Improving SBA Loan Performance in Colorado

David Olivero

April 27th 2020 cohort

Springboard

Thanks to Springboard mentor

Branko Kovac, Logikka

The Issue

- What would you guess the default rate is for a typical bank providing business loans?

- What would you guess is the default rate for a typical bank providing *SBA (Small Business Administration)* loans?

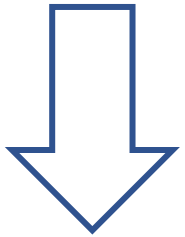
Enables small businesses

- Strategic goals: *(source: https://www.sba.gov/)*
  - **Support small business revenue and job growth;**
  - **Build healthy entrepreneurial ecosystems and create business friendly environments;**
  - **Restore small businesses and communities after disasters; and**
  - **Strengthen SBA's ability to serve small businesses.**

- Nearly $28 billion lent in FY19
- Every 1% of population default is roughly $280 million lost
- **The SBA loan default rate is > 15%**

The Question

If you wanted to reduce the SBA loan default rate, could you?



Can a model be built to predict whether a proposed loan will default?

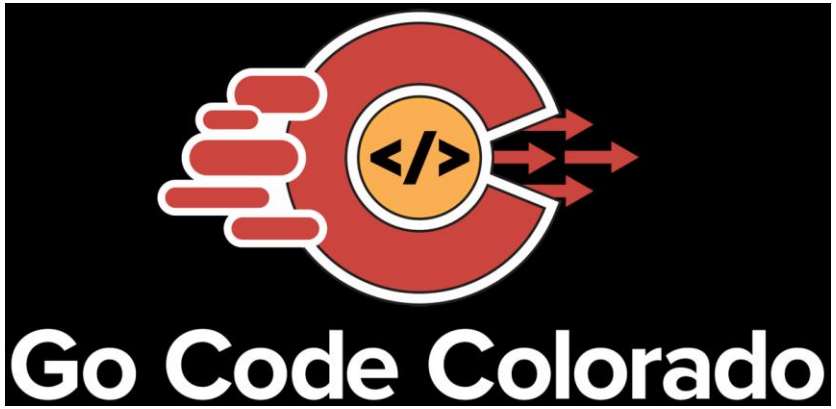
The Approach

Hypothesis: Information about place can help

Focus on loans in the state of Colorado (my home state)

GoCodeColorado :

