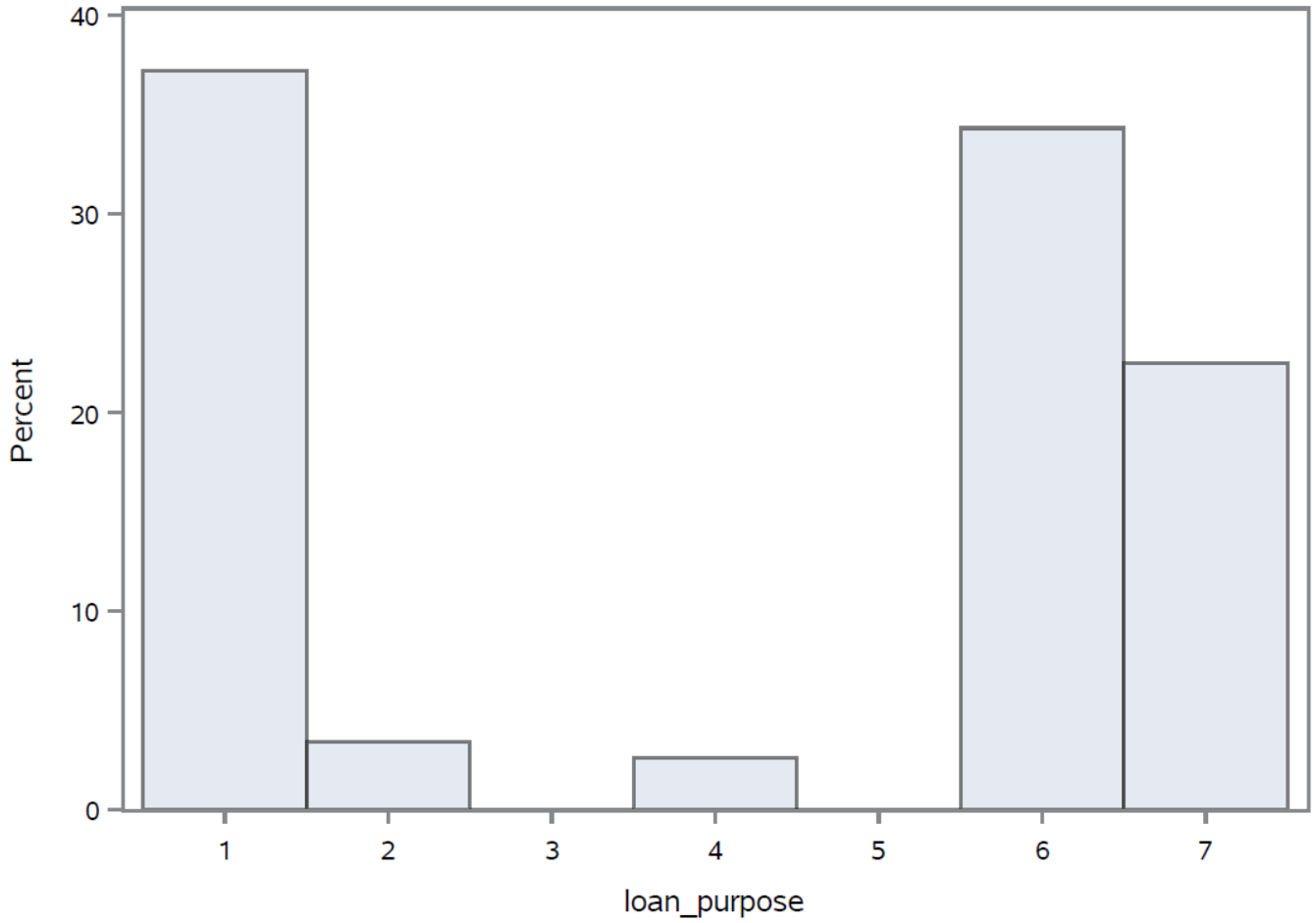
Interactive Dashboard

Univariate Analysis - Loan Type



1 – Conventional 2 – FHA 3 – VA 4 – RHS or FSA

Univariate Analysis - Loan Purpose



1 – Purchase

2 - Improvement

3 – Not Applicable

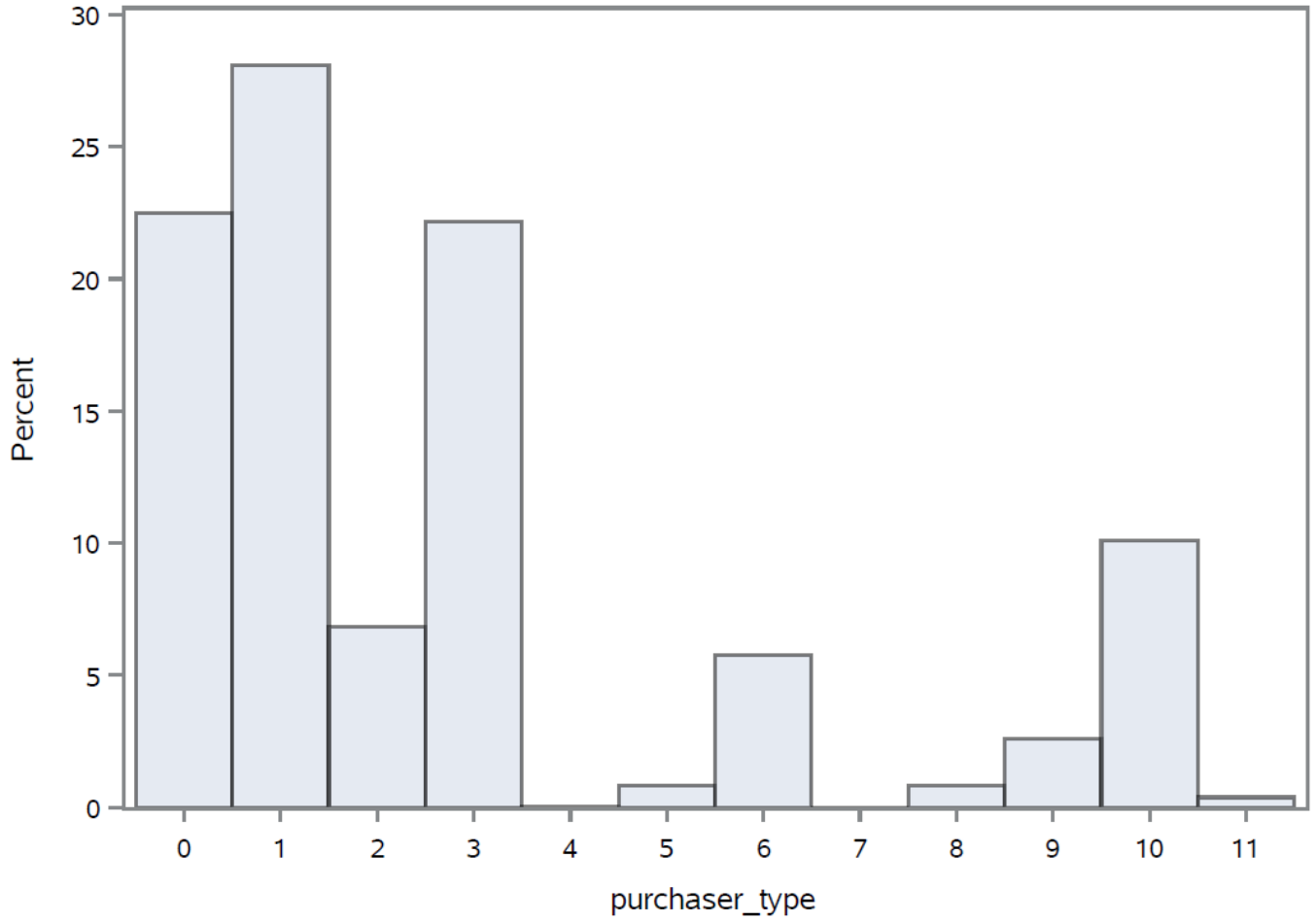
4 – Other

5 – Not Applicable

6 –Refinancing

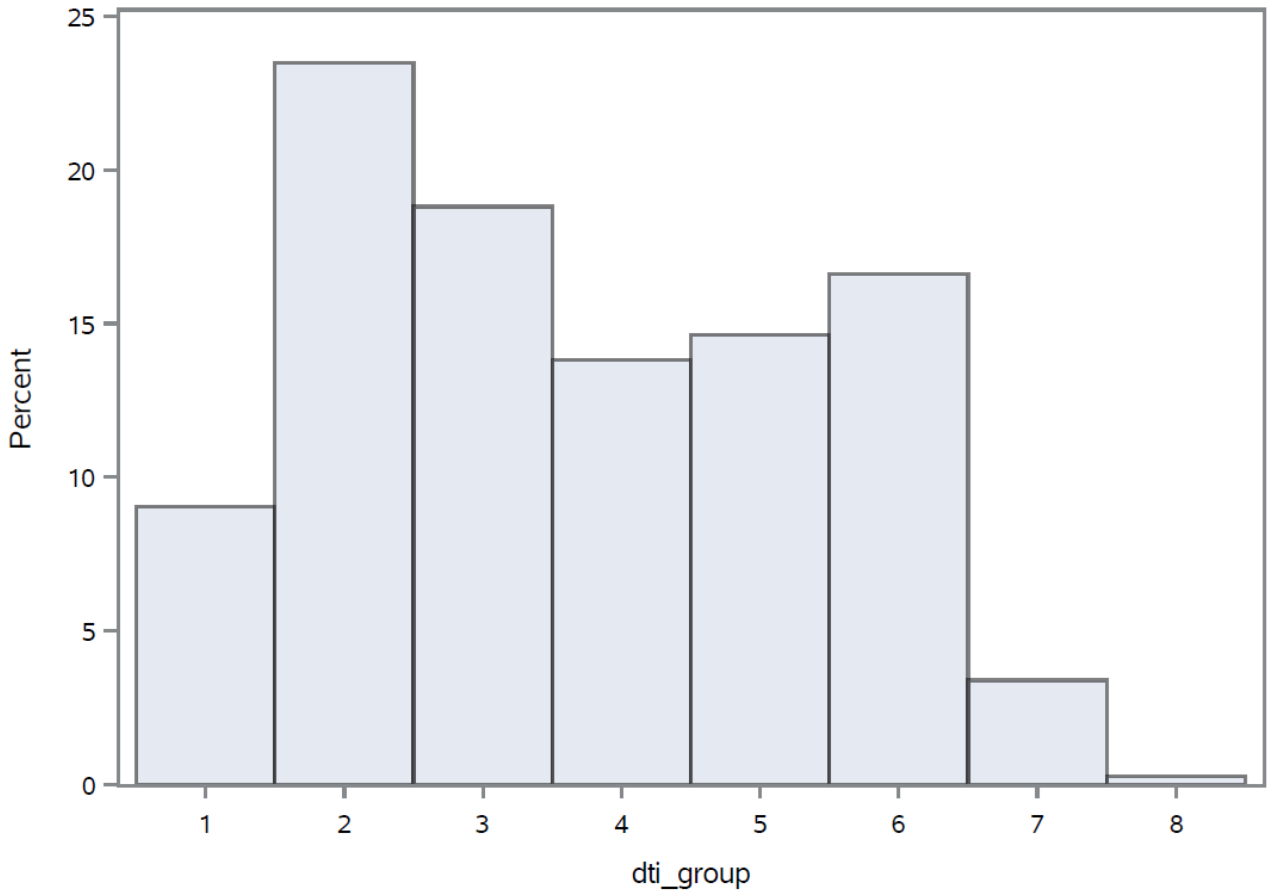
7 – Cash-out Refinancing

Univariate Analysis - Purchaser Type



- 0 – Not Applicable
- 1 – Fannie Mae
- 2 – Ginnie Mae
- 3 – Freddie Mac
- 4 – Farmer Mac
- 5 – Pvt. securitizer
- 6 – Banks
- 7 – Not Applicable
- 8 – Affiliate Inst
- 9 – Other
- 10 – Credit Union
- 11 – Life Insurance

Univariate Analysis - Debt to Income Group



1 – <20%

2 – 20%-30%

3 – 30%-36%

4 – 36%-39%

5 – 40%-43%

6 – 44%-49%

7 – 50%-60%

8 – >60%

Bivariate Analysis - Application Analysis

Rating	Application	Loans	Denied
Very High	2.48%	2.42%	16.13%
Relatively High	13.00%	12.89%	14.89%
Relatively Moderate	25.32%	25.25%	14.39%
Relatively Low	33.31%	33.35%	14.07%
Very Low	25.89%	26.09%	13.50%

Total Risk Rating

Rating	Application	Loans	Denied
Very High	1.19%	1.20%	13.46%
Relatively High	9.40%	9.49%	13.35%
Relatively Moderate	18.09%	18.24%	13.46%
Relatively Low	16.72%	16.68%	14.34%
Very Low	0.49%	0.46%	19.18%
No Expected Annual Losses	54.11%	53.92%	14.46%

Cold Wave Risk Rating

Rating	Application	Loans	Denied
Very High	1.97%	1.96%	14.62%
Relatively High	5.43%	5.38%	14.86%
Relatively Moderate	9.65%	9.58%	14.79%
Relatively Low	18.79%	18.75%	14.35%
Very Low	15.53%	15.60%	13.80%
No Expected Annual Losses	48.63%	48.73%	13.98%

Wildfire Risk Rating

Rating	Application	Loans	Denied
Very High	0.28%	0.28%	14.89%
Relatively High	1.72%	1.69%	15.49%
Relatively Moderate	6.68%	6.62%	14.99%
Relatively Low	19.09%	19.12%	14.03%
Very Low	34.11%	34.14%	14.10%
No Expected Annual Losses	38.12%	38.15%	14.08%

Drought Risk Rating

Bivariate Analysis - Race & Ethnicity

Rating	Hispanic or Latino	White (Not Hispanic or Latino)	Asian (Not Hispanic or Latino)	Black or African American (Not Hispanic or Latino)	American Indian or Alaska Native (Not Hispanic or Latino)	Native Hawaiian (Not Hispanic or Latino)
Very High	3.20%	2.25%	3.05%	1.94%	4.28%	4.95%
Relatively High	14.99%	12.09%	17.87%	9.97%	19.89%	17.94%
Relatively Moderate	28.70%	24.89%	26.36%	21.90%	28.56%	28.38%
Relatively Low	31.89%	34.46%	28.24%	34.87%	28.68%	32.39%
Very Low	21.22%	26.30%	24.48%	31.33%	18.60%	16.34%

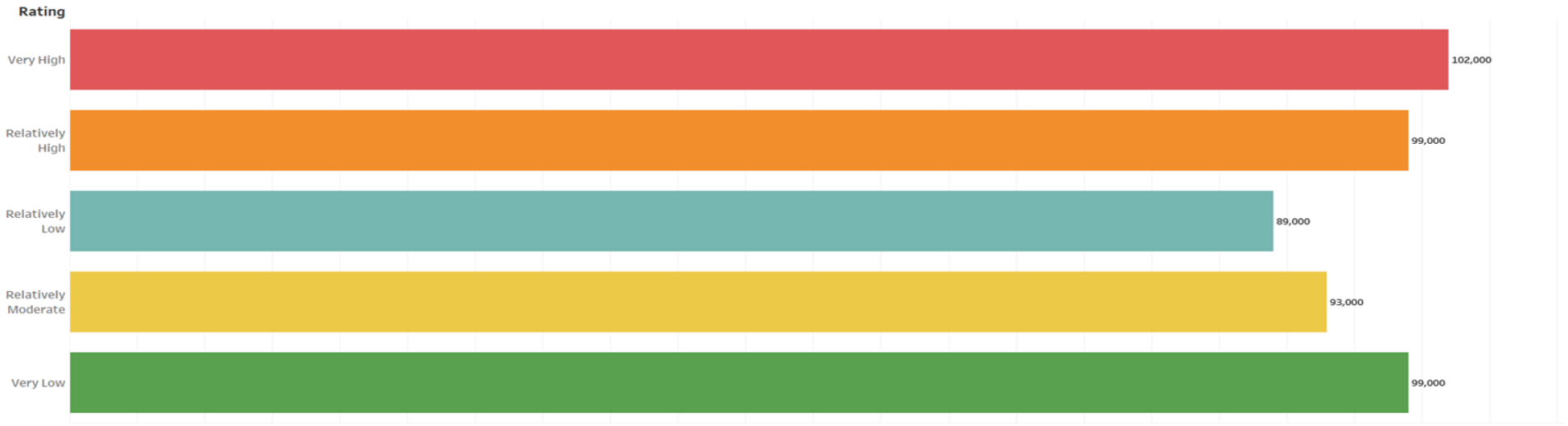
Total Risk Rating

Rating	Hispanic or Latino	White (Not Hispanic or Latino)	Asian (Not Hispanic or Latino)	Black or African American (Not Hispanic or Latino)	American Indian or Alaska Native (Not Hispanic or Latino)	Native Hawaiian (Not Hispanic or Latino)
Very High	3.16%	2.41%	4.16%	2.64%	2.81%	4.19%
Relatively High	14.05%	14.04%	20.77%	10.48%	19.42%	20.75%
Relatively Moderate	29.60%	29.19%	27.97%	25.65%	33.62%	28.36%
Relatively Low	32.83%	34.96%	29.09%	34.52%	28.43%	32.86%
Very Low	20.36%	19.39%	18.01%	26.71%	15.72%	13.85%