- Census data
- Economic data
- Health data



Combine data and build a decision tree machine learning model

Colorado

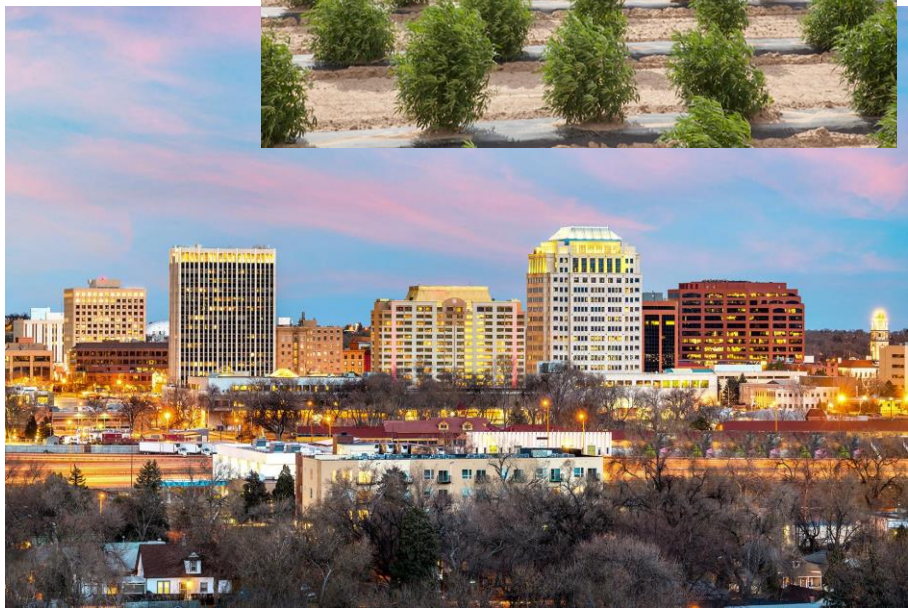
63 counties

Population 5.7 million

Several metropolitan areas

Large rural areas

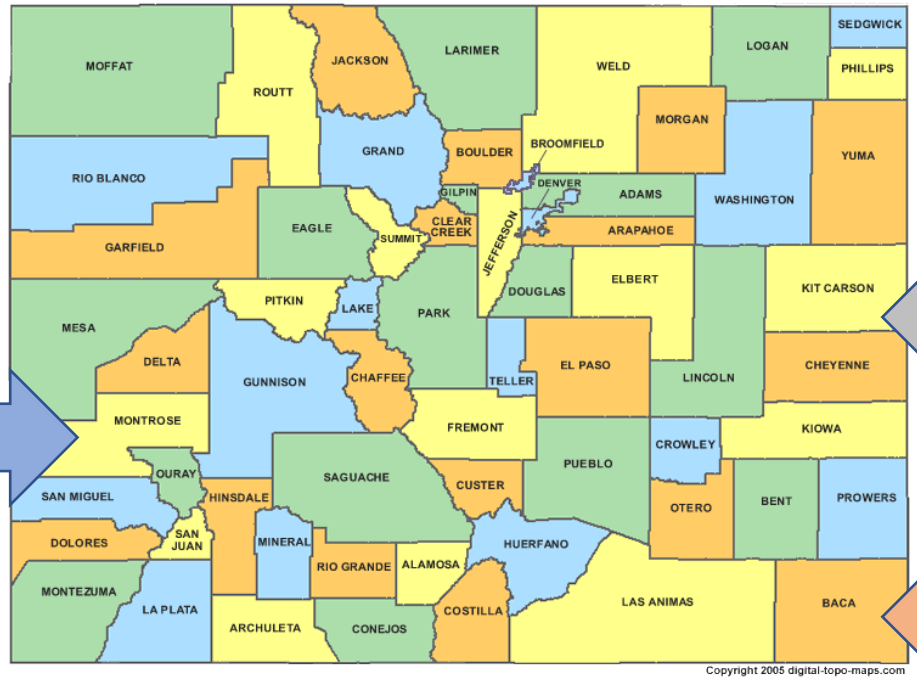
Diverse industrial sectors



Merged Data Set

SBA Data (source: https://www.kaggle.com/larsen0966/sba-loans-case-data-set)

- Loan Result
- Loan Amount
- Business City/Zip
- Loan Amount
- Low Documentation
- Revisable Line of Credit
- Approval FY
- New Business?
- Pay Delay
- NAICS (Industry Code)
- Number of Employees
- Loan Term

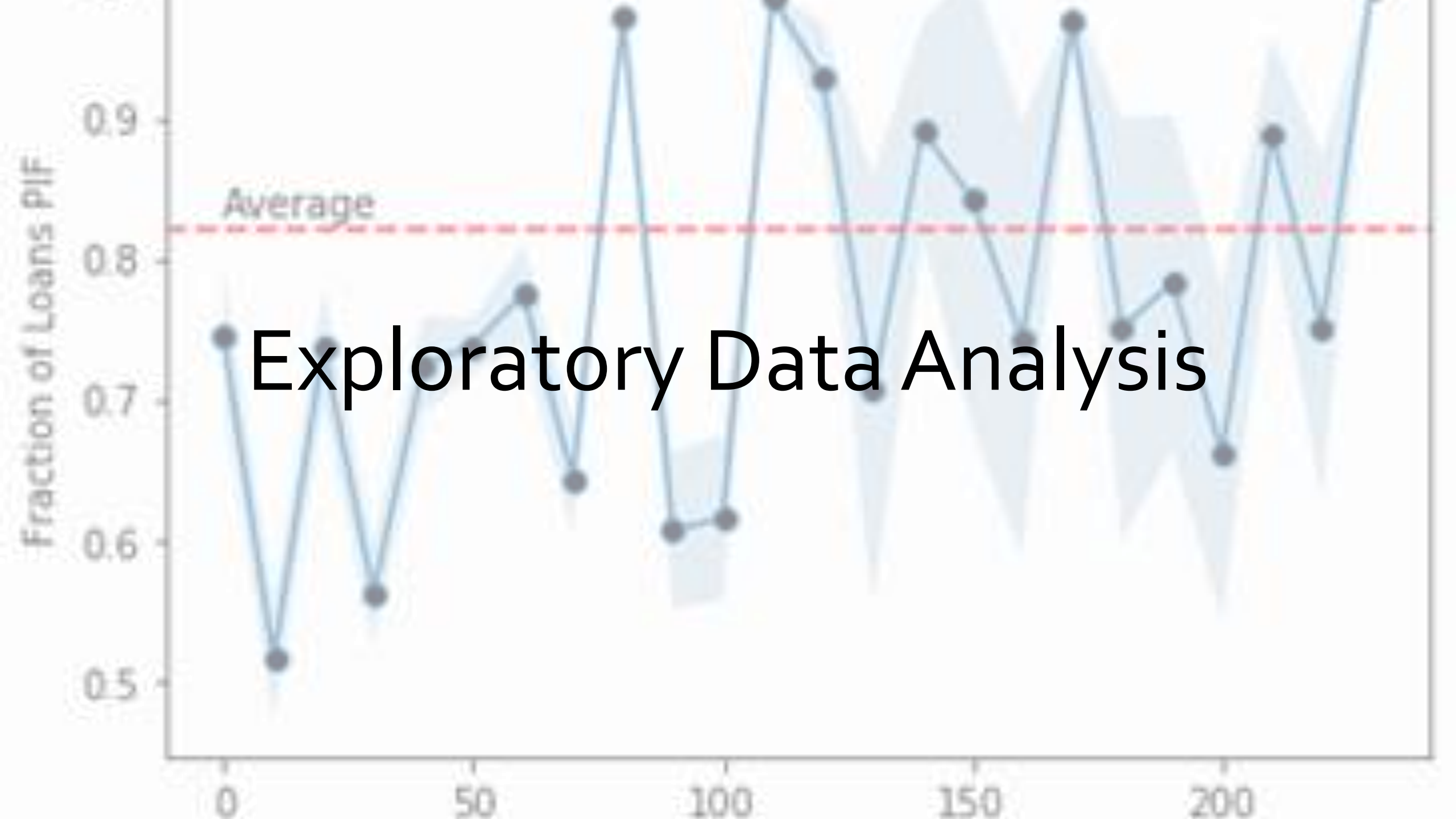


GoCodeColorado (source: https://data.colorado.gov/)

- Labor Force in County
- Unemployment Rate in County
- Median Income in County
- County Population
- % Population Change (10-yr)

CDC (source: https://wonder.cdc.gov/controller/datarequest)