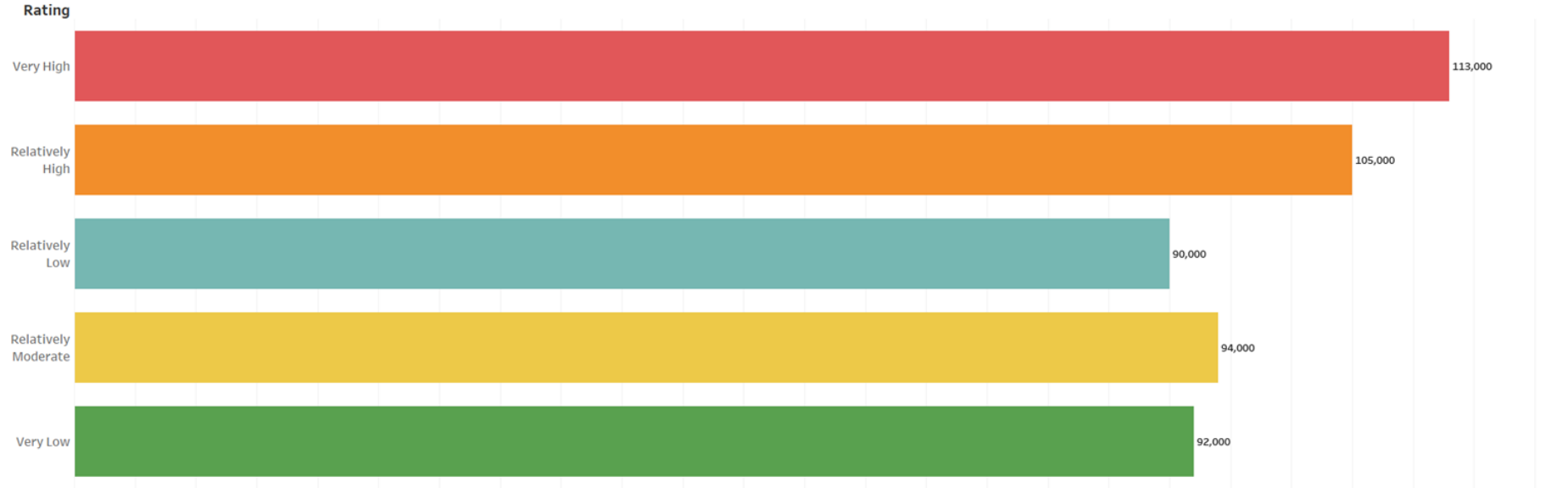
Total EAL Risk Rating

Bivariate Analysis - Median Family Income

Total Risk Rating



Total EAL Risk Rating



Interactive Dashboard

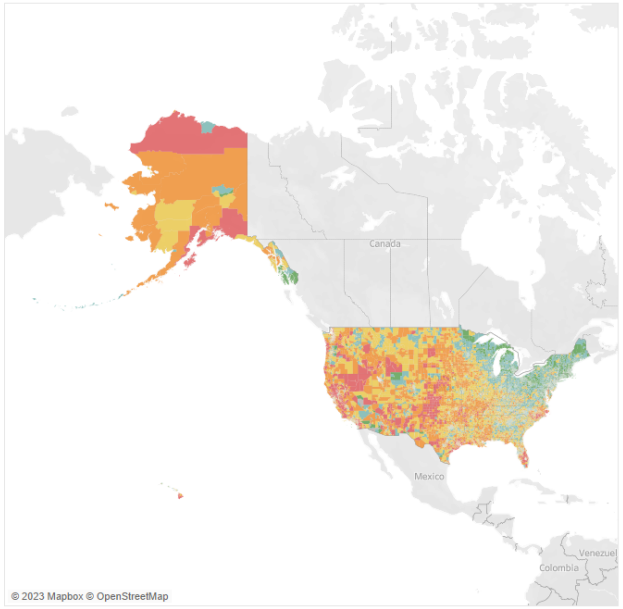
Interactive Dashboard

Mortgage Climate Risk Report

Hazards: Total Risk | Risk or Loss: National Risk Index Rating | Purchaser ..: All | State: All | County: All

Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low

Rating Map (All US Census Tract Based)

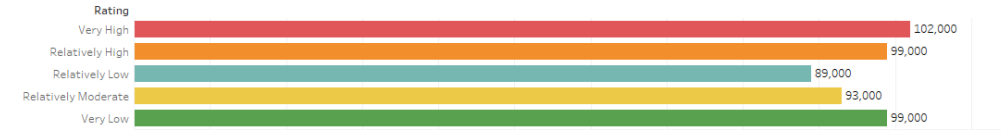


Rating	Application	Loans	Denied
Very High	2.48%	2.42%	16.13%
Relatively High	13.00%	12.89%	14.89%
Relatively Moderate	25.32%	25.25%	14.39%
Relatively Low	33.31%	33.35%	14.07%
Very Low	25.89%	26.09%	13.50%

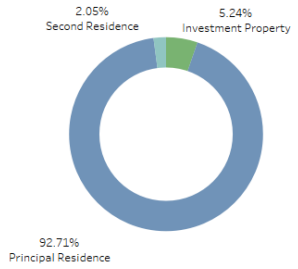
Purchaser Type

Rating	Freddie Mac	Fannie Mae	Ginnie Mae	Farmer Mac	Banks & Credit ..	Other
Very High	2.41%	2.52%	2.48%	0.68%	2.34%	2.55%
Relatively High	12.52%	13.34%	13.24%	5.88%	12.91%	13.02%
Relatively Moderate	24.10%	25.57%	27.18%	35.75%	25.98%	25.24%
Relatively Low	33.56%	33.05%	34.97%	50.90%	32.59%	33.36%
Very Low	27.41%	25.53%	22.13%	6.79%	26.20%	25.82%

Median Family Income



Occupancy Type



Race & Ethnicity

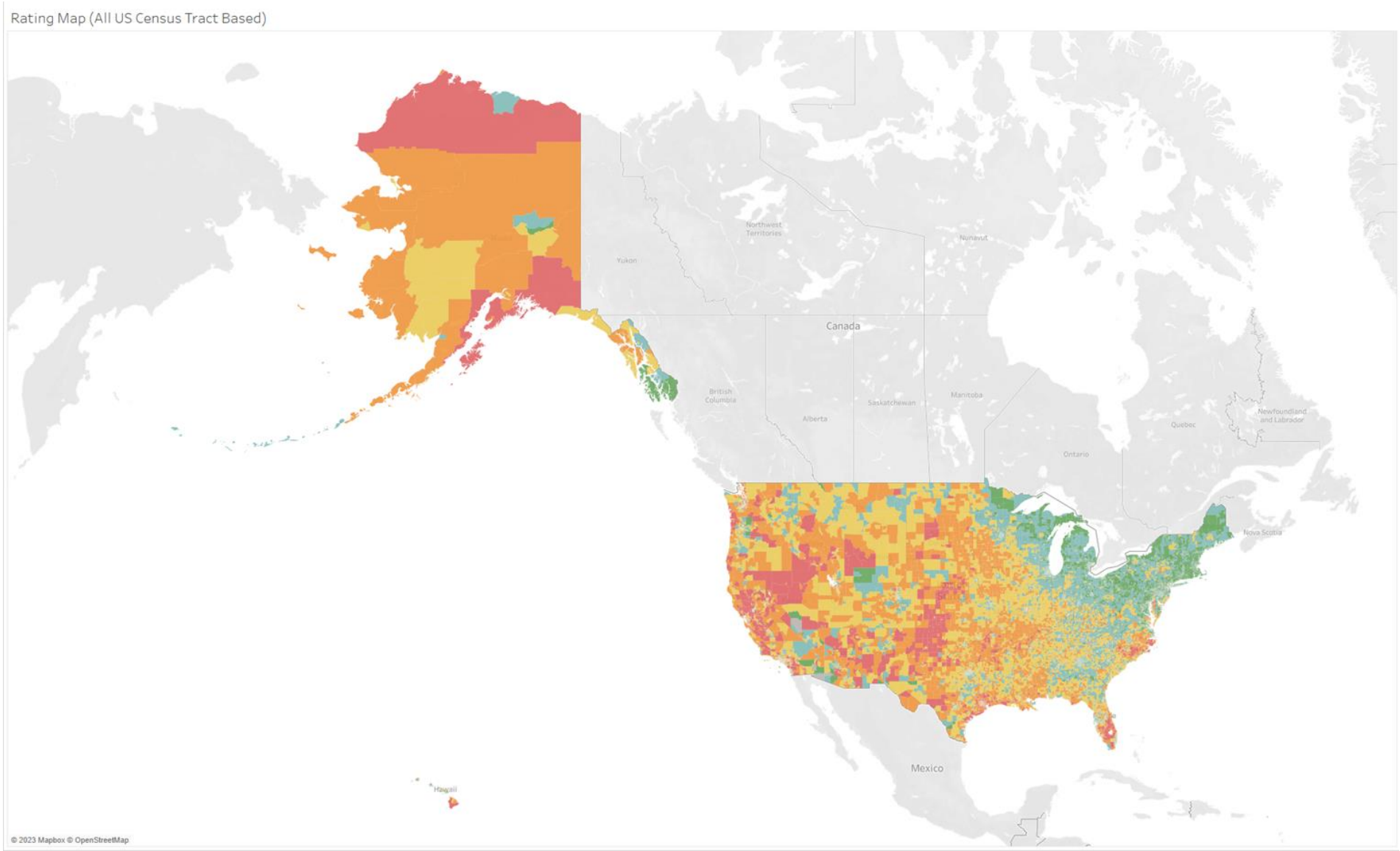
Rating	Hispanic or Latino	White (Not Hispanic or Latino)	Asian (Not Hispanic or Latino)	Black or African American (Not Hispanic or Latino)	American Indian or Alaska Native (Not Hispanic or La..	Native Hawaiian (Not Hispanic or Latino)
Very High	3.20%	2.25%	3.05%	1.94%	4.28%	4.95%
Relatively High	14.99%	12.09%	17.87%	9.97%	19.89%	17.94%
Relatively Moderate	28.70%	24.89%	26.36%	21.90%	28.56%	28.38%
Relatively Low	31.89%	34.46%	28.24%	34.87%	28.68%	32.39%
Very Low	21.22%	26.30%	24.48%	31.33%	18.60%	16.34%

Source: FFIEC, 2021 Home Mortgage Disclosure Act (HMDA) Data
FEMA, 2021 National Risk Index Data

[Dashboard Link]

Interactive Dashboard - Total Risk Rating Map

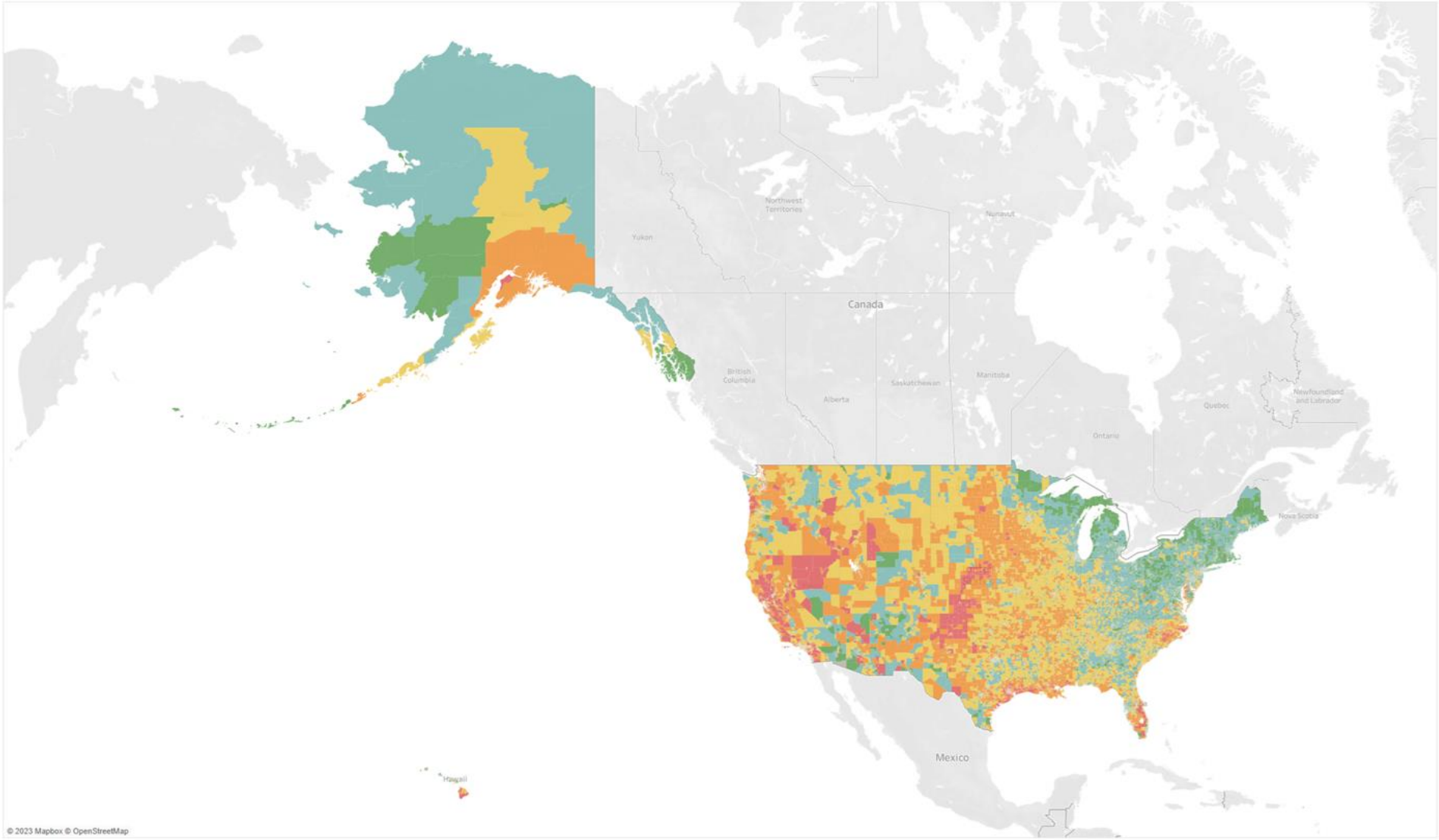
Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating



Interactive Dashboard - Total Risk EAL Rating Map

Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating

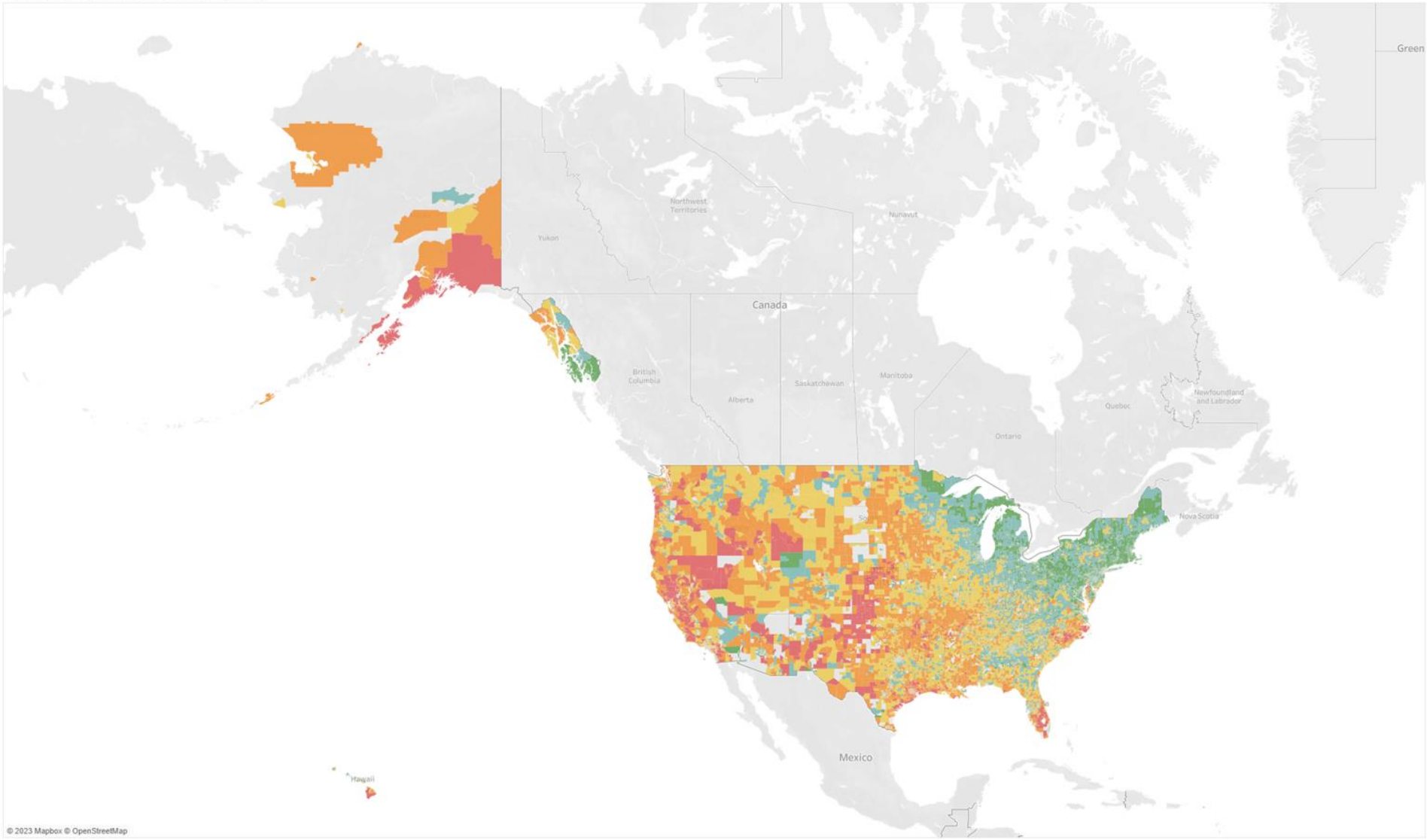
Rating Map (All US Census Tract Based)



Interactive Dashboard - Total Risk Rating Map (Freddie Mac)

Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating

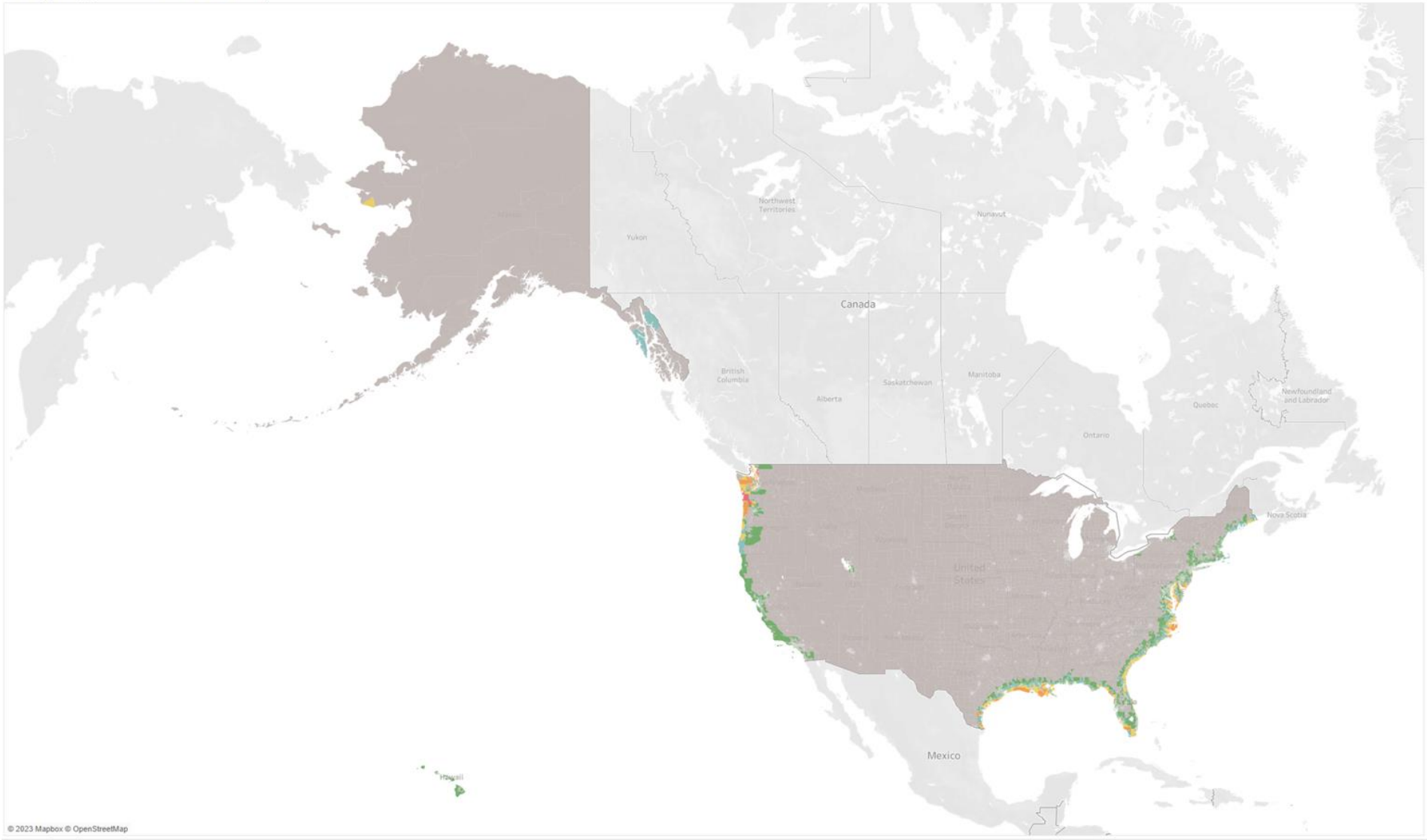
Rating Map (All US Census Tract Based)