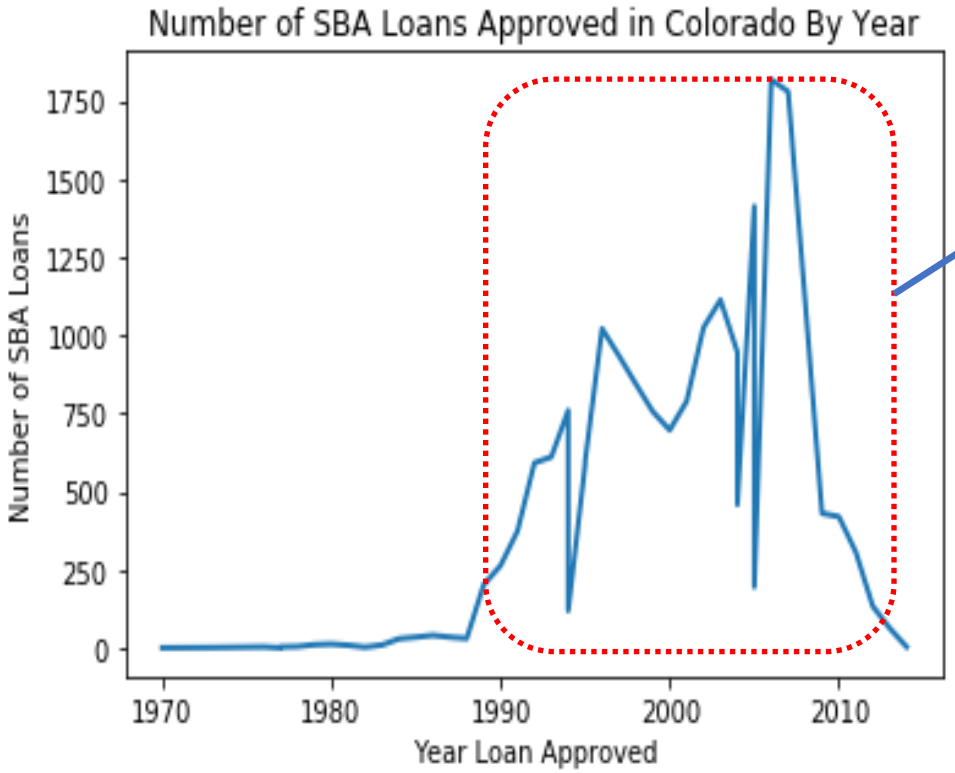
- All-cause mortality rate by county

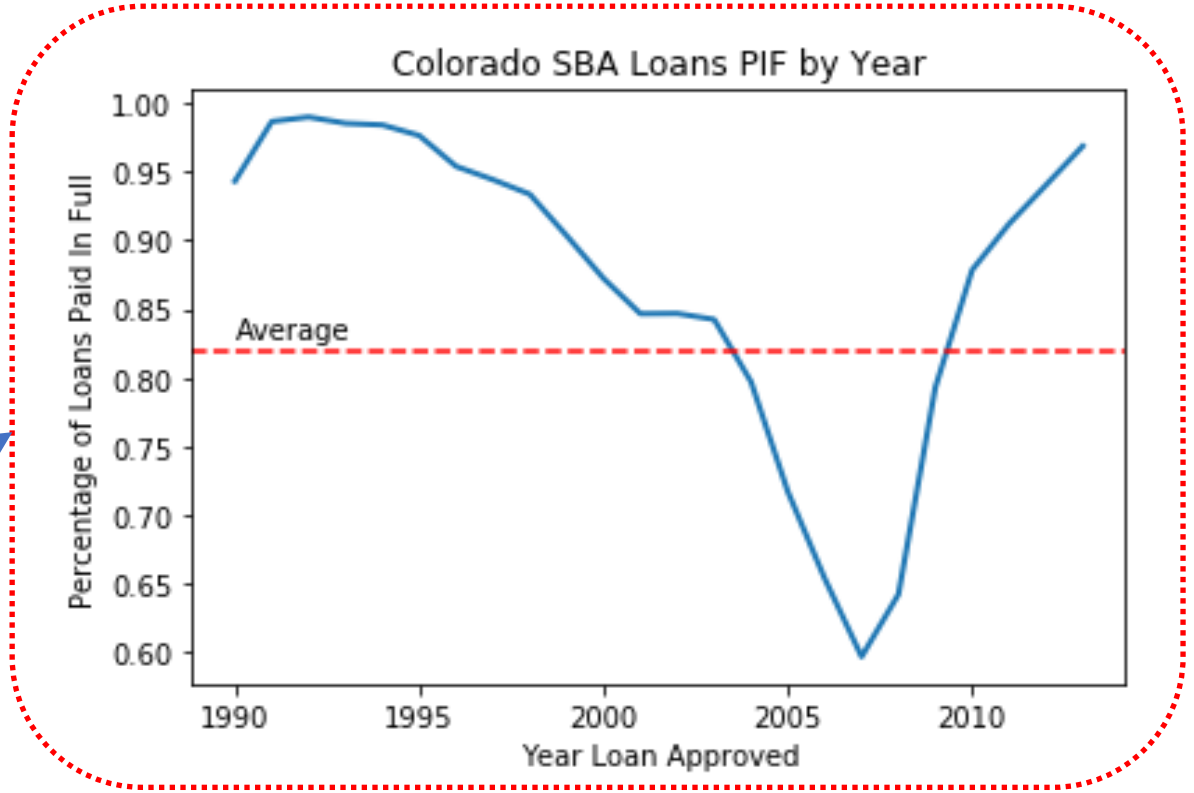