Interactive Dashboard - Coastal Flooding Risk Rating Map

Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating

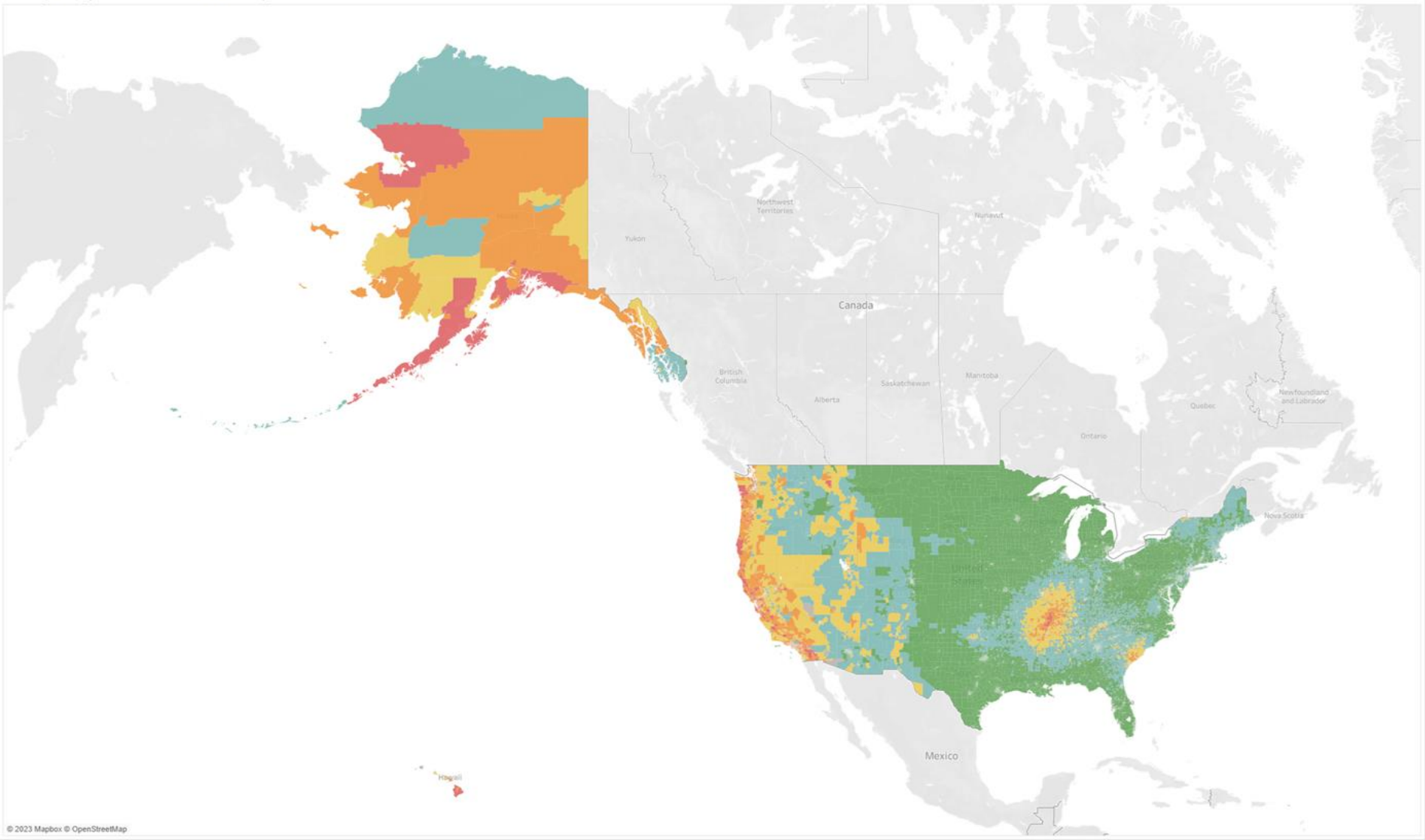
Rating Map (All US Census Tract Based)



Interactive Dashboard - Earthquake Risk Rating Map

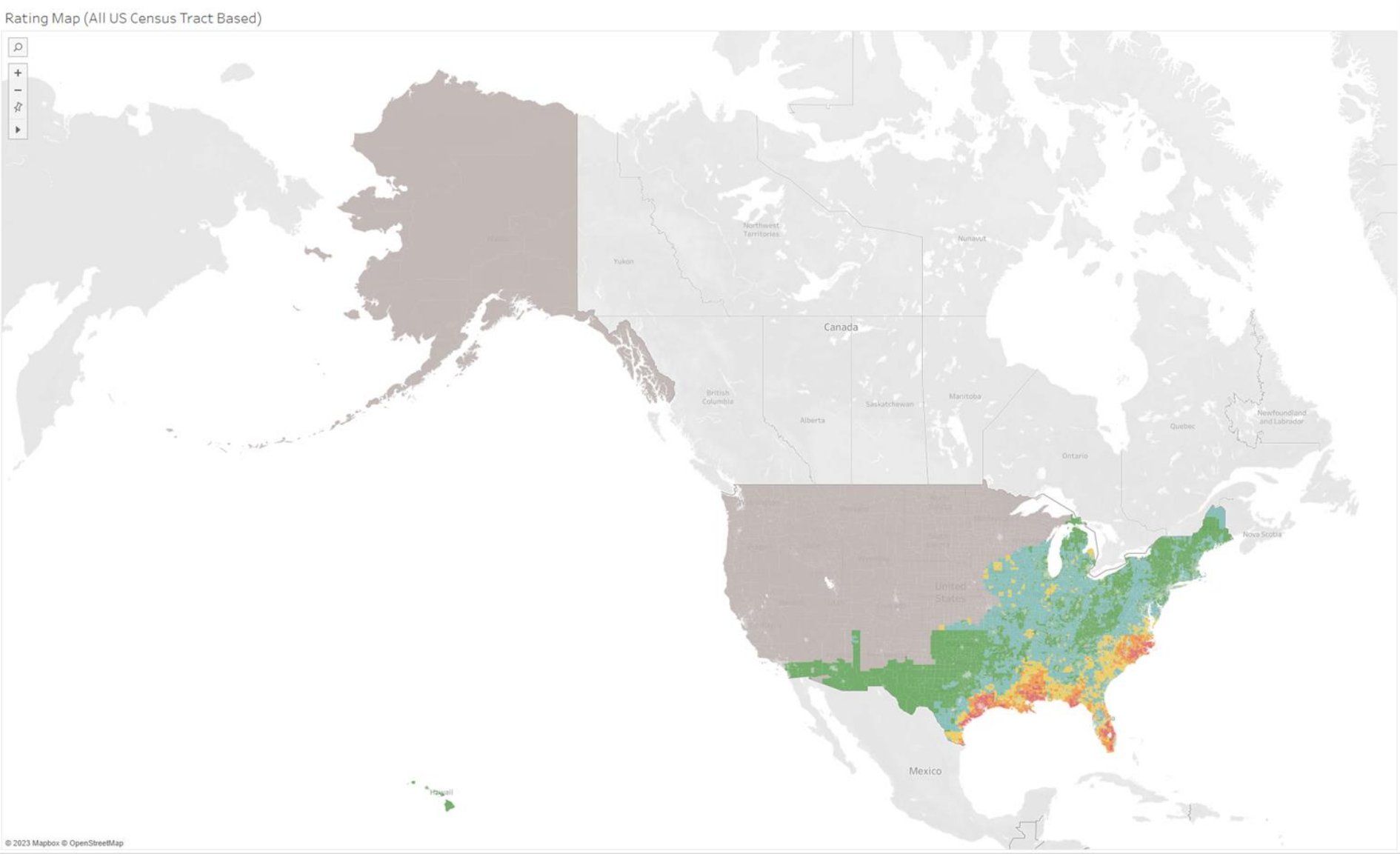
Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating

Rating Map (All US Census Tract Based)



Interactive Dashboard - Hurricane Risk Rating Map

Very High | Relatively High | Relatively Moderate | Relatively Low | Very Low | No Rating



4.1

Model Building and Validation

Formalize models that provide feature selection, have predictive power and enable inference

- Target Variables
- Features
- Candidate Models
- Feature Selection
- Model Results
- Champion Model
- Model Inference
- Further Refinement

Target Variables

Overall Risk/EAL Rating

Very High	Relatively High	Relatively Moderate	Relatively Low	Very Low

Overall Risk/EAL Score (Continuous Numerical Variable, from 0 to 100)

0 — 100

Relationship between Risk and Expected Annual Loss

$$\text{Risk} = \frac{\text{Expected Annual Loss} * \text{Social Vulnerability}}{\text{Community Resilience}}$$

RR (Risk Rating)	Risk Rating = very high, relatively high then RR = 1
RR Bisection	Risk Score > = 50% percentile then RS Bisection = 1
RR Reverse	Risk Rating = very low, relatively low then RR Reverse = 1
ER (EAL Rating)	EAL Rating = very high, relatively high then ER = 1
ES Bisection	EAL Score >= 50% percentile then ES Bisection = 1
ER Reverse	EAL Rating = very low, relatively low then ER Reverse = 1

Features

Tract-Level	Borrower's Characteristics	Loan Characteristics	Others
tract population	Age	Loan Amount	Property Value
tract minority population percent	Debt-to-Income Ratio	Loan Term	Total Units
ffiec msa md median family income	Combined Loan-to-Value	Loan Type	Hoepa
tract to msa income percentage	Income	Loan Purpose	Construction Method
tract owner occupied units	Number of Borrowers		Occupancy Type
tract one to four family homes	Race		Purchaser Type
tract median age of housing unit	Ethnicity		

Candidate Models

To fully test the relationship between the features and the target variables, a variety of statistical analysis were conducted

- Logistic Regression
- Ridge Regression
- Lasso Regression
- Elastic Net
- Ordered Logistic
- Multinomial Logistic
- Random Forest
- Naive Bayes Classifier
- eXtreme Gradient Boosting
- Support Vector Machine
- Neural Network
- K Nearest Neighbor

Candidate Models - Feature Selection

1

38 Features in Total

Too many input features at the same time.
Might have multicollinearity problem.

2

Multicollinearity Test

1. Correlation Matrix
2.Variance Inflation Factors Analysis

3

Refinement Required

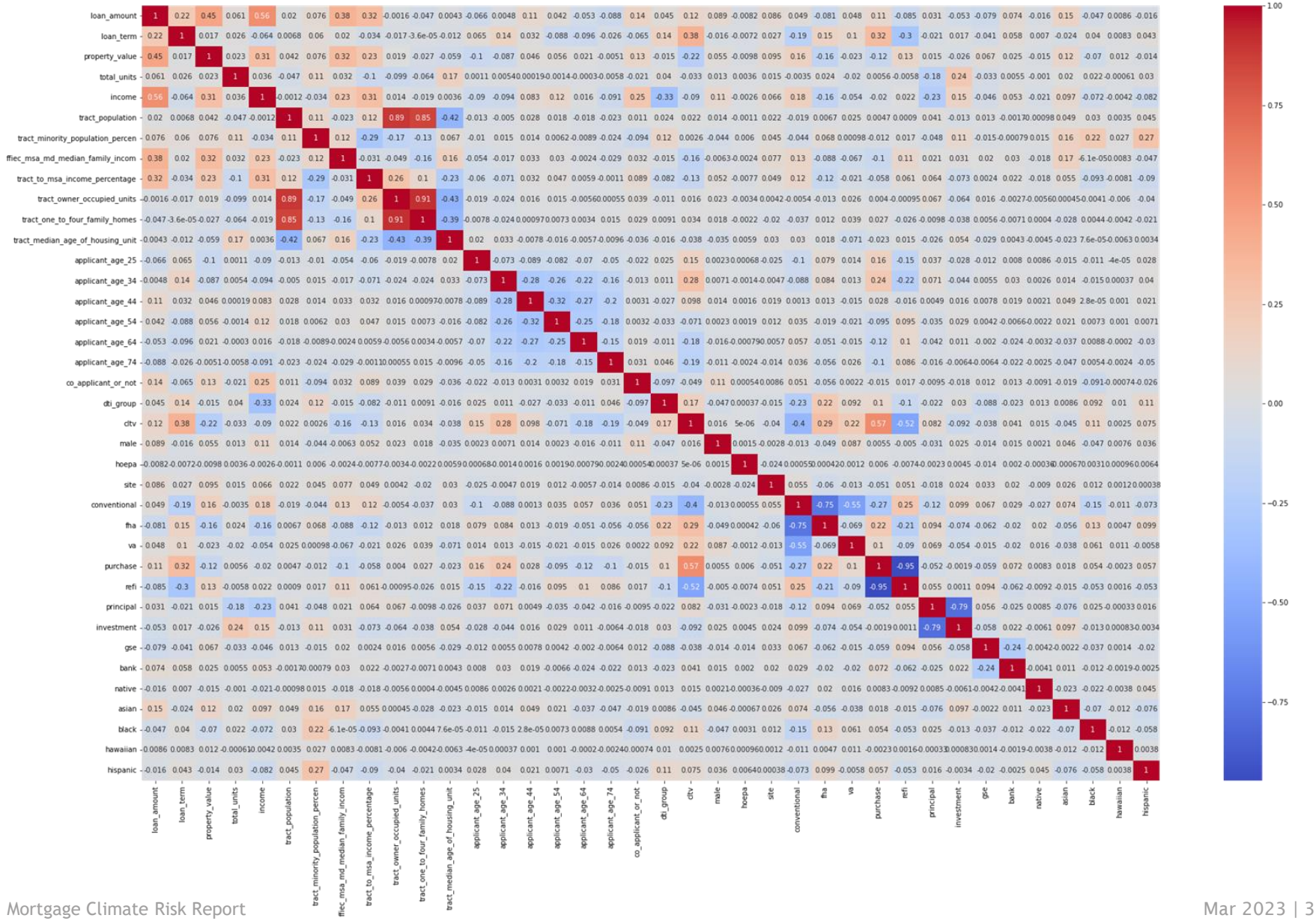
Different Models
generate
Different Important Features.
Select the Best Model in
Different Model Categories.

4

Champion Model

Compare the Best Model
select
Champion Model.

Candidate Models - Feature Selection



Candidate Models - Feature Selection

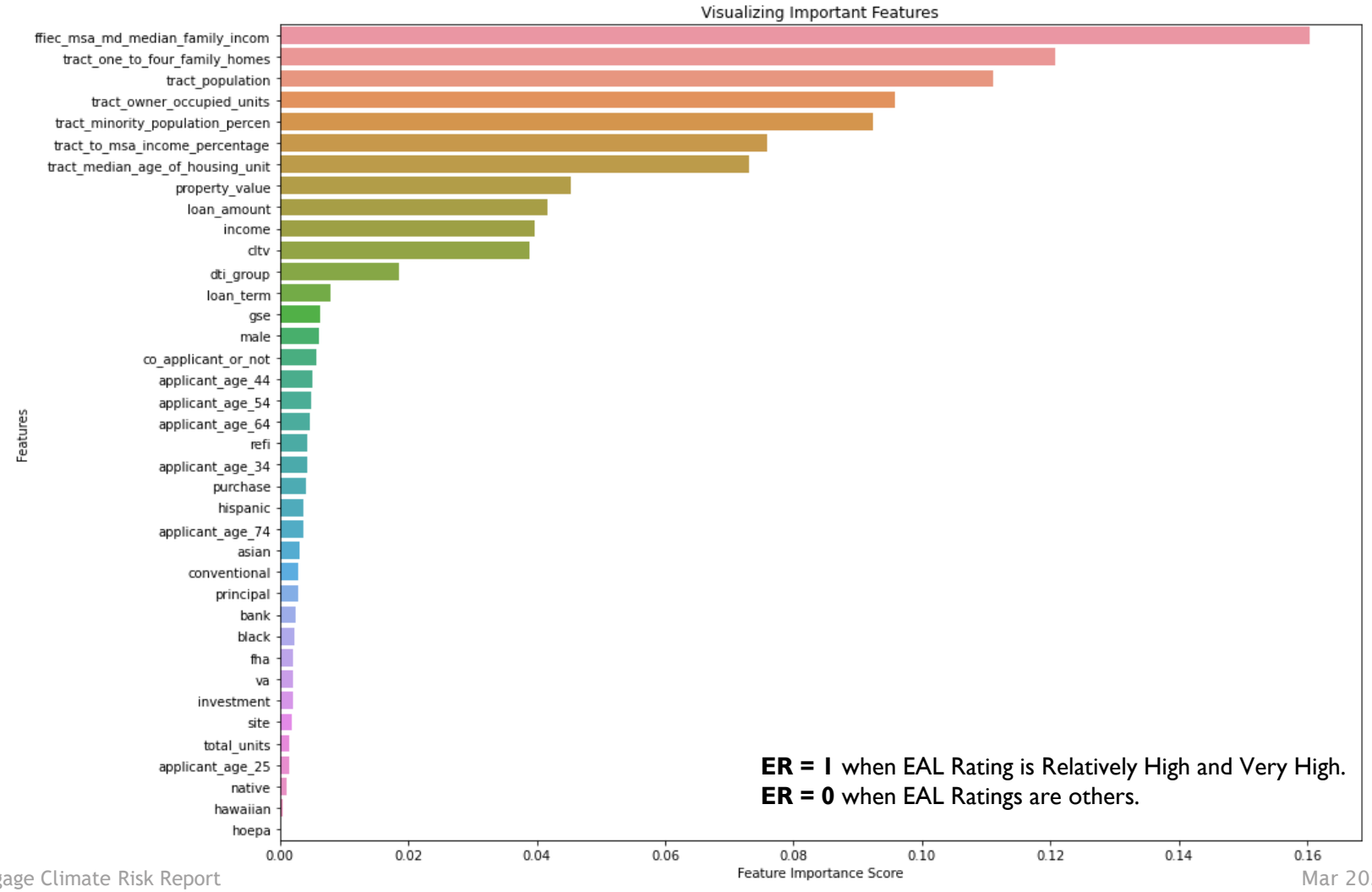
Multicollinearity Test (Variance Inflation Factor)

Variable	VIF
conventional	15.84
tract_owner_occupied_units	12.32
purchase	11.99
refi	11.14
fha	10.59
tract_one_to_four_family_homes	8.18
tract_population	7.39
va	7.13
applicant_age_44	7.03
applicant_age_54	6.48
applicant_age_34	6.05
applicant_age_64	5.45
applicant_age_74	3.75
principal	3.10
investment	2.89
loan_amount	2.69
income	2.30
cltv	2.24
tract_minority_population_percen	1.89
applicant_age_25	1.80
tract_to_msa_income_percentage	1.78
ffiec_msa_md_median_family_incom	1.57
property_value	1.54
tract_median_age_of_housing_unit	1.36
dti_group	1.35
loan_term	1.33
hispanic	1.15
black	1.14
gse	1.14
asian	1.13
total_units	1.12
co_applicant_or_not	1.11
bank	1.09
male	1.05
site	1.03
native	1.00
hawaiian	1.00
hoepa	1.00

Candidate Models – Feature Selection

Feature Importance Plot

Features: All
Target Variable: ER



ER = 1 when EAL Rating is Relatively High and Very High.
ER = 0 when EAL Ratings are others.

Model Results

- **Mean Squared Error (MSE):** Measures the amount of error in statistic models.

- **K-S Ratio:** Measures the degree of separation between the cumulative distribution functions of the predicted probabilities for the positive and negative outcomes in the model.

- **Accuracy Score:** Measures the proportion of correct predictions made by the mode out of all the predictions made.

- **Recall Score:** Measures the proportion of true positives (correctly predicted positive instances) out of all positive predictions made by the model.

- **Precision Score:** Measures the proportion of true positives out of all actual positive instances.

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Random Forest	All	RR		4.9394%	95.0606%	99.4156%	68.5878%	90.5123%
Random Forest	All	ER		5.9462%	94.0538%	99.43%	66.6216%	88.7553%
Random Forest	All	RS (Bisection)		6.1814%	93.8186%	95.2068%	92.2831%	87.7520%
Random Forest	All	ES (Bisection)		7.4120%	92.5880%	94.8618%	90.0532%	85.5384%
Random Forest	All	RR_Reverse		6.6232%	93.3768%	91.8691%	97.3908%	87.1418%
Random Forest	All	ER_Reverse		7.4094%	92.5906%	90.6129%	96.1139%	85.4771%
Random Forest	Top 11	RR		0.9428%	99.0572%	99.8278%	94.0896%	98.0281%
Random Forest	Top 11	ER		1.4758%	98.5242%	99.8813%	91.7298%	97.2109%
Random Forest	Top 11	RS (Bisection)		1.2866%	98.7134%	98.8923%	98.5304%	97.4984%
Random Forest	Top 11	ES (Bisection)		1.8290%	98.1710%	98.4151%	97.9187%	96.4568%
Random Forest	Top 11	RR_Reverse		1.2554%	98.7446%	98.4409%	99.4466%	97.5867%
Random Forest	Top 11	ER_Reverse		1.7060%	98.2940%	97.6914%	99.1559%	96.7629%
Random Forest	Tract	RR		0.1562%	99.8438%	99.8961%	99.0970%	99.3088%
Random Forest	Tract	ER		0.1188%	99.8812%	99.9737%	99.3516%	99.4622%
Random Forest	Tract	RS (Bisection)		0.4838%	99.5162%	99.4251%	99.6084%	99.0368%
Random Forest	Tract	ES (Bisection)		0.6650%	99.3350%	99.1057%	99.5684%	98.6800%
Random Forest	Tract	RR_Reverse		0.2834%	99.7166%	99.6626%	99.8576%	99.4041%
Random Forest	Tract	ER_Reverse		0.2884%	99.7116%	99.5898%	99.8726%	99.4179%
Logistics	All	RR		15.3176%	84.6824%	54.6765%	7.8170%	35.2612%
Logistics	All	ER		16.6790%	83.3210%	64.4270%	11.8358%	35.0353%
Logistics	All	RS (Bisection)		31.9992%	68.0008%	68.9595%	65.4712%	36.1645%
Logistics	All	ES (Bisection)		33.1424%	66.8576%	67.9872%	63.7131%	33.8332%
Logistics	All	RR_Reverse		31.5922%	68.4078%	69.2541%	83.5175%	35.4336%
Logistics	All	ER_Reverse		32.9548%	67.0452%	66.6230%	77.0115%	33.9234%
Logistics	Top10	RR		15.3768%	84.6232%	53.6075%	7.1020%	34.4727%
Logistics	Top10	ER		16.7606%	83.2394%	63.2760%	11.5281%	34.5828%
Logistics	Top10	RS (Bisection)		32.7366%	67.2634%	68.3212%	64.3752%	34.7148%
Logistics	Top10	ES (Bisection)		33.6692%	66.3308%	67.6660%	62.5471%	32.7712%
Logistics	Top10	RR_Reverse		32.6276%	67.3724%	67.9127%	84.7053%	33.0255%
Logistics	Top10	ER_Reverse		33.4072%	66.5928%	66.0037%	77.5051%	32.8685%
Logistics	Top11+p	ER		16.7838%	83.2162%	64.5813%	10.4209%	32.2494%
Logistics	Top10+p	ER		16.8260%	83.1740%	63.0743%	10.7730%	33.2461%
Neural Network	All	RR		9.40%	87.42%	73.81%	29.42%	52.38%
Neural Network	All	ER		9.78%	87.01%	78.56%	36.10%	52.94%
Neural Network	All	RS (Bisection)		15.64%	76.67%	76.94%	76.16%	53.39%
Neural Network	All	ES (Bisection)		16.22%	75.33%	76.01%	74.03%	50.79%
Neural Network	All	RR_Reverse		14.97%	78.00%	78.74%	85.89%	54.24%
Neural Network	All	ER_Reverse		15.90%	76.15%	75.44%	82.18%	51.79%
Neural Network	Top10	RR		10.24%	86.75%	72.75%	23.42%	44.09%
Neural Network	Top11	ER		7.66%	89.91%	80.57%	56.27%	64.26%
Neural Network	Top14	RS (Bisection)		12.46%	81.74%	81.66%	81.86%	63.54%
Neural Network	Top12	ES (Bisection)		12.12%	82.31%	81.68%	83.30%	64.88%
Neural Network	Top15	RR_Reverse		11.50%	83.49%	83.92%	89.08%	65.52%
Neural Network	Top12	ER_Reverse		12.07%	82.44%	83.16%	84.25%	64.62%

source: Model Results.xlsx

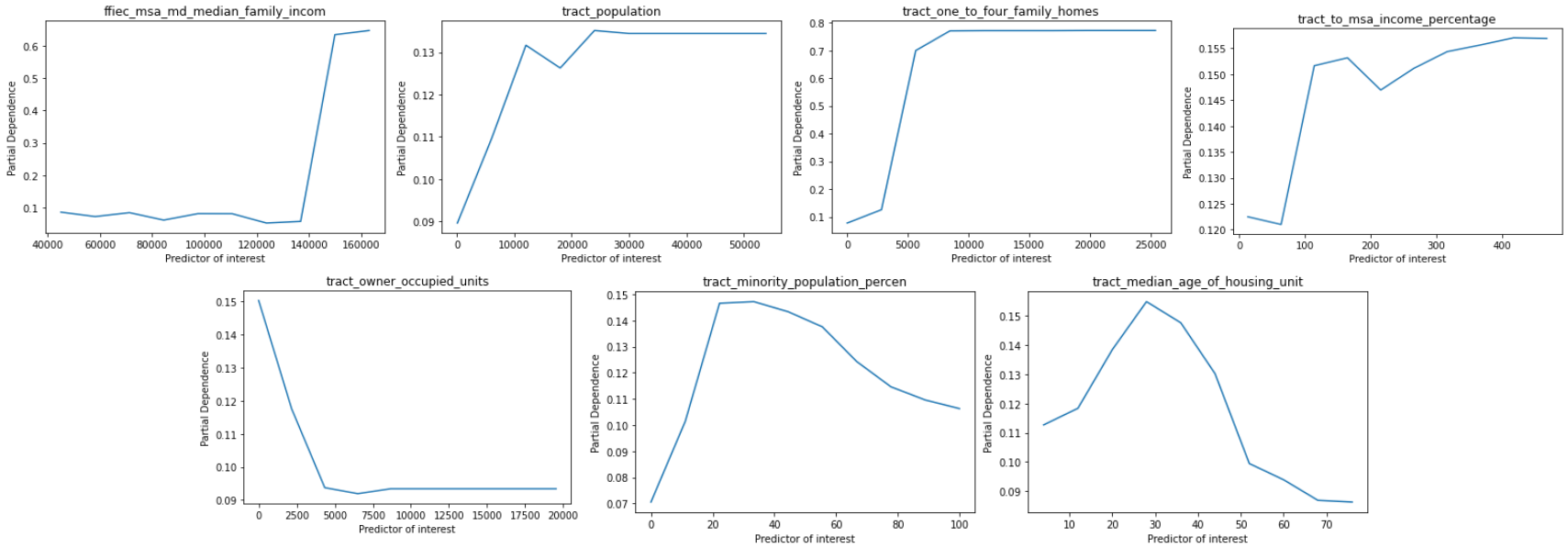
Champion Model

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Random Forest	All	RR		4.9394%	95.0606%	99.4156%	68.5878%	90.5123%
Random Forest	All	ER		5.9462%	94.0538%	99.4289%	66.6216%	88.7553%
Random Forest	All	RS (Bisection)		6.1814%	93.8186%	95.2068%	92.2831%	87.7520%
Random Forest	All	ES (Bisection)		7.4120%	92.5880%	94.8618%	90.0532%	85.5384%
Random Forest	All	RR_Reverse		6.6232%	93.3768%	91.8691%	97.3908%	87.1418%
Random Forest	All	ER_Reverse		7.4094%	92.5906%	90.6129%	96.1139%	85.4771%
Random Forest	Top 11	RR		0.9428%	99.0572%	99.8278%	94.0896%	98.0281%
Random Forest	Top 11	ER		1.4758%	98.5242%	99.8813%	91.7298%	97.2109%
Random Forest	Top 11	RS (Bisection)		1.2866%	98.7134%	98.8923%	98.5304%	97.4984%
Random Forest	Top 11	ES (Bisection)		1.8290%	98.1710%	98.4151%	97.9187%	96.4568%
Random Forest	Top 11	RR_Reverse		1.2554%	98.7446%	98.4409%	99.4466%	97.5867%
Random Forest	Top 11	ER_Reverse		1.7060%	98.2940%	97.6914%	99.1559%	96.7629%
Random Forest	Tract	RR		0.1562%	99.8438%	99.8961%	99.0970%	99.3088%
Random Forest	Tract	ER		0.1188%	99.8812%	99.9737%	99.3516%	99.4622%
Random Forest	Tract	RS (Bisection)		0.4838%	99.5162%	99.4251%	99.6084%	99.0368%
Random Forest	Tract	ES (Bisection)		0.6650%	99.3350%	99.1057%	99.5684%	98.6800%
Random Forest	Tract	RR_Reverse		0.2834%	99.7166%	99.6626%	99.8576%	99.4041%
Random Forest	Tract	ER_Reverse		0.2884%	99.7116%	99.5898%	99.8726%	99.4179%

Logistics	All	RR		15.3176%	84.6824%	54.6765%	7.8170%	35.2612%
Logistics	All	ER		16.6790%	83.3210%	64.4270%	11.8358%	35.0353%
Logistics	All	RS (Bisection)		31.9992%	68.0008%	68.9595%	65.4712%	36.1645%
Logistics	All	ES (Bisection)		33.1424%	66.8576%	67.9872%	63.7131%	33.8332%
Logistics	All	RR_Reverse		31.5922%	68.4078%	69.2541%	83.5175%	35.4336%
Logistics	All	ER_Reverse		32.9548%	67.0452%	66.6230%	77.0115%	33.9234%
Logistics	Top10	RR		15.3768%	84.6232%	53.6075%	7.1020%	34.4727%
Logistics	Top10	ER		16.7606%	83.2394%	63.2760%	11.5281%	34.5828%
Logistics	Top10	RS (Bisection)		32.7366%	67.2634%	68.3212%	64.3752%	34.7148%
Logistics	Top10	ES (Bisection)		33.6692%	66.3308%	67.6660%	62.5471%	32.7712%
Logistics	Top10	RR_Reverse		32.6276%	67.3724%	67.9127%	84.7053%	33.0255%
Logistics	Top10	ER_Reverse		33.4072%	66.5928%	66.0037%	77.5051%	32.8685%

source: Model Results.xlsx

Statistical Inference - Partial Dependence Plot



Positive: ffiec msa md median family income, tract population, tract one to four family homes, tract to msa income percentage

Negative: tract owner occupied units

Different Trend in Different Bucket: tract minority population percent, tract median age of housing unit

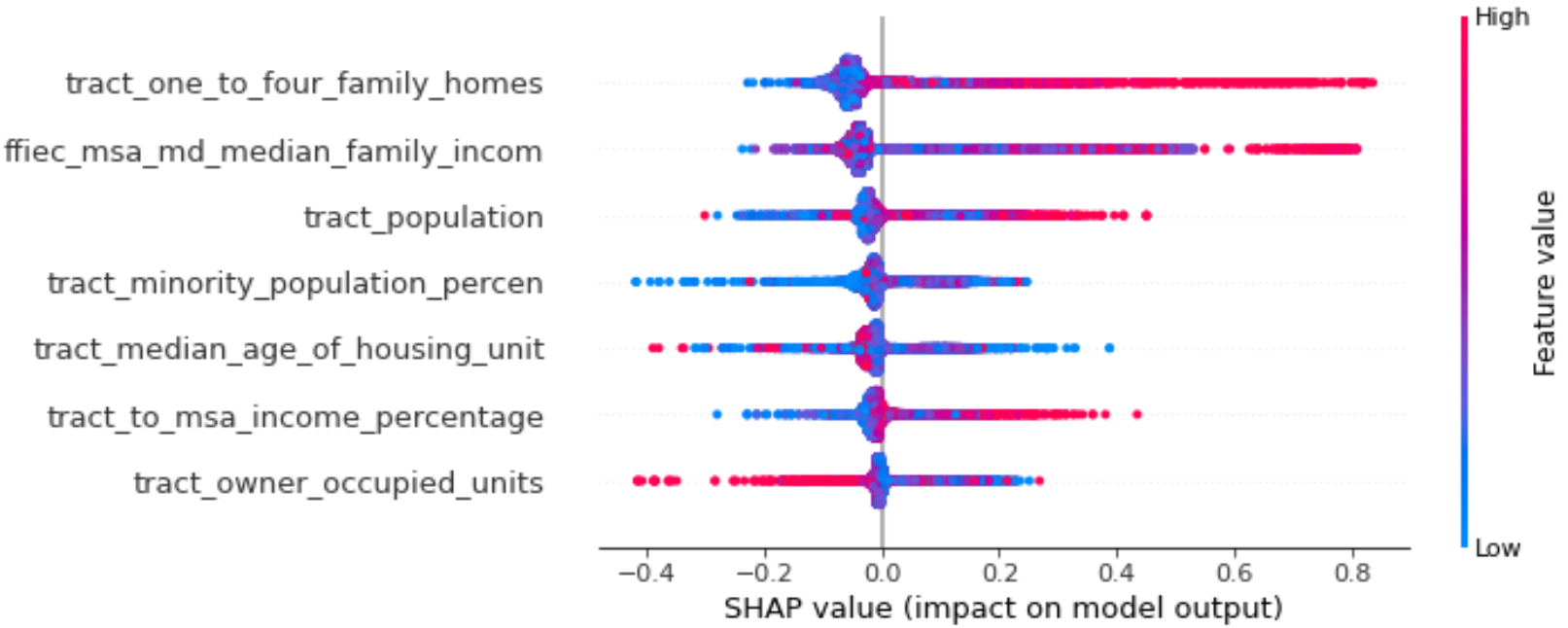
Statistical Inference - Beeswarm Plot

The location of the cluster decides the relationship between features and Target Variable.

If the cluster falls onto the **left** to the 0, then the feature has the **Negative** relationship with the target variable.

If the cluster falls onto the **Right** to the 0, then the feature has the **Positive** relationship with the target variable.

Feature Importance Rank: The **more scattered** the points, the **more important** the variable is.



Statistical Inference - Change in Odds

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Logistics	All	RR		15.3176%	84.6824%	54.6765%	7.8170%	35.2612%
Logistics	All	ER		16.6790%	83.3210%	64.4270%	11.8358%	35.0353%
Logistics	All	RS (Bisection)		31.9992%	68.0008%	68.9595%	65.4712%	36.1645%
Logistics	All	ES (Bisection)		33.1424%	66.8576%	67.9872%	63.7131%	33.8332%
Logistics	All	RR_Reverse		31.5922%	68.4078%	69.2541%	83.5175%	35.4336%
Logistics	All	ER_Reverse		32.9548%	67.0452%	66.6230%	77.0115%	33.9234%
Logistics	Top10	RR		15.3768%	84.6232%	53.6075%	7.1020%	34.4727%
Logistics	Top10	ER		16.7606%	83.2394%	63.2760%	11.5281%	34.5828%
Logistics	Top10	RS (Bisection)		32.7366%	67.2634%	68.3212%	64.3752%	34.7148%
Logistics	Top10	ES (Bisection)		33.6692%	66.3308%	67.6660%	62.5471%	32.7712%
Logistics	Top10	RR_Reverse		32.6276%	67.3724%	67.9127%	84.7053%	33.0255%
Logistics	Top10	ER_Reverse		33.4072%	66.5928%	66.0037%	77.5051%	32.8685%

Variable	Coefficient	P-value	Change in Odds
tract_one_to_four_family_homes	0.0007	<0.001	0.0700%
tract_population	0.0001	<0.001	0.0100%
ffiec_msa_md_median_family_incom	0.00000523	<0.001	0.0005%
loan_amount	0.000002103	<0.001	0.0002%
tract_owner_occupied_units	-0.0007	<0.001	-0.0700%
tract_median_age_of_housing_unit	-0.0165	<0.001	-1.6365%
cltv	-0.0172	<0.001	-1.7053%
fha	-1.9408	<0.001	-85.6411%
va	-2.067	<0.001	-87.3435%
conventional	-2.1144	<0.001	-87.9294%

The coefficients of the logistic regression in terms of the change in odds can be interpreted as if the census tract has 1 more unit of 1-4 family home, then the loan is exposed to **0.07% more odds** of being in a high loss area.