Exploratory Data Analysis

Loan Trends Over Time

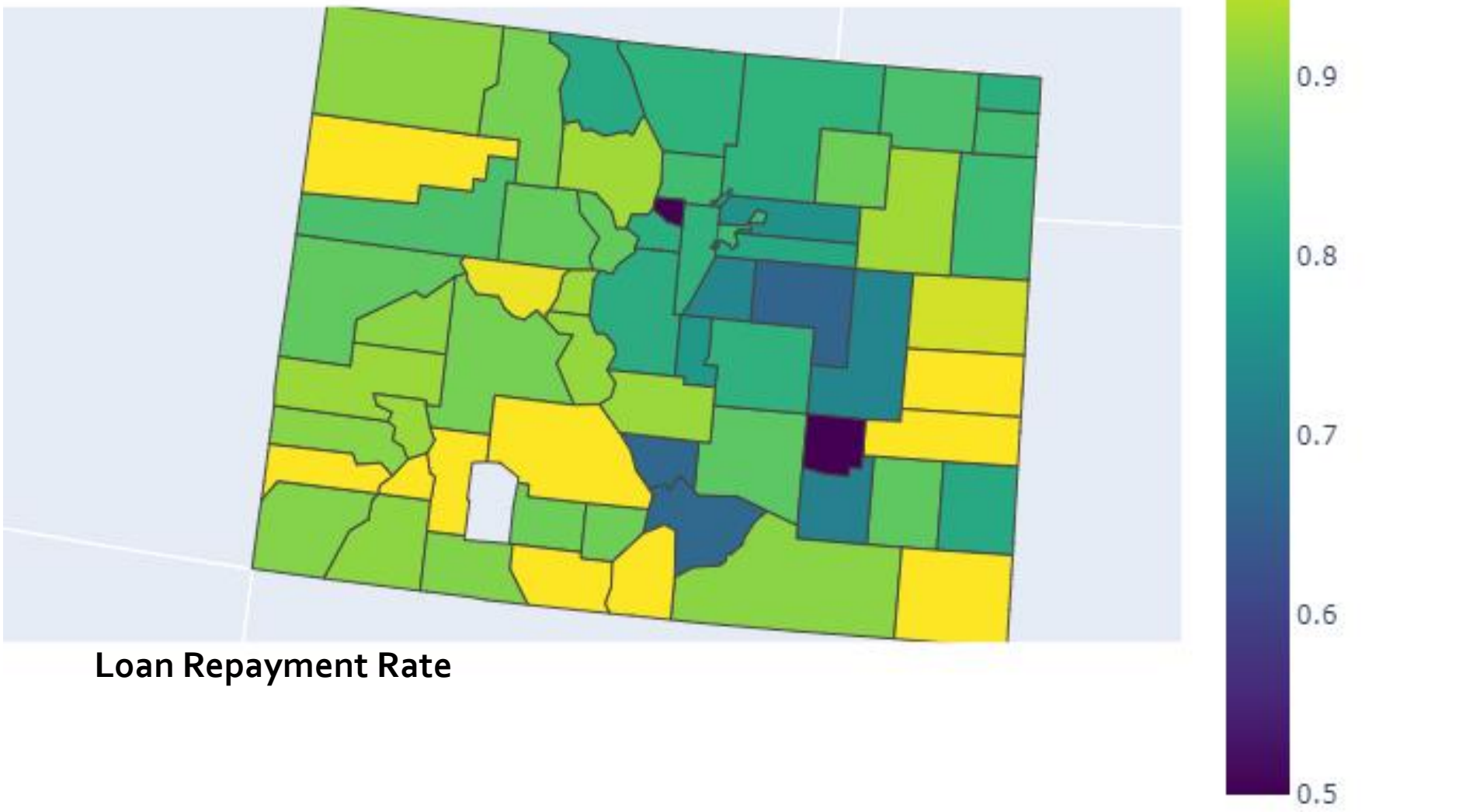
- Focus on 1990-2013
- Contains ~20,000 loans