A second example would be if loan is issued and insured by VA then the odds of a loan in this tract will have a **decreased odds of 87.34%**.

Further Refinement

Target Variable: Overall EAL Rating	
Variable	**Count**
loan_amount	8
loan_term	0
property_value	5
total_units	3
income	2
tract_population	11
tract_minority_population_percen	7
ffiec_msa_md_median_family_income	10
tract_to_msa_income_percentage	7
tract_owner_occupied_units	8
tract_one_to_four_family_homes	11
tract_median_age_of_housing_unit	10
applicant_age_25	0
applicant_age_34	0
applicant_age_44	0
applicant_age_54	0
applicant_age_64	0
applicant_age_74	0
co_applicant_or_not	0
dti_group	0
cltv	7
male	0
hoepa	0
site	3
conventional	3
fha	3
va	3
purchase	0
refi	0
principal	0
investment	0
gse	0
bank	0
native	0
asian	3
black	3
hawaiian	0
hispanic	0

Target Variable: Overall Risk Rating	
Variable	**Count**
loan_amount	8
loan_term	3
property_value	5
total_units	2
income	1
tract_population	11
tract_minority_population_percen	9
ffiec_msa_md_median_family_income	11
tract_to_msa_income_percentage	9
tract_owner_occupied_units	7
tract_one_to_four_family_homes	11
tract_median_age_of_housing_unit	9
applicant_age_25	0
applicant_age_34	0
applicant_age_44	0
applicant_age_54	0
applicant_age_64	0
applicant_age_74	0
co_applicant_or_not	0
dti_group	0
cltv	7
male	0
hoepa	0
site	3
conventional	3
fha	3
va	3
purchase	0
refi	0
principal	3
investment	0
gse	0
bank	0
native	0
asian	3
black	3
hawaiian	0
hispanic	0

source: Model Results.xlsx

Further Refinement

Variable	VIF
conventional	15.84
tract_owner_occupied_units	12.32
purchase	11.99
refi	11.14
fha	10.59
tract_one_to_four_family_homes	8.18
tract_population	7.39
va	7.13
applicant_age_44	7.03
applicant_age_54	6.48
applicant_age_34	6.05
applicant_age_64	5.45
applicant_age_74	3.75
principal	3.10
investment	2.89
loan_amount	2.69
income	2.30
cltv	2.24
tract_minority_population_percen	1.89
applicant_age_25	1.80
tract_to_msa_income_percentage	1.78
ffiec_msa_md_median_family_incom	1.57
property_value	1.54
tract_median_age_of_housing_unit	1.36
dti_group	1.35
loan_term	1.33
hispanic	1.15
black	1.14
gse	1.14
asian	1.13
total_units	1.12
co_applicant_or_not	1.11
bank	1.09
male	1.05
site	1.03
native	1.00
hawaiian	1.00
hoepa	1.00

variable	VIF
tract_owner_occupied_units	11.60
tract_one_to_four_family_homes	7.52
tract_population	7.20
tract_minority_population_percen	1.60
tract_to_msa_income_percentage	1.34
tract_median_age_of_housing_unit	1.31
ffiec_msa_md_median_family_incom	1.15

Further Refinement - 3-Factor Model

Sharpley value of tract_own_occupied_units ≈ 0

#	MSE_Test	Accuracy_Score	Recall_Score	Precision_Score	KS
1	0.1226%	99.8774%	99.9064%	99.3970%	99.3771%
6	0.1274%	99.8726%	99.9623%	99.3141%	99.3061%
30	0.1644%	99.8356%	99.6483%	99.4175%	99.3425%
36	6.0450%	93.9550%	92.0005%	71.9326%	70.5955%
Best	# 1	# 1	# 6	# 30	# 1
Tract	0.1188%	99.8812%	99.9737%	99.3516%	99.4622%

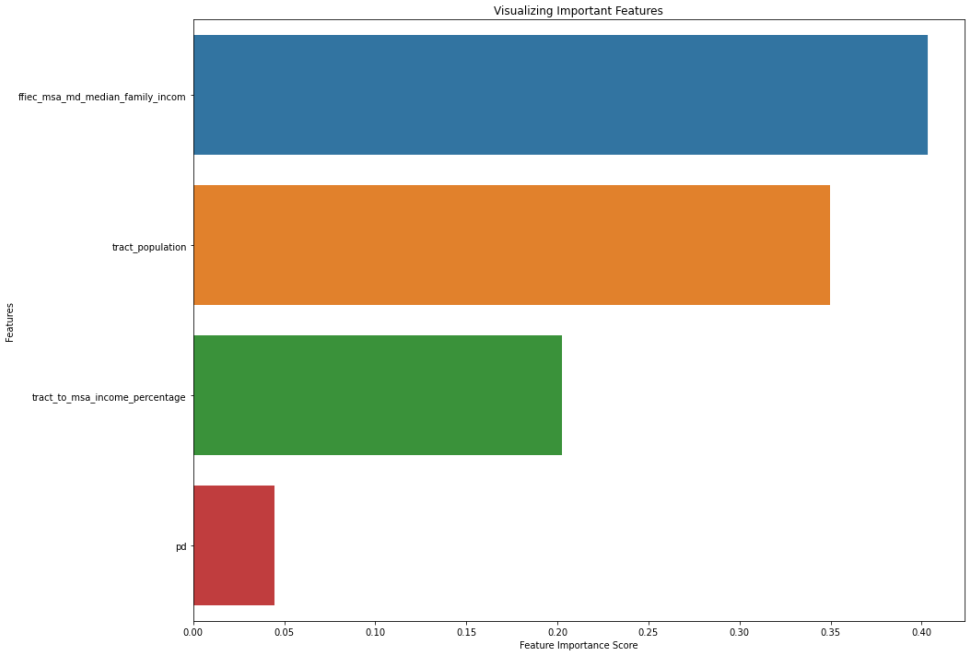
Rank	Variable 1	Variable 2	Variable 3
# 1	tract population	tract to msa income percentage	ffiec msa md median family income
# 6	tract one to four family homes	tract owner occupied units	tract minority population percent
# 30	ffiec msa md median family income	tract median age of housing unit	tract minority population percent
# 36	cltv	tract one to four family homes	ffiec msa md median family income

Further Refinement - Probability of Default

If individual borrower's attribute does not have a significant explanatory power to the loss rating, an aggregate variable might boost the predictability.

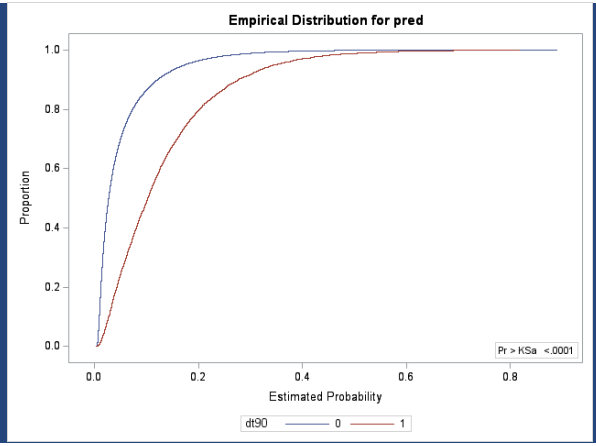
An aggregate variable would be the probability of default calculated by the borrower's attribute. By using a GSE dataset from 2017, we have obtained the logit coefficients of some of the main borrower's attributes, and the default model has achieved a K-S statistics of 45.74%. With the obtained coefficients, we applied the logit model in our HMDA data, and produced a "probability of default" variable for each loan.

The variable is then used in the Random Forest model with the top 3 tract-only variables. However, the borrower's importance remains a low importance in predictability. Which led to a worse result than the 3-factor model, and the model is discarded.



Exhibit

Default Probability Prediction



Table

Default Coefficient

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	7.9158	0.1834	1863.3767	<.0001
cscore_b	1	-0.0157	0.000257	3741.1952	<.0001
dti_group	1	0.1915	0.00768	621.5610	<.0001
loan_purpose_p	1	-0.1762	0.0328	28.8664	<.0001
occ_stat_i	1	0.2786	0.0571	23.7794	<.0001
cnt_borrower_2	1	-0.4351	0.0307	201.2007	<.0001

4.2

Hazard Modeling

Testing if any variable has any significance to specific hazards

- Hazard Selection
- Data Cleaning
- Modeling Results

Hazard Selection

Based on the Natural Language Processing and topic modelling on over 5,000 random tweets, the more relevant topics are shown in the graphs to the left. People express their concern and direct their conversation towards **Tornado** and **Fire**.

Hazard Selection

Variable	N	Median	Sum
AVLN_EALT	1867855	40393.56	587422418287
CFLD_EALT	5925195	0.7968158	325466369368
CWAV_EALT	13257046	0	77298605238
DRGT_EALT	13257046	7.3198588	576127341773
ERQK_EALT	13257046	9716.64	1.8272436E12
HAIL_EALT	13257046	1120.83	281633948055
HWAV_EALT	13257046	2139.43	159179870264
HRCN_EALT	9562853	1860.72	1.1313771E12
ISTM_EALT	11258254	1966.07	83344363316
LNDS_EALT	13177171	133.2561699	50116246573
LTNG_EALT	13177171	4906.14	121001288844
RFLD_EALT	13257046	13281.34	1.0583587E12
SWND_EALT	13257046	6709.74	243103035670
TRND_EALT	13257046	37347.51	1.1755214E12
TSUN_EALT	1684767	0	675523940
VLCN_EALT	700945	25520.82	38838891992
WFIR_EALT	13177171	1.1173335	582318334315
WNTW_EALT	13257046	1307.64	58332347214

Variable	N	Median	Sum
AVLN EALT	1867855	40393.56	587422418287
CFLD EALT	5925195	0.7968158	325466369368
DRGT EALT	13257046	7.3198588	576127341773
ERQK EALT	13257046	9716.64	1.8272436E12
HRCN EALT	9562853	1860.72	1.1313771E12
TRND EALT	13257046	37347.51	1.1755214E12
WFIR EALT	13177171	1.1173335	582318334315

Sum of Loss

The 6 hazards with the highest sum of expected loss are **Earthquake, Hurricane, River Flooding, Tornado, Avalanche** and **Drought.**

Median of Loss

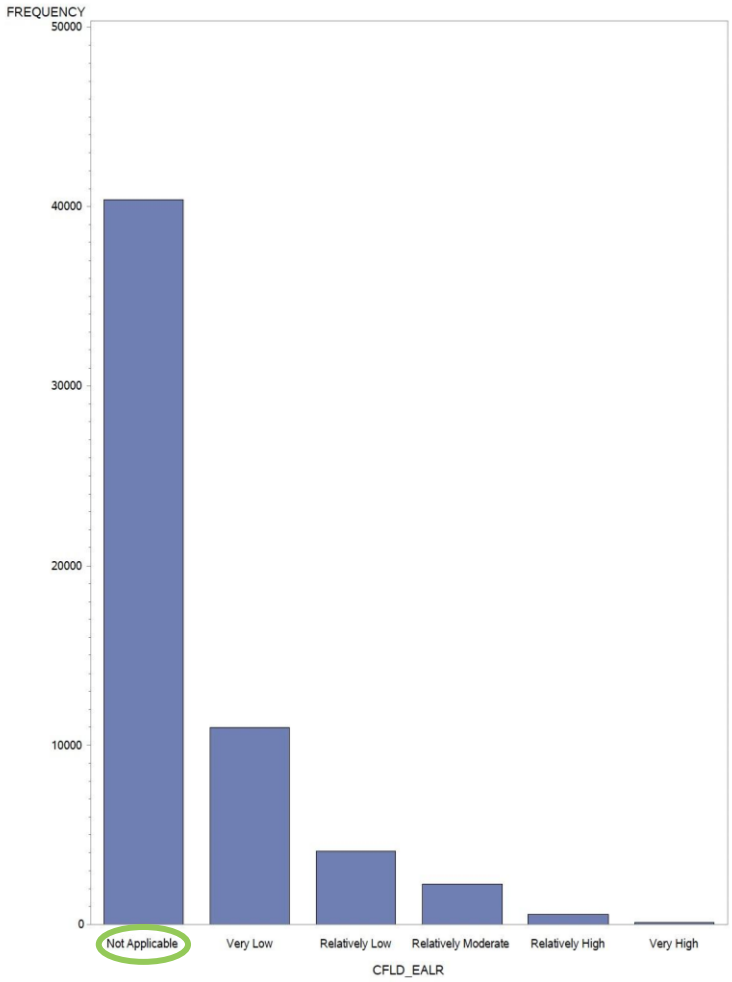
The 6 hazards with the highest median of expected loss are **Avalanche, Tornado, Volcano, River Flooding, Earthquake** and **Strong Wind.**

Selection

The number of observations for **Avalanche** and **Volcano** are relatively low, the two hazards are dropped in our selection

Hazard Modelling - Data Cleaning

To avoid skewed dataset, missing data were dropped.



Very High	Very Low
Relatively High	No Expected Annual Loss
Relatively Moderate	Not Applicable
Relatively Low	Insufficient Data

Very High
Relatively High
Relatively Moderate
Relatively Low
Very Low

Hazard Modelling - Results

Hazard	Model	X	Y	MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
HRCN (Hurricane)	Random Forest	All	ER	2.45%	97.55%	99.73%	73.54%	95.11%
	Random Forest	Top 11	ER	0.30%	99.70%	99.85%	96.88%	99.26%
	Random Forest	Tract	ER	0.07%	99.93%	99.93%	99.33%	99.61%
CFLD (Coastal Flooding)	Random Forest	All	ER	2.21%	97.79%	100.00%	57.34%	94.97%
	Random Forest	Top 11	ER	0.22%	99.78%	99.93%	95.84%	99.23%
	Random Forest	Tract	ER	0.03%	99.97%	100.00%	99.37%	99.68%
RFLD (Riverine Flooding)	Random Forest	All	ER	6.43%	93.57%	97.68%	51.94%	85.56%
	Random Forest	Top 11	ER	1.07%	98.93%	99.88%	93.40%	98.12%
	Random Forest	Tract	ER	0.22%	99.78%	99.90%	98.71%	99.00%
DRGT (Drought)	Random Forest	All	ER	1.84%	98.16%	99.84%	40.17%	40.17%
	Random Forest	Top 11	ER	0.40%	99.60%	99.86%	87.08%	87.07%
	Random Forest	Tract	ER	0.06%	99.94%	99.91%	98.09%	98.09%
WFIR (Wildfire)	Random Forest	All	ER	3.22%	96.78%	96.65%	60.24%	85.86%
	Random Forest	Top 11	ER	1.07%	98.93%	96.60%	89.21%	94.50%
	Random Forest	Tract	ER	0.22%	99.78%	97.47%	99.68%	99.47%
ERQK (Earthquake)	Random Forest	All	ER	5.64%	94.36%	95.19%	72.16%	86.66%
	Random Forest	Top 11	ER	1.87%	98.13%	98.56%	90.90%	95.79%
	Random Forest	Tract	ER	0.06%	99.94%	99.98%	99.72%	99.77%
TRND (Tornado)	Random Forest	All	ER	7.76%	92.24%	96.30%	75.13%	82.84%
	Random Forest	Top 11	ER	2.00%	98.00%	99.56%	93.25%	96.45%
	Random Forest	Tract	ER	0.21%	99.79%	99.89%	99.37%	99.47%

Thank You For Listening

Mar 2023

Robert H. Smith School of Business

5

Appendix

Model Refining

Confusion Matrix:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

Naïve Bayes

Definition:

Naïve Bayes is a simple learning algorithm that utilizes Bayes' rule together with a strong assumption that the attributes are conditionally independent given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy. Coupled with its computational efficiency and many other desirable features, this leads to naïve Bayes being widely applied in practice.

Assumption:

The **naive Bayes** or **idiot Bayes** assumption is that all the features are conditionally independent given the class label:

$$p(x|y=c) = \prod_{i=1}^{D} p(x_i|y=c)$$

Even though this is usually false (since features are usually dependent), the resulting model is easy to fit and works surprisingly well. In the case of Gaussian data, we get

$$p(x|y=c,\theta_c) = \prod_{i=1}^{D} \mathcal{N}(x_i|\mu_{ic},\sigma_{ic})$$

so we just have to estimate $C \times D$ separate Gaussian parameters, μ_{ic}, σ_{ic}. In the case of binary data, we get

$$p(x|y=c,\theta_c) = \prod_{i=1}^{D} Be(x_i|\theta_{ic})$$

so we just have to estimate $C \times D$ separate Bernoulli parameters, θ_{ic}. We discuss the case of binary (and K-ary) data in more detail below.

Naïve Bayes

Principle of classification:

For categorical attributes, the required probabilities $P(y)$ and $P(x_i \mid y)$ are normally derived from frequency counts stored in arrays whose values are calculated by a single pass through the training data at training time. These arrays can be updated as new data are acquired, supporting incremental learning. Probability estimates are usually derived from the frequency counts using smoothing functions such as the Laplace estimate or an m-estimate.

For numeric attributes, either the data are discretized (see discretization), or probability density estimation is employed.

Naïve Bayes

Principle of classification:

Naïve Bayes is a form of Bayesian Network Classifier based on Bayes' rule

$$P(y \mid \mathbf{x}) = P(y)P(\mathbf{x} \mid y)/P(\mathbf{x}) \quad (1)$$

together with an assumption that the attributes are conditionally independent given the class. For attribute-value data, this assumption entitles

$$P(\mathbf{x} \mid y) = \prod_{i=1}^{n} P(x_i \mid y) \quad (2)$$

where x_i is the value of the i^{th} attribute in $\mathbf{x}$, and n is the number of attributes.

$$P(\mathbf{x}) = \prod_{i=1}^{k} P(c_i)P(\mathbf{x} \mid c_i) \quad (3)$$

where k is the number of classes and c_i is the i^{th} class. Thus, (1) can be calculated by normalizing the numerators of the right-hand-side of the equation.

The resulting classifier uses a linear model, equivalent to that used by logistic regression, differing only in the manner in which the parameters are chosen.

Model Results of Naïve Bayes

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Naive_bayes	All	RR		17.3550%	82.6450%	39.1479%	21.2583%	26.7057%
Naive_bayes	All	ER		19.1838%	80.8162%	43.6829%	30.8438%	29.9300%
Naive_bayes	All	RS (Bisection)		38.7788%	61.2212%	71.7324%	37.0364%	28.4601%
Naive_bayes	All	ES (Bisection)		37.2676%	62.7324%	71.4117%	42.4603%	28.2790%
Naive_bayes	All	RR_Reverse		34.8972%	65.1028%	65.0682%	88.1700%	27.2199%
Naive_bayes	All	ER_Reverse		36.0036%	63.9964%	62.0843%	84.0872%	28.2737%
Naive_bayes	Top5	RR		16.9208%	83.0792%	40.6371%	19.5076%	23.9719%
Naive_bayes	Top6	ER		19.2800%	80.7200%	42.5115%	26.8717%	27.9930%
Naive_bayes	Top5	RS (Bisection)		39.2730%	60.7270%	70.1739%	37.3116%	24.8883%
Naive_bayes	Top6	ES (Bisection)		38.5862%	61.4138%	70.5270%	39.2110%	24.9023%
Naive_bayes	Top5	RR_Reverse		35.4812%	64.5188%	64.7787%	87.3270%	24.1563%
Naive_bayes	Top6	ER_Reverse		37.2244%	62.7756%	60.9458%	84.7807%	25.1704%
Naive_bayes	Tract	RR		17.2154%	82.7846%	38.7802%	18.8146%	26.0241%
Naive_bayes	Tract	ER		18.4524%	81.5476%	45.4298%	23.6989%	28.7987%
Naive_bayes	Tract	RS (Bisection)		40.0036%	59.9964%	71.1019%	33.6808%	25.8613%
Naive_bayes	Tract	ES (Bisection)		38.5464%	61.4536%	72.7904%	36.5773%	25.6284%
Naive_bayes	Tract	RR_Reverse		35.8092%	64.1908%	64.0272%	89.6690%	24.5161%
Naive_bayes	Tract	ER_Reverse		37.1070%	62.8930%	60.5944%	87.7151%	25.5141%

Model Results of Random Combination of Random Forest

When selecting 2 variables:

#	MSE_Test	Accuracy_Score	Recall_Score	Precision_Score	KS
1	0.1684%	99.8316%	99.7513%	99.2914%	99.2385%
2	0.1810%	99.8190%	99.6865%	99.2846%	99.2178%
3	0.1858%	99.8142%	99.6603%	99.2835%	99.2111%
4	0.1974%	99.8026%	99.6590%	99.2187%	99.1462%
5	0.1994%	99.8006%	99.6351%	99.2312%	99.1536%
6	0.3874%	99.6126%	98.9853%	98.8133%	98.5968%
7	0.5088%	99.4912%	98.4126%	98.7032%	98.3629%
...	...	...	...	...	...
49	21.3982%	78.6018%	37.0039%	30.5997%	19.4633%
50	22.2838%	77.7162%	22.1860%	10.5777%	2.6466%
51	22.3478%	77.6522%	22.4851%	10.9853%	2.8894%
52	23.3436%	76.6564%	21.6147%	12.3889%	2.7842%
53	23.4158%	76.5842%	29.7210%	24.1452%	11.9397%
54	23.5444%	76.4556%	20.4409%	11.6451%	1.9557%
55	23.5554%	76.4446%	19.0182%	10.3562%	0.9291%

#	Remain_variables
1	tract_minority_population_percen,tract_one_to_four_family_homes
2	tract_minority_population_percen,tract_population
3	tract_minority_population_percen,tract_owner_occupied_units
4	tract_population,tract_one_to_four_family_homes
5	tract_owner_occupied_units,tract_population
6	tract_owner_occupied_units,tract_one_to_four_family_homes
7	ffiec_msa_md_median_family_incom,tract_population
...	...
49	tract_population,income
50	property_value,cltv
51	loan_amount,cltv
52	tract_median_age_of_housing_unit,cltv
53	tract_minority_population_percen,income
54	cltv,tract_to_msa_income_percentage
55	income,cltv

Model Results of Random Combination of Random Forest

When selecting 3 variables:

#	MSE_Test	Accuracy_Score	Recall_Score	Precision_Score	KS
1	0.1226%	99.8774%	99.9064%	99.3970%	99.3771%
2	0.1254%	99.8746%	99.8893%	99.3982%	99.3746%
3	0.1258%	99.8742%	99.9360%	99.3493%	99.3357%
4	0.1260%	99.8740%	99.9087%	99.3754%	99.3560%
5	0.1270%	99.8730%	99.9531%	99.3255%	99.3155%
6	0.1274%	99.8726%	99.9623%	99.3141%	99.3061%
7	0.1274%	99.8726%	99.8927%	99.3834%	99.3606%
...	...	...	...	...	...
159	21.5616%	78.4384%	27.0743%	13.2394%	5.6159%
160	21.9188%	78.0812%	22.2489%	9.8009%	2.4789%
161	21.9528%	78.0472%	27.0238%	14.4919%	6.1258%
162	22.0418%	77.9582%	27.3187%	15.1448%	6.5311%
163	22.2172%	77.7828%	25.7974%	13.9332%	5.3656%
164	22.2736%	77.7264%	22.7152%	11.0160%	3.0036%
165	23.4618%	76.5382%	27.0821%	19.6223%	8.3279%

#	remain_variables
1	tract_population,tract_to_msa_income_percentage,ffiec_msa_md_median_family_incom
2	tract_minority_population_percen,tract_one_to_four_family_homes,ffiec_msa_md_median_family_incom
3	tract_population,tract_to_msa_income_percentage,tract_minority_population_percen
4	tract_minority_population_percen,tract_population,ffiec_msa_md_median_family_incom
5	tract_population,tract_owner_occupied_units,tract_minority_population_percen
6	tract_one_to_four_family_homes,tract_owner_occupied_units,tract_minority_population_percen
7	tract_minority_population_percen,tract_owner_occupied_units,ffiec_msa_md_median_family_incom
159	cltv,property_value,tract_median_age_of_housing_unit
160	income,tract_to_msa_income_percentage,loan_amount
161	income,property_value,tract_median_age_of_housing_unit
162	cltv,tract_median_age_of_housing_unit,loan_amount
163	income,tract_median_age_of_housing_unit,loan_amount
164	cltv,property_value,loan_amount
165	property_value,tract_minority_population_percen,loan_amount

Model Results of Random Combination of Random Forest

When selecting 8 variables:

#	MSE_Test	Accuracy_Score	Recall_Score	Precision_Score	KS
1	0.2356%	99.7644%	99.9781%	98.6839%	98.6793%
2	0.3282%	99.6718%	99.9746%	98.1615%	98.1562%
3	0.4796%	99.5204%	99.9673%	97.3087%	97.3019%
4	0.4814%	99.5186%	99.9673%	97.2985%	97.2917%
5	0.8694%	99.1306%	99.9749%	95.0876%	95.0825%
6	0.8804%	99.1196%	99.9534%	95.0456%	95.0361%
7	0.8864%	99.1136%	99.9391%	95.0252%	95.0128%
...	...	...	...	...	...
159	7.9098%	92.0902%	96.7182%	57.0251%	56.6114%
160	8.0040%	91.9960%	96.3154%	56.7253%	56.2614%
161	8.1494%	91.8506%	95.9218%	56.1155%	55.6055%
162	8.1668%	91.8332%	96.0960%	55.9020%	55.4165%
163	8.3068%	91.6932%	96.6288%	54.7460%	54.3377%
164	8.4456%	91.5544%	95.9131%	54.3645%	53.8693%
165	8.5430%	91.4570%	95.8301%	53.8376%	53.3368%

#	Drop_Variables
1	property_value,income,loan_amount
2	property_value,cltv,loan_amount
3	cltv,income,loan_amount
4	property_value,cltv,income
5	property_value,loan_amount,tract_one_to_four_family_homes
6	property_value,loan_amount,tract_owner_occupied_units
7	property_value,loan_amount,tract_to_msa_income_percentage
...	...
159	tract_median_age_of_housing_unit,tract_one_to_four_family_homes,ffiec_msa_md_median_family_incom
160	tract_population,tract_owner_occupied_units,ffiec_msa_md_median_family_incom
161	tract_minority_population_percen,tract_owner_occupied_units,ffiec_msa_md_median_family_incom
162	tract_one_to_four_family_homes,tract_owner_occupied_units,ffiec_msa_md_median_family_incom
163	tract_population,tract_one_to_four_family_homes,ffiec_msa_md_median_family_incom
164	tract_population,tract_minority_population_percen,ffiec_msa_md_median_family_incom
165	tract_minority_population_percen,tract_one_to_four_family_homes,ffiec_msa_md_median_family_incom

Model Results of Random Combination of Random Forest

When selecting 9 variables:

#	MSE_Test	Accuracy_Score	Recall_Score	Precision_Score	KS
1	0.3302%	99.6698%	99.9746%	98.1502%	98.1448%
2	0.5108%	99.4892%	99.9428%	97.1554%	97.1436%
3	0.5574%	99.4426%	99.9555%	96.8784%	96.8691%
4	0.5638%	99.4362%	99.9590%	96.8386%	96.8301%
5	0.6190%	99.3810%	99.9447%	96.5388%	96.5274%
6	0.6808%	99.3192%	99.9410%	96.1914%	96.1792%
7	0.9744%	99.0256%	99.9148%	94.5482%	94.5310%
...	...	...	...	...	...
49	3.0140%	96.9860%	99.1079%	83.6401%	83.4792%
50	3.8338%	96.1662%	98.5484%	79.4022%	79.1522%
51	3.9094%	96.0906%	98.4404%	79.0559%	78.7881%
52	3.9930%	96.0070%	98.5873%	78.4529%	78.2126%
53	4.2756%	95.7244%	98.5441%	76.8597%	76.6170%
54	4.3874%	95.6126%	98.4269%	76.3090%	76.0483%
55	4.4644%	95.5356%	98.1076%	76.1205%	75.8066%

#	Drop_variables
1	loan_amount,property_value
2	property_value,cltv
3	loan_amount,income
4	loan_amount,cltv
5	property_value,income
6	income,cltv
7	loan_amount,tract_owner_occupied_units
...	...
49	ffiec_msa_md_median_family_incom,property_value
50	tract_to_msa_income_percentage,ffiec_msa_md_median_family_incom
51	ffiec_msa_md_median_family_incom,tract_owner_occupied_units
52	ffiec_msa_md_median_family_incom,tract_median_age_of_housing_unit
53	ffiec_msa_md_median_family_incom,tract_population
54	ffiec_msa_md_median_family_incom,tract_one_to_four_family_homes
55	ffiec_msa_md_median_family_incom,tract_minority_population_percen

Ridge, Lasso, Elastic Net

Overview

- **Used for regularization and feature selection**
  - **Ridge:** deals with multicollinearity, shrinks regression coefficient towards zero
  - **Lasso:** performs feature selection, causes regression coefficient to be exact zero
  - **ElasticNet:** combines ridge and lasso, improve the regularization of statistical model

Why use Ridge, Lasso, ElasticNet?

- **Overfitting concern**
- **Feature selection**
- **Improve performance**
- **Interpretability**

Ridge Regression

Definition:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

The cost function for ridge regression:

$$Min(||Y - X(theta)||^2 + \lambda||theta||^2)$$

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. The higher the values of alpha, the bigger is the penalty and therefore the magnitude of coefficients is reduced.

Source: https://www.mygreatlearning.com/blog/what-is-ridge-regression/

Lasso Regression

Definition:

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The cost function for lasso regression:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

Source: https://www.statisticshowto.com/lasso-regression/

Elastic Net Regression

Definition:

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

The elastic net technique solves this regularization problem. For an alpha strictly between 0 and 1, and a nonnegative lambda, elastic net solves the problem:

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-x_i^T\beta\right)^2+\lambda P_\alpha(\beta)\right),$$

where:

$$P_\alpha(\beta)=\frac{(1-\alpha)}{2}\|\beta\|_2^2+\alpha\|\beta\|_1=\sum_{j=1}^{p}\left(\frac{(1-\alpha)}{2}\beta_j^2+\alpha|\beta_j|\right).$$

Source: https://www.mathworks.com/help/stats/lasso-and-elastic-net.html
https://corporatefinanceinstitute.com/resources/data-science/elastic-net/

Ridge Regression: alpha = 1

Data Set: Train: 500,000
Test: 500,000
X: All variables after standardization
Y: EAL Rating / Risk Rating

Top Variables: Variables after shrinkage
Selection Method: Coefficient and p-value

Model	X	Y		MSE	Train R2	Test R2	K-S Ratio
Ridge	All	RR		11.7681%	9.8300%	10.2700%	35.7359%
Ridge	All	ER		12.7206%	11.5600%	12.34%	35.3337%
Ridge	All	RS (Bisection)		20.9395%	15.5300%	16.2400%	36.0132%
Ridge	All	ES (Bisection)		21.3530%	14.1200%	14.5900%	33.5977%
Ridge	All	RR_Reverse		20.5091%	14.5700%	15.2300%	35.2327%
Ridge	All	ER_Reverse		21.2607%	13.9400%	14.5300%	33.6782%
Ridge	Top 10	RR		11.8749%	8.8400%	9.4500%	34.1424%
Ridge	Top 10	ER		12.7661%	11.0700%	12.0200%	34.8726%
Ridge	Top 11	RS (Bisection)		21.2135%	14.2500%	15.1500%	34.7820%
Ridge	Top 10	ES (Bisection)		21.4994%	13.3500%	14.0000%	32.7045%
Ridge	Top 10	RR_Reverse		21.0025%	12.6200%	13.1900%	32.9456%
Ridge	Top 10	ER_Reverse		21.3940%	13.2000%	14.0000%	32.7310%

Lasso Regression: alpha = 0.01

Model	X	Y		MSE	Train R2	Test R2	K-S Ratio
Lasso	All	RR		12.0479%	7.4800%	8.1400%	32.0831%
Lasso	All	ER		12.9651%	9.8700%	10.6500%	33.5276%
Lasso	All	RS (Bisection)		21.3739%	13.7100%	14.5000%	34.6128%
Lasso	All	ES (Bisection)		21.7245%	12.5100%	13.1000%	31.3856%
Lasso	All	RR_Reverse		20.9927%	12.4800%	13.2300%	33.4077%
Lasso	All	ER_Reverse		21.6510%	12.2700%	12.9600%	31.4097%
Lasso	Top 10	RR		11.9865%	7.8200%	8.6000%	31.7478%
Lasso	Top 11	ER		12.8875%	10.2300%	11.1900%	33.2388%
Lasso	Top 14	RS (Bisection)		21.3299%	13.6500%	14.6800%	33.9445%
Lasso	Top 12	ES (Bisection)		21.6240%	12.7600%	13.5000%	31.9396%
Lasso	Top 15	RR_Reverse		20.9370%	12.4200%	13.4600%	32.6420%
Lasso	Top 12	ER_Reverse		21.5359%	12.5300%	13.4300%	31.9200%

Elastic Net Regression: alpha = 0.1, l1 ratio = 0.1

Model	X	Y		MSE	Train R2	Test R2	K-S Ratio
ElasticNet	All	RR		12.0912%	7.2000%	7.8100%	32.1662%
ElasticNet	All	ER		12.9981%	9.6900%	10.4200%	33.5611%
ElasticNet	All	RS (Bisection)		21.5367%	13.1200%	13.8500%	34.4436%
ElasticNet	All	ES (Bisection)		21.8128%	12.2100%	12.7500%	31.0826%
ElasticNet	All	RR_Reverse		21.1416%	11.9200%	12.6100%	33.3196%
ElasticNet	All	ER_Reverse		21.7325%	12.0000%	12.6400%	31.1667%
ElasticNet	Top 10	RR		11.9865%	7.8200%	8.6000%	31.7486%
ElasticNet	Top 11	ER		12.8875%	10.2300%	11.1900%	33.2401%
ElasticNet	Top 14	RS (Bisection)		21.3299%	13.6500%	14.6800%	33.9457%
ElasticNet	Top 12	ES (Bisection)		21.6241%	12.7600%	13.5000%	31.9377%
ElasticNet	Top 15	RR_Reverse		20.9370%	12.4200%	13.4600%	32.6402%
ElasticNet	Top 11	ER_Reverse		21.6062%	12.3000%	13.1400%	31.1045%

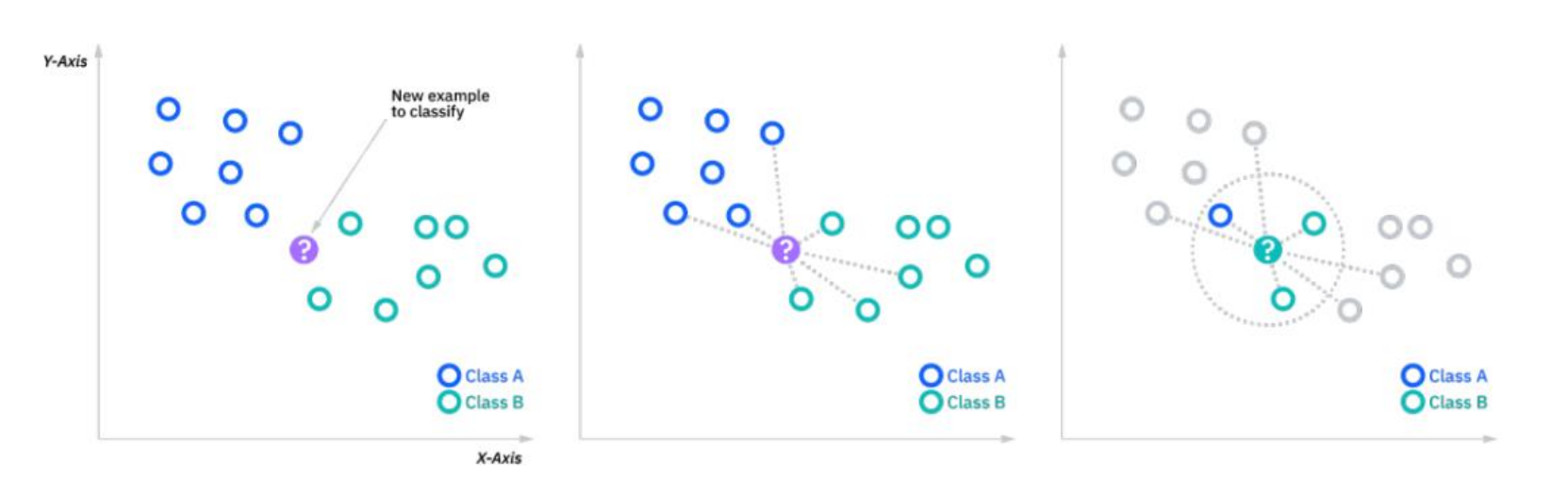
KNN (K-Nearest Neighbors Algorithm)

Overview

- On-parametric, supervised learning classifier
- Make classifications or predictions about the grouping of an individual data point

Why KNN?