Loan default rates have varied widely over the years

Differences between counties

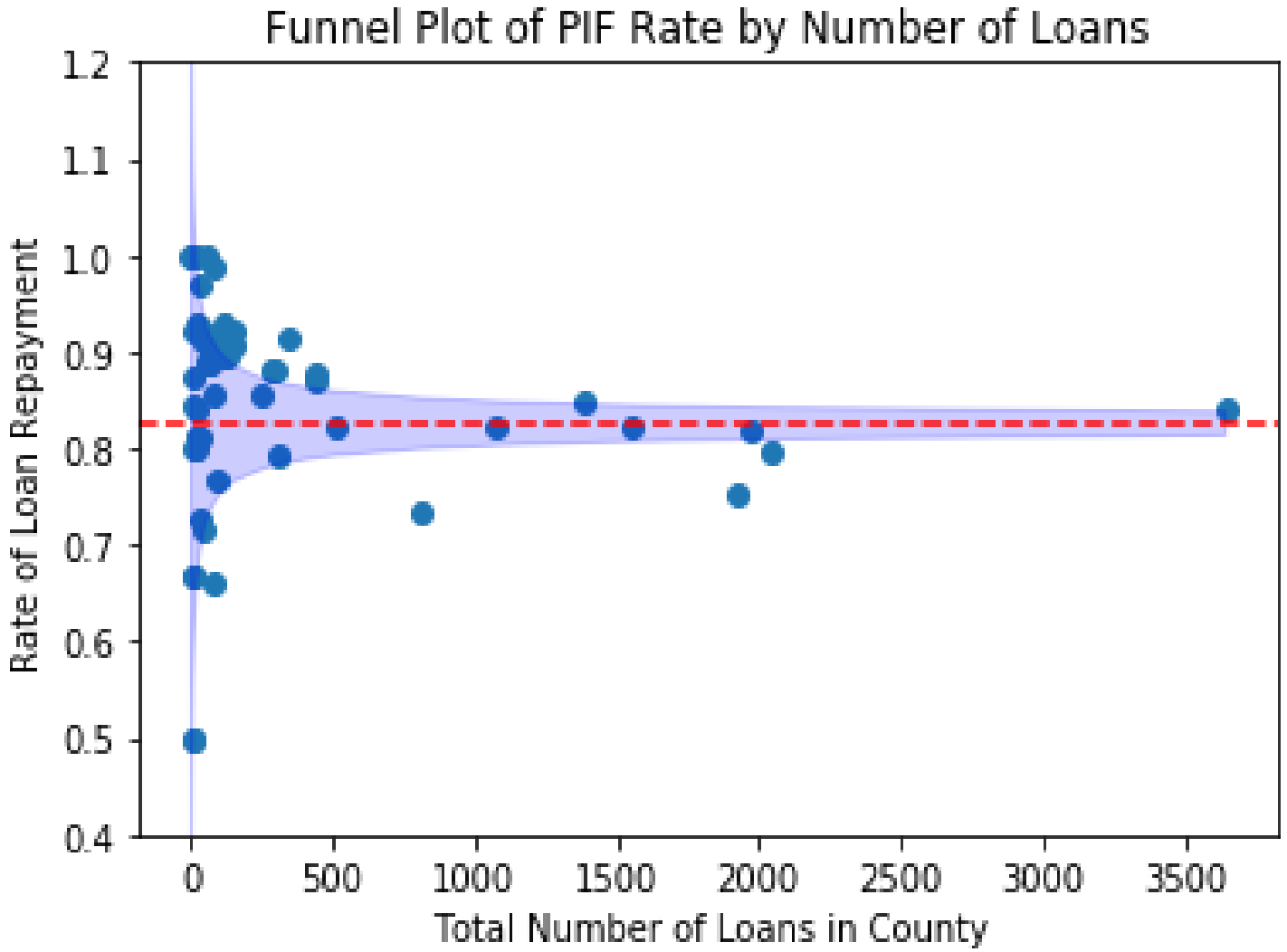


Loan Repayment Rate

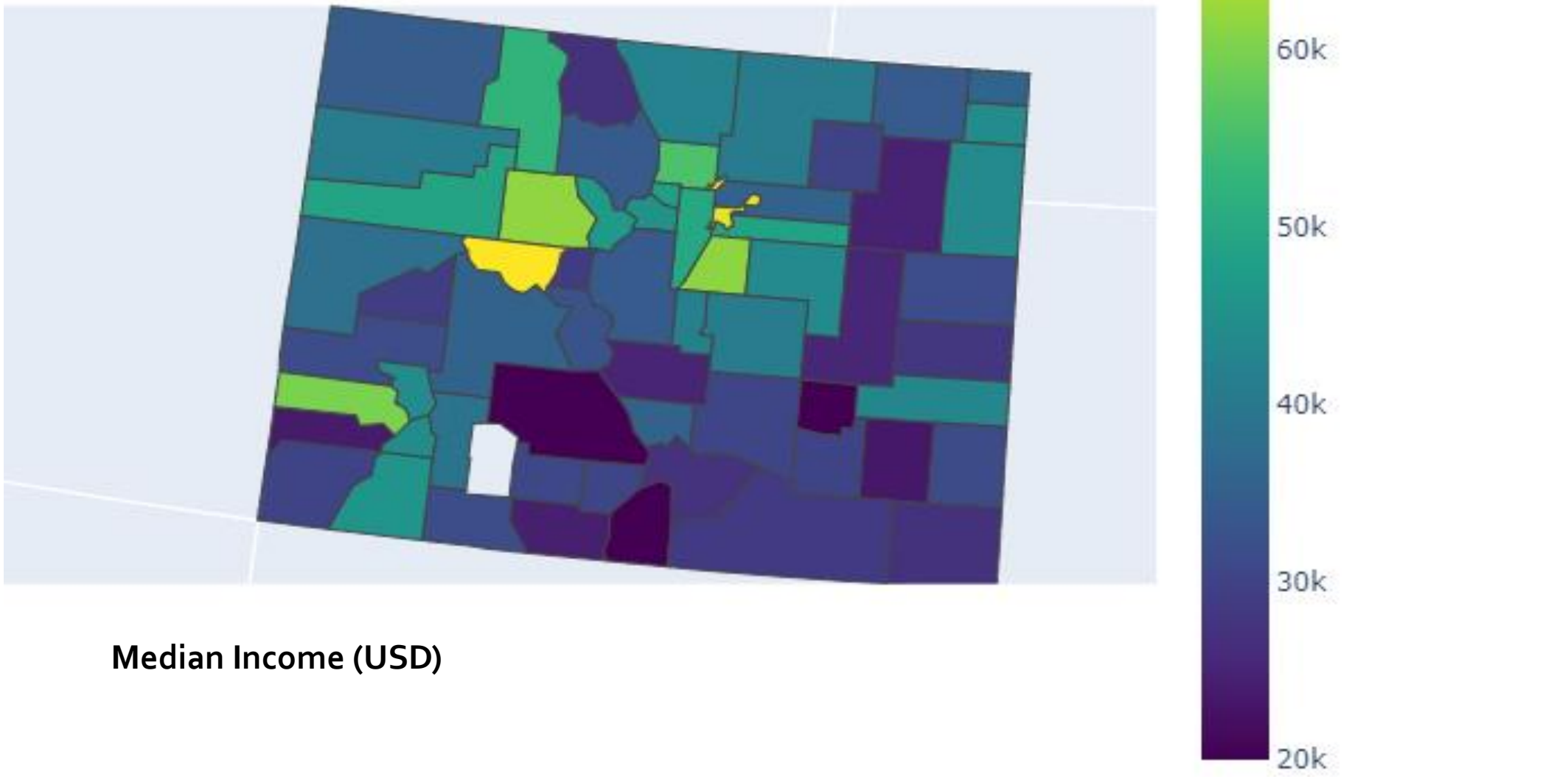
Are differences by county significant?

- More variation with low sample size
- Counties vary widely in quantity of loans

- Shaded region: Expected variation around the population mean
- Dots: Data for each county