- Simplicity
- Versatility
- Robustness
- Interpretability
- Effectiveness

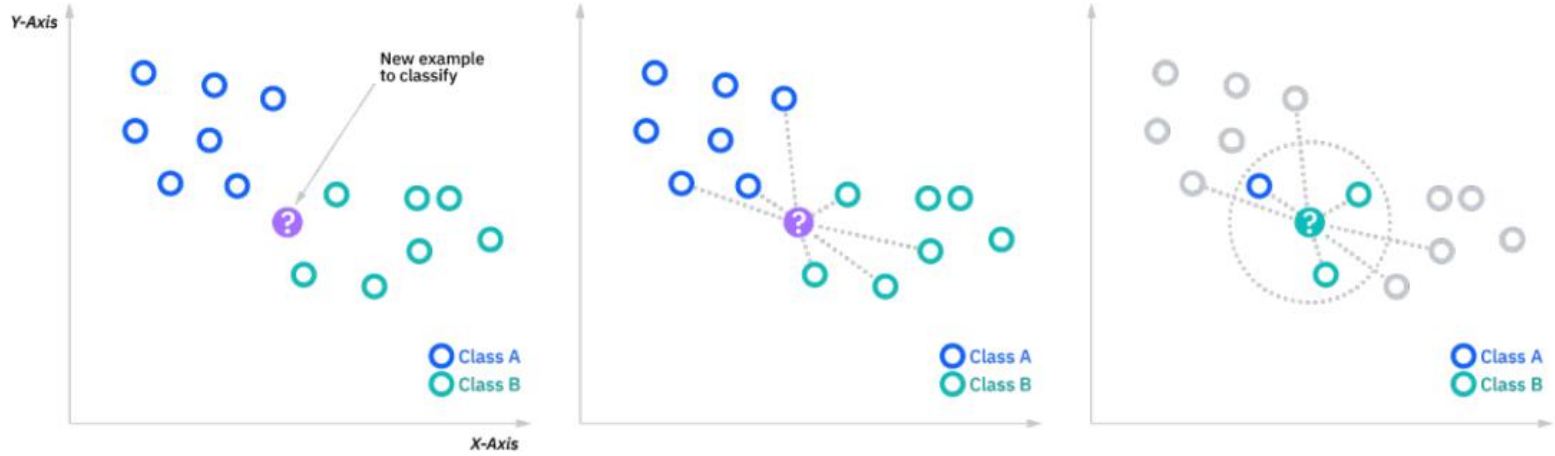


KNN (K-Nearest Neighbors Algorithm)

Definition:

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used. While this is technically considered “plurality voting”, the term, “majority vote” is more commonly used in literature. The distinction between these terminologies is that “majority voting” technically requires a majority of greater than 50%, which primarily works when there are only two categories.



KNN (K-Nearest Neighbors Algorithm)

Defining k:

The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. Due to the different variables we use, the optimal k varys from 18 to 27. To have a same stander, we use k=25 as the default value for all knn models.

```
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=25)
```

Determine distance metrics:

Euclidean distance (p=2): This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

$$d(A,B) = \sqrt{(x_{1a} - x_{1b})^2 + (x_{2a} - x_{2b})^2 + \cdots + (x_{3a} - x_{3b})^2}$$
$$= \sqrt{\sum_{i=1}^{n} (x_{ia} - x_{ib})^2}$$

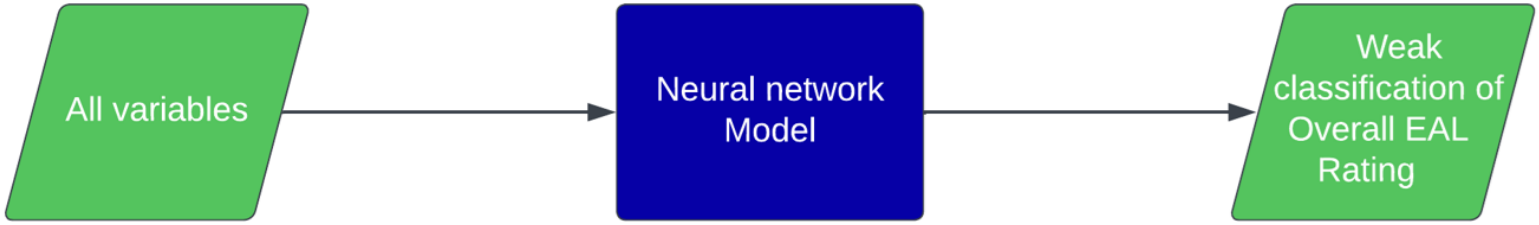
KNN (K-Nearest Neighbors Algorithm)

Data Set: Train: 500,000
Test: 500,000
X: All variables after standardization
Y: Eal Rating or Risk Rating
Top Variables: Tract Only

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
KNN	All	RR		15.4466%	84.5534%	60.6852%	1.4377%	18.7139%
KNN	All	ER		17.2154%	82.7846%	63.0383%	5.4518%	21.2554%
KNN	All	RS (Bisection)		38.7356%	61.2644%	61.8596%	58.7527%	22.5287%
KNN	All	ES (Bisection)		39.1028%	60.8972%	61.2961%	59.1237%	21.7943%
KNN	All	RR_Reverse		36.5234%	63.4766%	65.2140%	81.6116%	22.1095%
KNN	All	ER_Reverse		38.6768%	61.3232%	61.9574%	71.8666%	21.8881%
KNN	Tract	RR		5.6784%	94.3216%	86.2965%	75.3974%	85.1384%
KNN	Tract	ER		5.5080%	94.4920%	88.3859%	79.1240%	85.8591%
KNN	Tract	RS (Bisection)		10.4122%	89.5878%	89.9725%	89.1063%	79.1756%
KNN	Tract	ES (Bisection)		10.3548%	89.6452%	89.7708%	89.4864%	79.2904%
KNN	Tract	RR_Reverse		9.7932%	90.2068%	90.9549%	92.6061%	79.9468%
KNN	Tract	ER_Reverse		10.0512%	89.9488%	90.0317%	91.3336%	79.6876%

Neural Network Classification Model

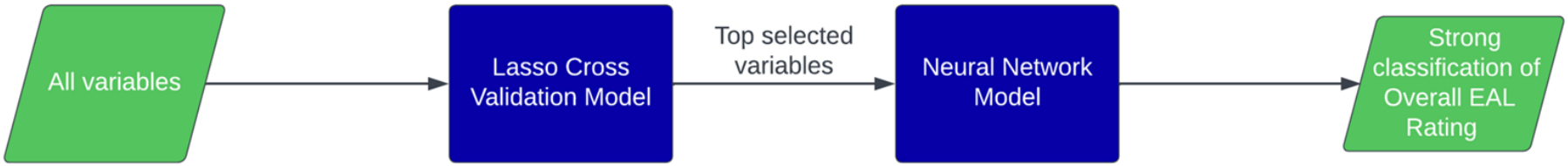
Using All Variables



Reasons for weak results

- *Overfitting*
- *Noisy Features*
- *Multicollinearity*
- *Curse of dimensionality*

Using Feature Selection



Selected Features

- *All Tract variables*
- *CLTV*
- *Race: Asian & Black*
- *Property values*
- *Loan Amount*
- *Loan Amount*

Neural Network Classification Model

Output with all features

				MSE	Accuracy	Recall	Precision	K-S Ratio
Neural Network	All	RR		9.40%	87.42%	73.81%	29.42%	52.38%
Neural Network	All	ER		9.78%	87.01%	78.56%	36.10%	52.94%
Neural Network	All	RS (Bisection)		15.64%	76.67%	76.94%	76.16%	53.39%
Neural Network	All	ES (Bisection)		16.22%	75.33%	76.01%	74.03%	50.79%
Neural Network	All	RR_Reverse		14.97%	78.00%	78.74%	85.89%	54.24%
Neural Network	All	ER_Reverse		15.90%	76.15%	75.44%	82.18%	51.79%

Output with selected features

				MSE	Accuracy	Recall	Precision	K-S Ratio
Neural Network	Top10	RR		10.24%	86.75%	72.75%	23.42%	44.09%
Neural Network	Top11	ER		7.66%	89.91%	80.57%	56.27%	64.26%
Neural Network	Top14	RS (Bisection)		12.46%	81.74%	81.66%	81.86%	63.54%
Neural Network	Top12	ES (Bisection)		12.12%	82.31%	81.68%	83.30%	64.88%
Neural Network	Top15	RR_Reverse		11.50%	83.49%	83.92%	89.08%	65.52%
Neural Network	Top12	ER_Reverse		12.07%	82.44%	83.16%	84.25%	64.62%

XGBoost (eXtreme Gradient Boosting)

Overview

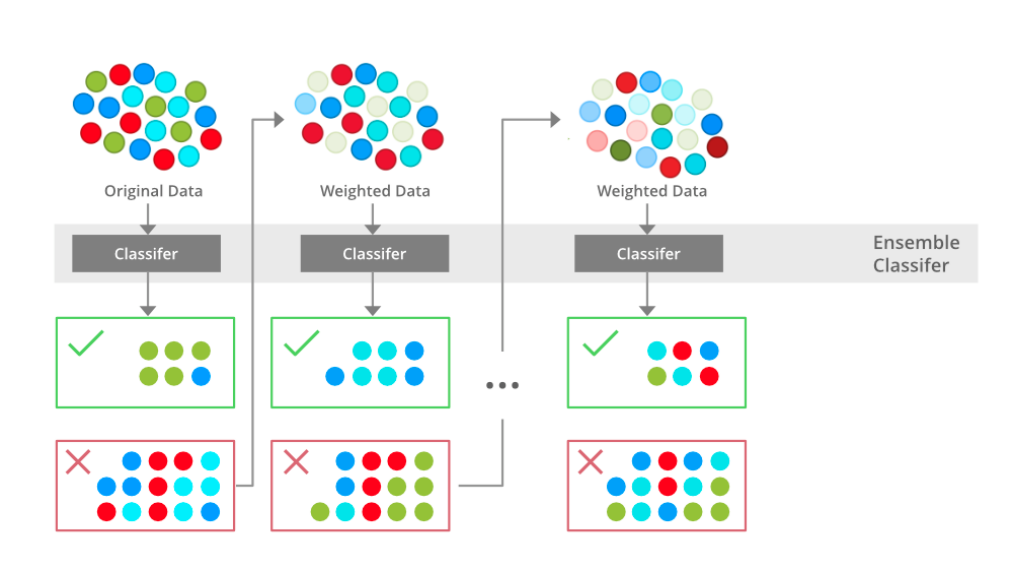
- XGBoost is a open-source machine learning library that provides implementation of gradient boosting algorithms
- Used for classification, regression and ranking problems

Principle

- Creating a series of decision trees which are trained sequentially
- Each subsequent tree tries to correct the errors of the previous trees
- Combining weak models to generate a collectively strong model

Why XGBoost?

- **Performance**
- **Flexibility**
- **Feature Importance**
- **Regularization**
- **Handling Missing Values**
- **Built-in CV**
- **Interpretability**



XGBoost Output

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
XGBoost	All	RR		6.2870%	93.7130%	91.3298%	65.7460%	79.9244%
XGBoost	All	ER		5.9912%	94.0088%	94.2221%	70.2940%	80.6712%
XGBoost	All	RS (Bisection)		11.4614%	88.5386%	89.0817%	87.8435%	77.0908%
XGBoost	All	ES (Bisection)		11.6546%	88.3454%	88.4849%	88.1632%	76.7551%
XGBoost	All	RR_Reverse		10.9008%	89.0992%	89.0024%	93.0120%	77.9035%
XGBoost	All	ER_Reverse		11.0296%	88.9704%	88.5850%	91.1371%	77.6917%
XGBoost	Top 11	RR		6.2984%	93.7016%	91.4442%	65.5657%	80.1852%
XGBoost	Top 11	ER		6.0060%	93.9940%	94.1067%	70.3019%	81.1272%
XGBoost	Top 11	RS (Bisection)		11.0480%	88.9520%	89.3832%	88.4043%	77.9264%
XGBoost	Top 11	ES (Bisection)		11.1208%	88.8792%	89.4283%	88.1820%	77.8379%
XGBoost	Top 11	RR_Reverse		10.8464%	89.1536%	88.9899%	93.1340%	77.5658%
XGBoost	Top 11	ER_Reverse		10.8966%	89.1034%	88.7267%	91.2335%	78.0584%

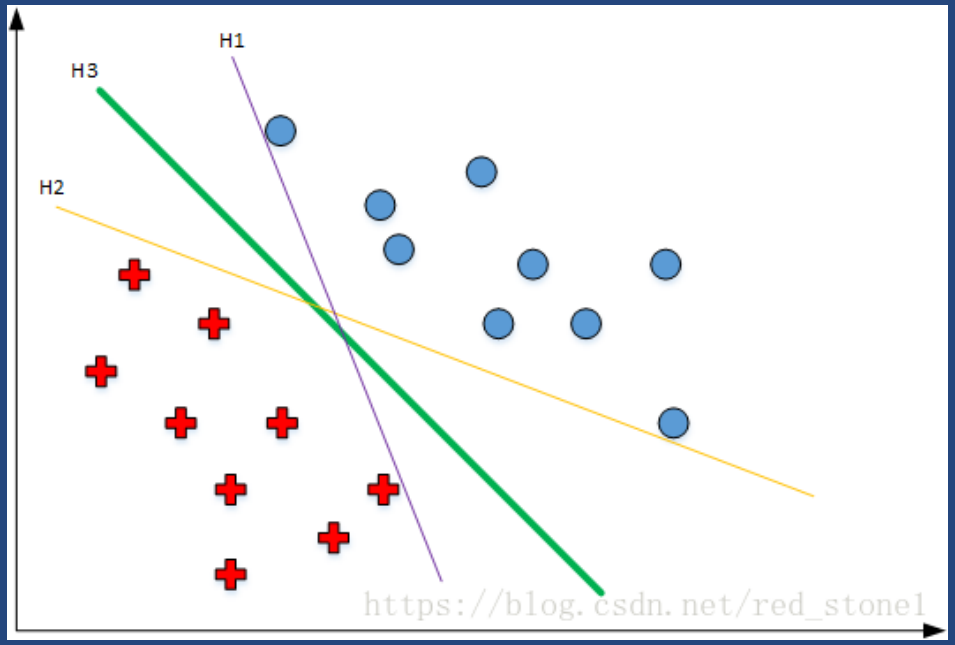
Support Vector Machines (SVM)

Strengths:

- High dimensional data
- Non-linear relationship

Weakness:

- Very computationally consuming
- Not interpretive



Support Vector Machines (SVM)

Support Vector Machines (SVMs) are a type of machine learning model that is widely used for classification and regression tasks. SVMs work by identifying a hyperplane in a high-dimensional space that best separates the data into different classes. The hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the closest data points from each class.

SVMs are particularly useful when dealing with high-dimensional data and when the data is not linearly separable. They can also handle noisy data and outliers well. SVMs have been successfully used in a variety of applications, including text classification, image classification, and bioinformatics.

One of the strengths of SVMs is their ability to handle complex decision boundaries. This makes them particularly useful when dealing with non-linearly separable data. SVMs also have a strong theoretical foundation, which makes them well-suited for problems where interpretability is important.

Support Vector Machines (SVM)

Data Set: Train: 500,000
Test: 500,000
X: All variables after standardization
Y: Eal Rating or Risk Rating
Kernel Function: RBF (Non-linear)
Top Variables: Tract Only
Selection Method: Forward Stepwise

Model	X	Y	Variables	MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
SVM	All	RR	All	15.111%	84.889%	81.225%	3.467%	31.288%
SVM	All	ER	All	16.412%	83.588%	78.677%	9.348%	31.244%
SVM	All	RS (Bisection)	All	29.374%	70.626%	71.369%	68.886%	41.260%
SVM	All	ES (Bisection)	All	30.409%	69.591%	70.589%	67.163%	39.194%
SVM	All	RR_Reverse	All	28.818%	71.182%	71.496%	85.050%	40.326%
SVM	All	ER_Reverse	All	30.417%	69.583%	68.852%	78.835%	38.934%
SVM	Tract	RR	Tract	14.320%	85.680%	85.913%	9.287%	35.738%
SVM	Tract	ER	Tract	15.720%	84.280%	84.538%	13.150%	36.193%
SVM	Tract	RS (Bisection)	Tract	29.606%	70.394%	73.081%	64.573%	41.552%
SVM	Tract	ES (Bisection)	Tract	29.939%	70.061%	72.706%	64.234%	40.338%
SVM	Tract	RR_Reverse	Tract	28.407%	71.593%	70.596%	88.843%	41.018%
SVM	Tract	ER_Reverse	Tract	29.734%	70.266%	68.692%	81.672%	40.541%

Borrower Characteristics versus ER / RR

Objective: Explore relation between only borrower features and Risk rating / EAL rating.
Method: Exclude tract features from independent variables to highlight the importance of borrower features.
Target Variable: RR/ER
Independent Variables: All variables at first except tract ones
Train Set Size: 500,000
Test Set Size: 500,000
Model: Random Forest. Use grid search to adjust parameters. Set f1 score as standard

X	Y	MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Borrower	RR	15.5506%	84.4494%	42.6590%	0.4754%	12.8061%
Borrower	ER	17.6178%	82.3822%	49.7013%	2.4562%	15.2629%

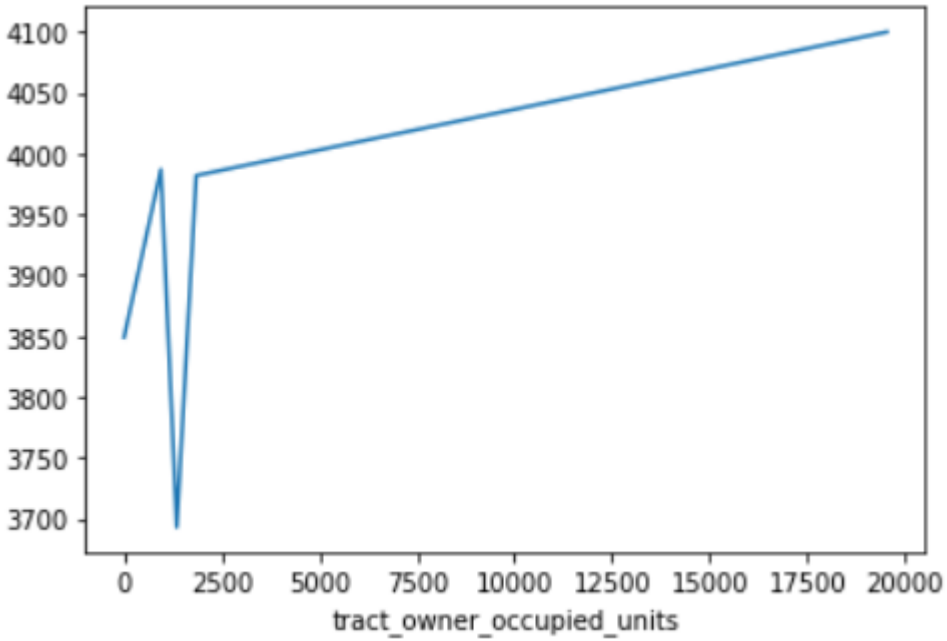
Individual loan features versus Tract Features

Objective: Individual loan features - Tract features
Target Variable: Each Tract Feature
Independent Variables: All variables except seven tract-level ones
Model: Random Forest

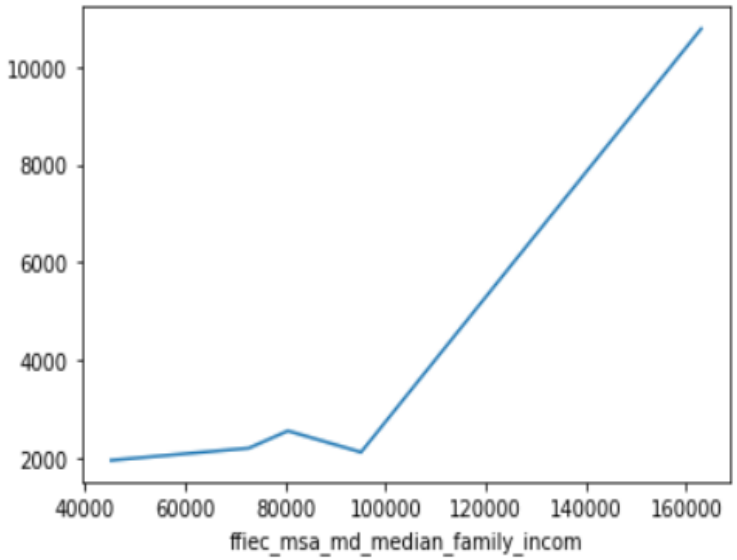
X	Y	MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Loan	ffiec_msa_md_median_family_incom	35.7630%	64.2370%	63.3876%	65.4789%	28.6004%
Loan	tract_one_to_four_family_homes	43.2536%	56.7464%	55.9991%	62.5273%	13.5495%
Loan	tract_population	43.9368%	56.0632%	54.8465%	68.5007%	12.1496%
Loan	tract_minority_population_percen	34.3080%	65.6920%	71.0678%	52.9174%	31.4366%
Loan	tract_owner_occupied_units	42.2308%	57.7692%	56.4258%	67.9624%	15.5996%
Loan	tract_to_msa_income_percentage	34.8894%	65.1106%	64.4539%	66.2250%	30.2988%
Loan	tract_median_age_of_housing_unit	38.6596%	61.3404%	66.4742%	44.0514%	22.5630%

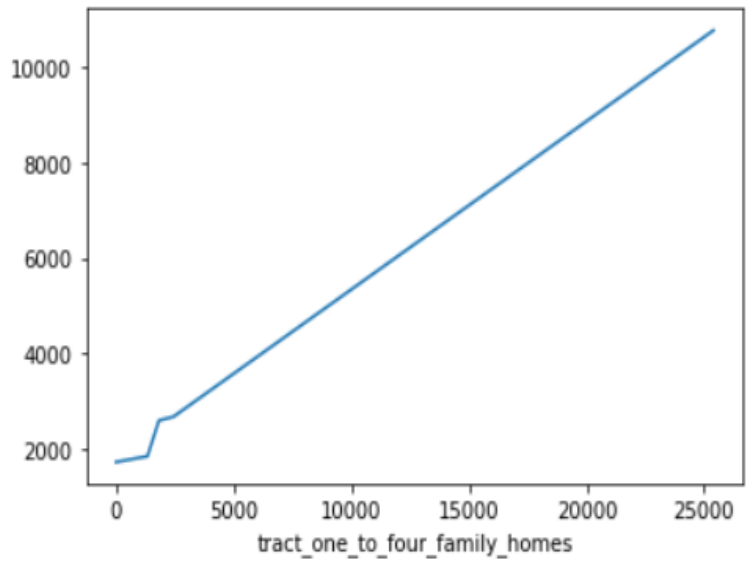
Model Inference

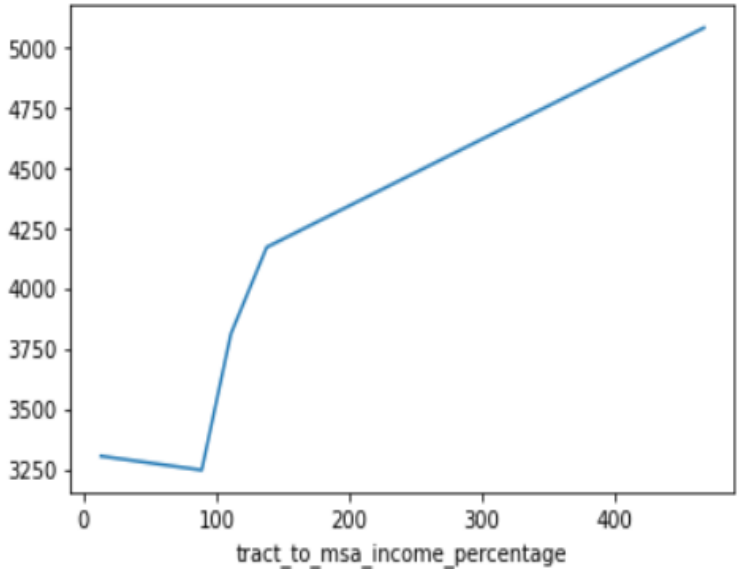
Objective: Explain how champion model made classification.
Target Variable: ER
Independent Variables: Tract variables.
Method: Set 5 quantiles as the representative values of each independent variable. Create cartesian product and make predictions. Count the number of data with a dependent variable of 1 at each level.
Model: Random Forest.



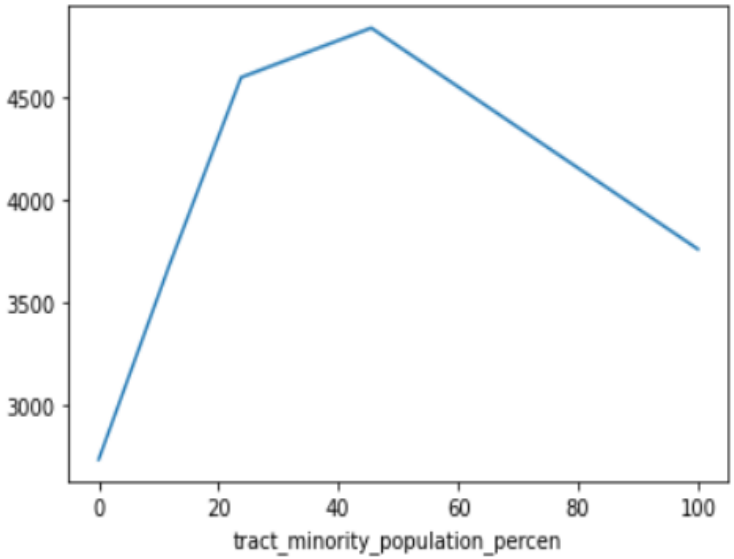
Model Inference

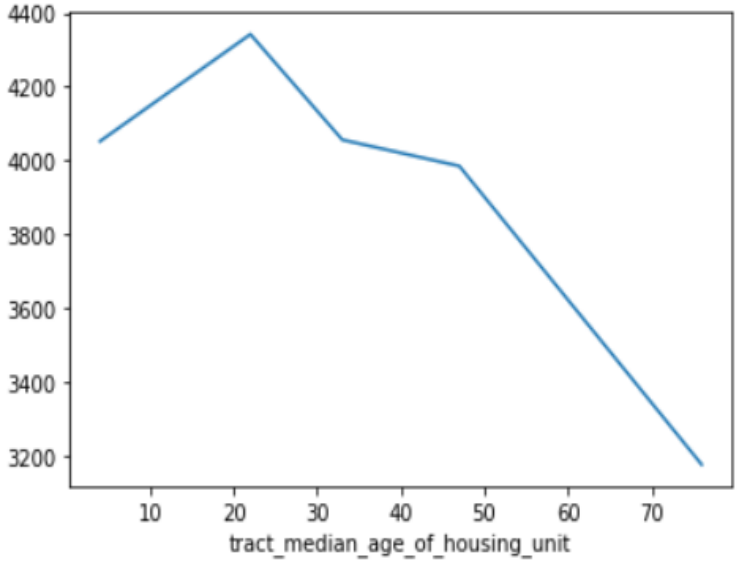


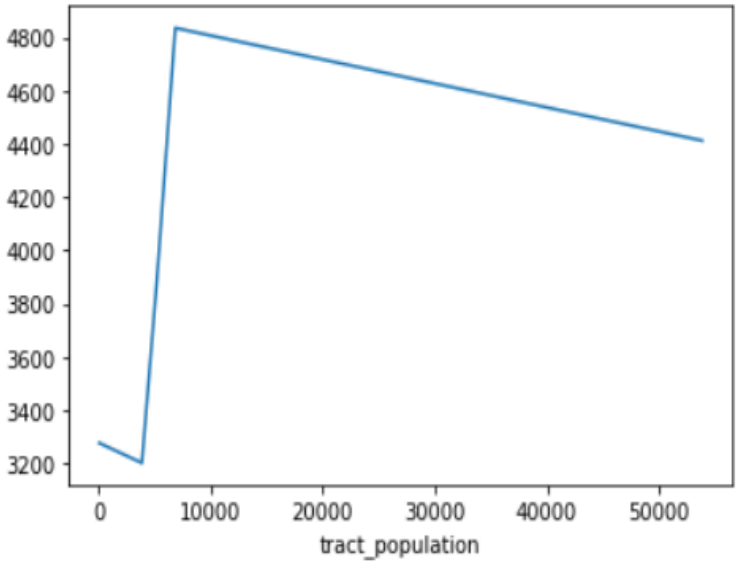




Model Inference

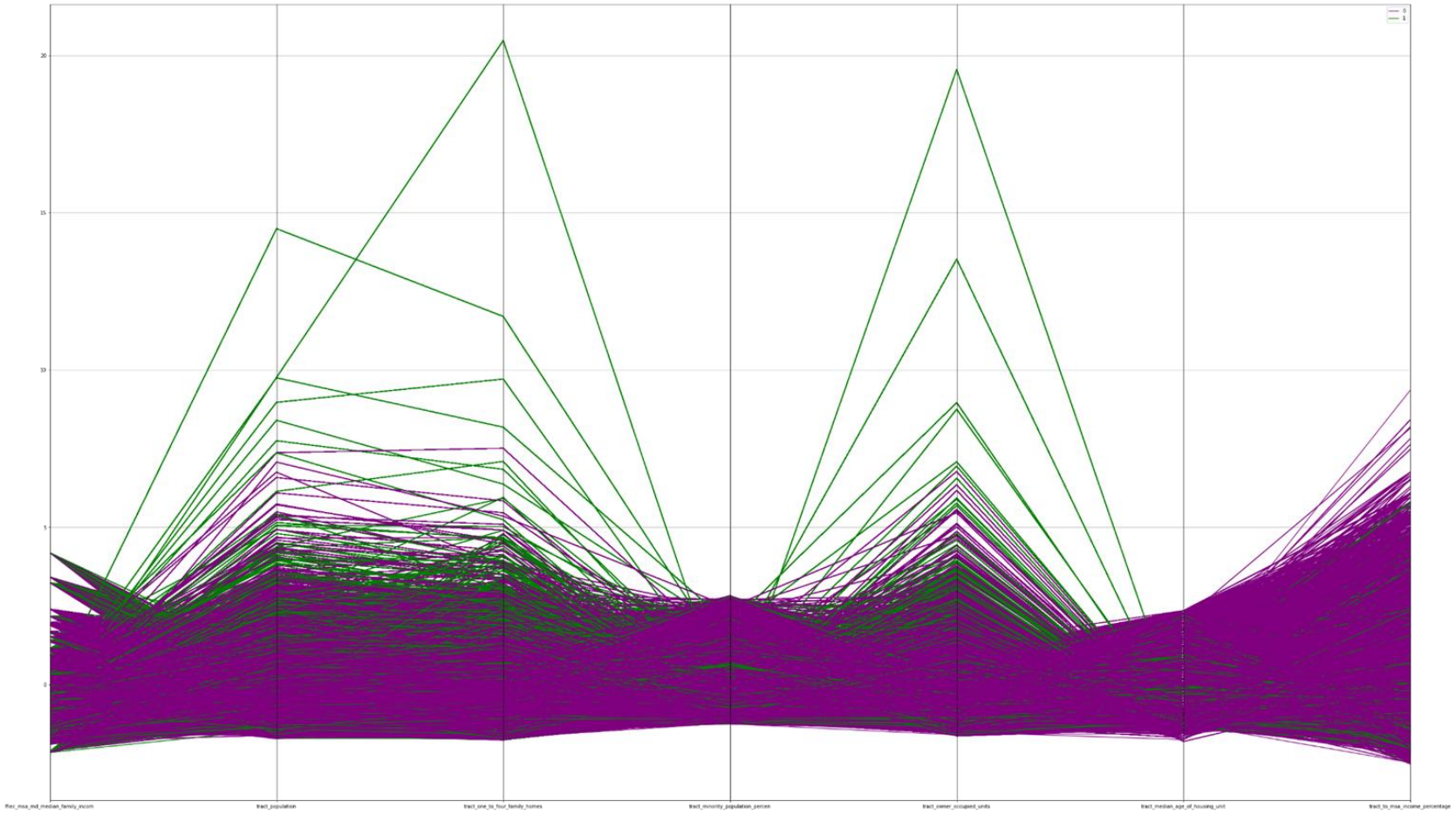






Parallel Coordinate Chart of Tract Variables

Greenline: RR = 1, Purple Line: RR = 0



Logistics

Definition:

In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.

Model:

The logistic function is of the form:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

where μ is a location parameter (the midpoint of the curve, where $p(\mu) = 1/2$) and s is a scale parameter. This expression may be rewritten as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where $\beta_0 = -\mu/s$ and is known as the intercept (it is the *vertical* intercept or *y*-intercept of the line $y = \beta_0 + \beta_1 x$), and $\beta_1 = 1/s$ (inverse scale parameter or rate parameter): these are the *y*-intercept and slope of the log-odds as a function of *x*. Conversely, $\mu = -\beta_0/\beta_1$ and $s = 1/\beta_1$.

Logistics Model Results

Model	X	Y		MSE	Accuracy Score	Recall Score	Precision Score	K-S Ratio
Logistics	All	RR		15.3176%	84.6824%	54.6765%	7.8170%	35.2612%
Logistics	All	ER		16.6790%	83.3210%	64.4270%	11.8358%	35.0353%
Logistics	All	RS (Bisection)		31.9992%	68.0008%	68.9595%	65.4712%	36.1645%
Logistics	All	ES (Bisection)		33.1424%	66.8576%	67.9872%	63.7131%	33.8332%
Logistics	All	RR_Reverse		31.5922%	68.4078%	69.2541%	83.5175%	35.4336%
Logistics	All	ER_Reverse		32.9548%	67.0452%	66.6230%	77.0115%	33.9234%
Logistics	Top10	RR		15.3768%	84.6232%	53.6075%	7.1020%	34.4727%
Logistics	Top10	ER		16.7606%	83.2394%	63.2760%	11.5281%	34.5828%
Logistics	Top10	RS (Bisection)		32.7366%	67.2634%	68.3212%	64.3752%	34.7148%
Logistics	Top10	ES (Bisection)		33.6692%	66.3308%	67.6660%	62.5471%	32.7712%
Logistics	Top10	RR_Reverse		32.6276%	67.3724%	67.9127%	84.7053%	33.0255%
Logistics	Top10	ER_Reverse		33.4072%	66.5928%	66.0037%	77.5051%	32.8685%

Logistics ER High

	All Variables	Coefficient	P-value
1st	tract one to four family homes	0.6472	0.00
2nd	tract owner occupied units	-0.4768	0.00
3rd	tract population	0.3472	0.00
4th	loan amount	0.2057	0.00
5th	conventional	-0.1545	0.00
6th	cltv	-0.1277	0.00
7th	fha	-0.1089	0.00
8th	va	-0.1034	0.00
9th	ffiec msa md median family incom	0.1023	0.00
10th	tract median age of housing unit	-0.0613	0.00
11th	tract to msa income percentage	0.0594	0.00
12th	site	-0.0549	0.00
13th	total units	-0.0474	0.00
...	...	...	...

Top 10 Variables	Coefficient	P-value
tract one to four family homes	0.6033	0.00
tract owner occupied units	-0.4299	0.00
tract population	0.3437	0.00
loan amount	0.2214	0.00
conventional	-0.1592	0.00
cltv	-0.1220	0.00
fha	-0.1165	0.00
va	-0.1091	0.00
ffiec msa md median family incom	0.0915	0.00
tract median age of housing unit	-0.0789	0.00

Logistics ER Low

	All Variables	Coefficient	P-value
1st	tract one to four family homes	-0.7328	0.00
2nd	tract population	-0.6241	0.00
3rd	tract owner occupied units	0.5593	0.00
4th	loan amount	-0.4348	0.00
5th	conventional	0.3537	0.00
6th	cltv	0.2582	0.00
7th	fha	0.2424	0.00
8th	tract median age of housing unit	0.2394	0.00
9th	va	0.2147	0.00
10th	ffiec msa md median family incom	0.1623	0.00
11th	site	0.0981	0.00
12th	total units	0.0966	0.00
13th	income	0.0870	0.00
…	…	…	…

Top 10 Variables	Coefficient	P-value
tract one to four family homes	-0.7675	0.00
tract population	-0.6354	0.00
tract owner occupied units	0.6239	0.00
loan amount	-0.3852	0.00
conventional	0.3620	0.00
tract median age of housing unit	0.2572	0.00
fha	0.2562	0.00
cltv	0.2322	0.00
va	0.2185	0.00
ffiec msa md median family incom	0.1350	0.00

Logistics ER Bisection

	All Variables	Coefficient	P-value
1st	tract one to four family homes	0.6815	0.00
2nd	tract population	0.6166	0.00
3rd	tract owner occupied units	-0.5292	0.00
4th	loan amount	0.4342	0.00
5th	conventional	-0.3609	0.00
6th	tract median age of housing unit	-0.2840	0.00
7th	cltv	-0.2593	0.00
8th	fha	-0.2486	0.00
9th	va	-0.2175	0.00
10th	ffiec msa md median family incom	-0.1997	0.00
11th	total units	-0.0983	0.00
12th	site	-0.0957	0.00
13th	income	-0.0889	0.00
...	...	...	...

Top 10 Variables	Coefficient	P-value
tract one to four family homes	0.7341	0.00
tract population	0.6305	0.00
tract owner occupied units	-0.6162	0.00
loan amount	0.3733	0.00
conventional	-0.3698	0.00
tract median age of housing unit	-0.2986	0.00
fha	-0.2643	0.00
cltv	-0.2266	0.00
va	-0.2222	0.00
ffiec msa md median family incom	-0.1660	0.00