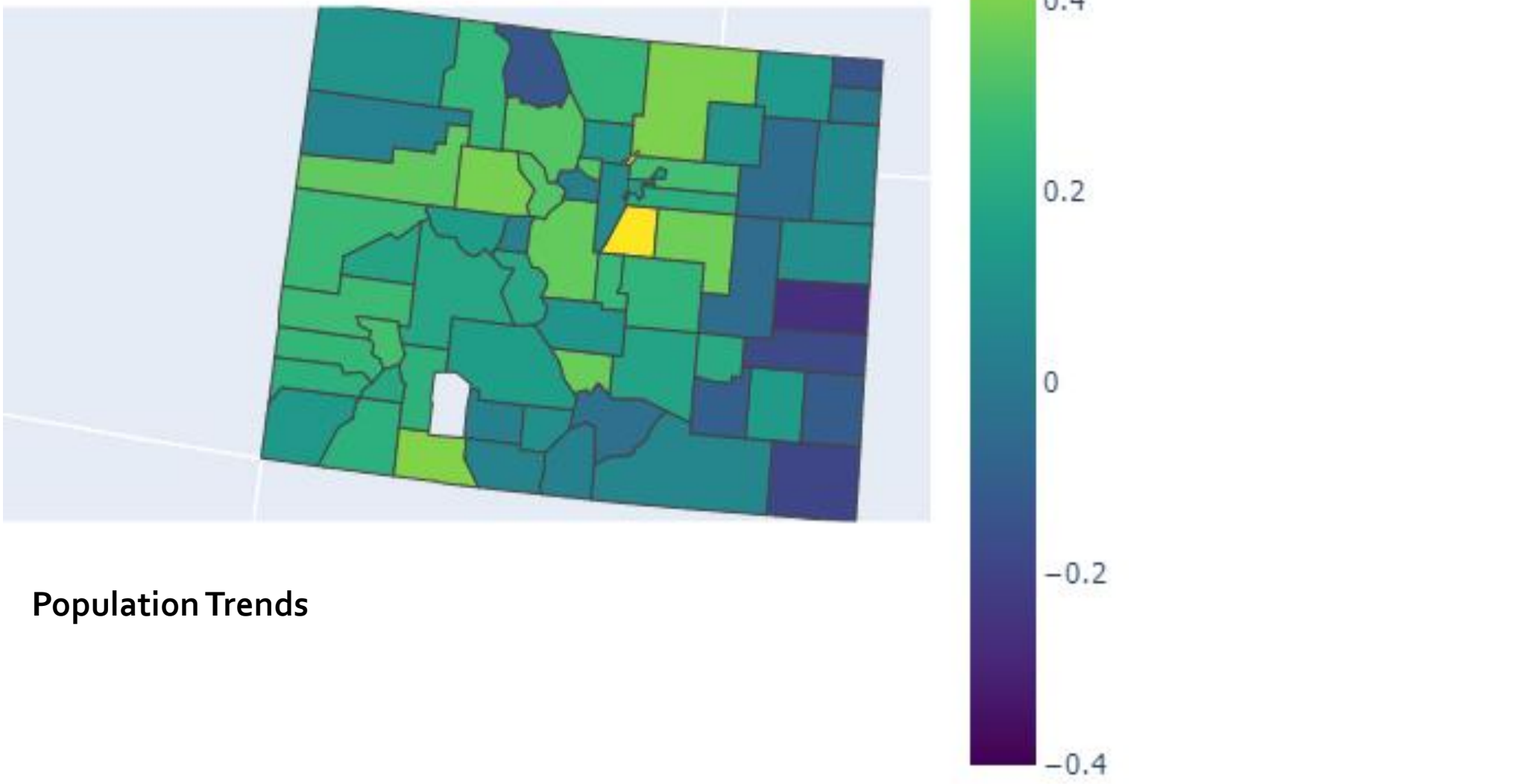


Differences between counties



Median Income (USD)

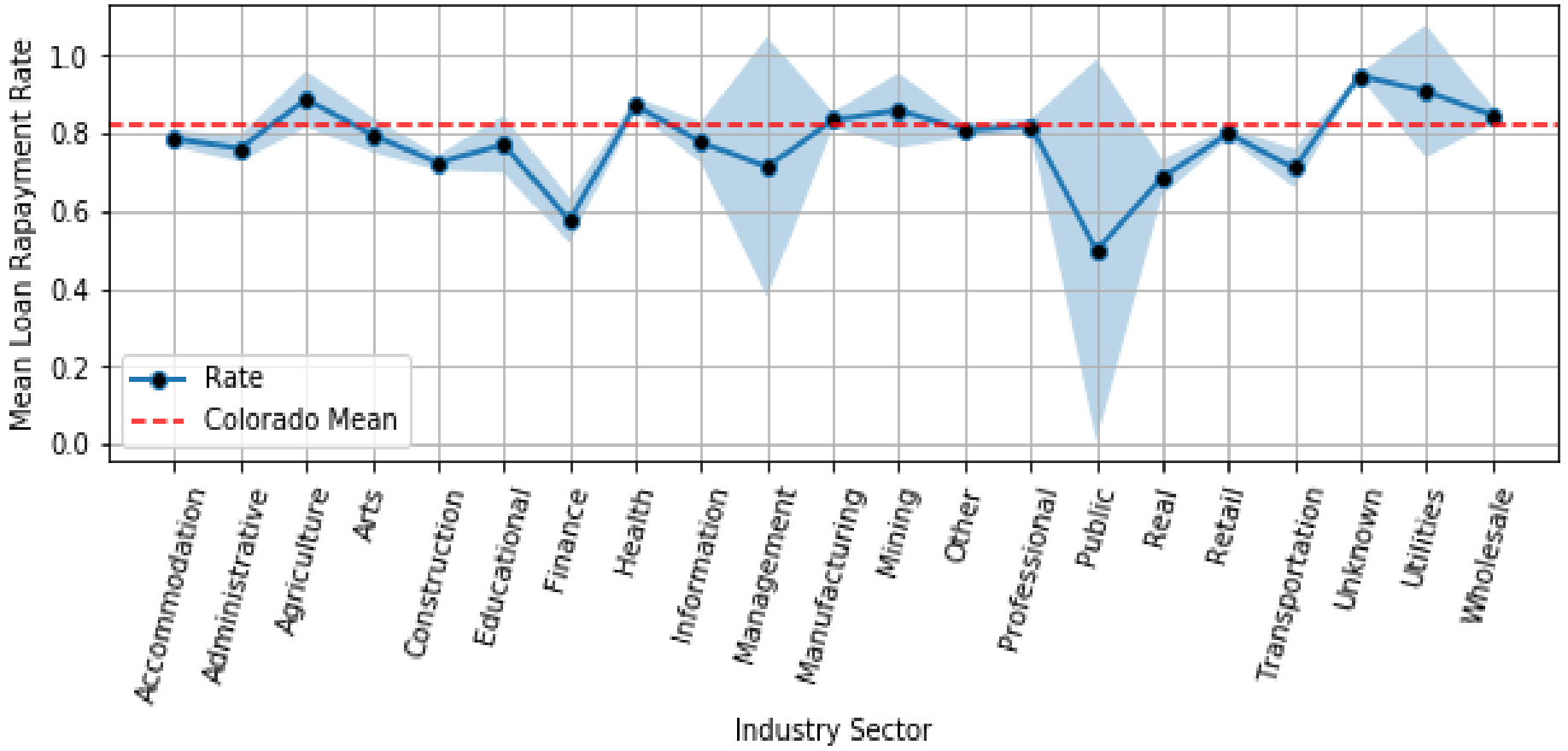
Differences between counties



Population Trends

NAICS (North American Industry Classification System)

- Codes simplified into 21 Broad Categories
- One-hot encoded into dataset

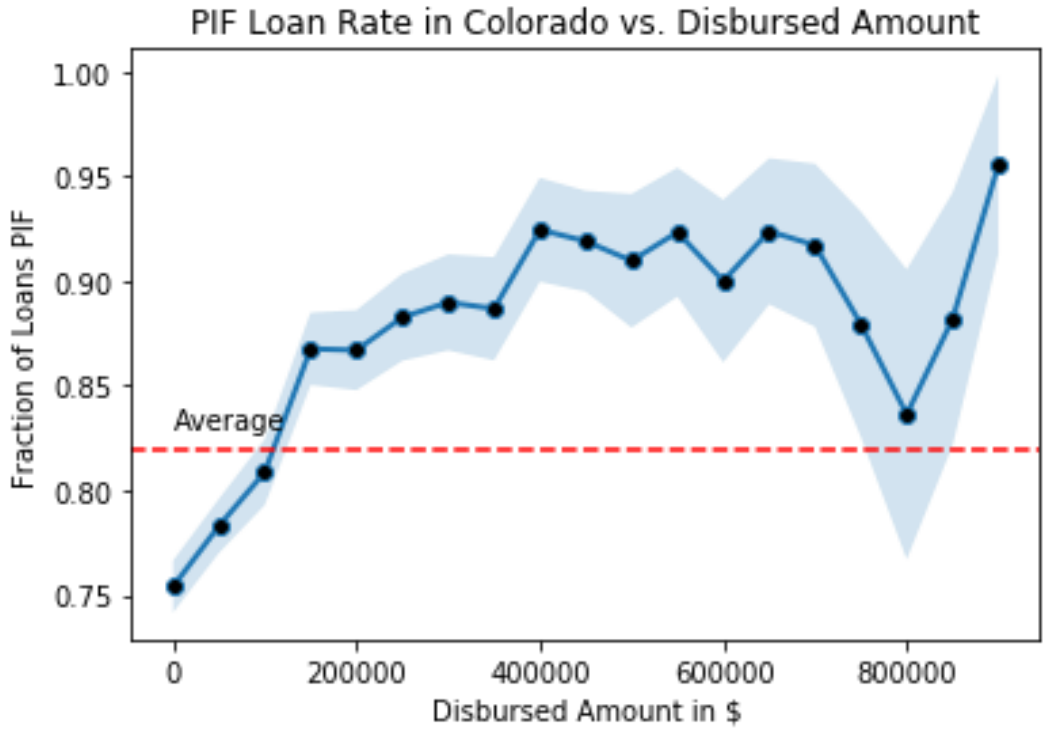


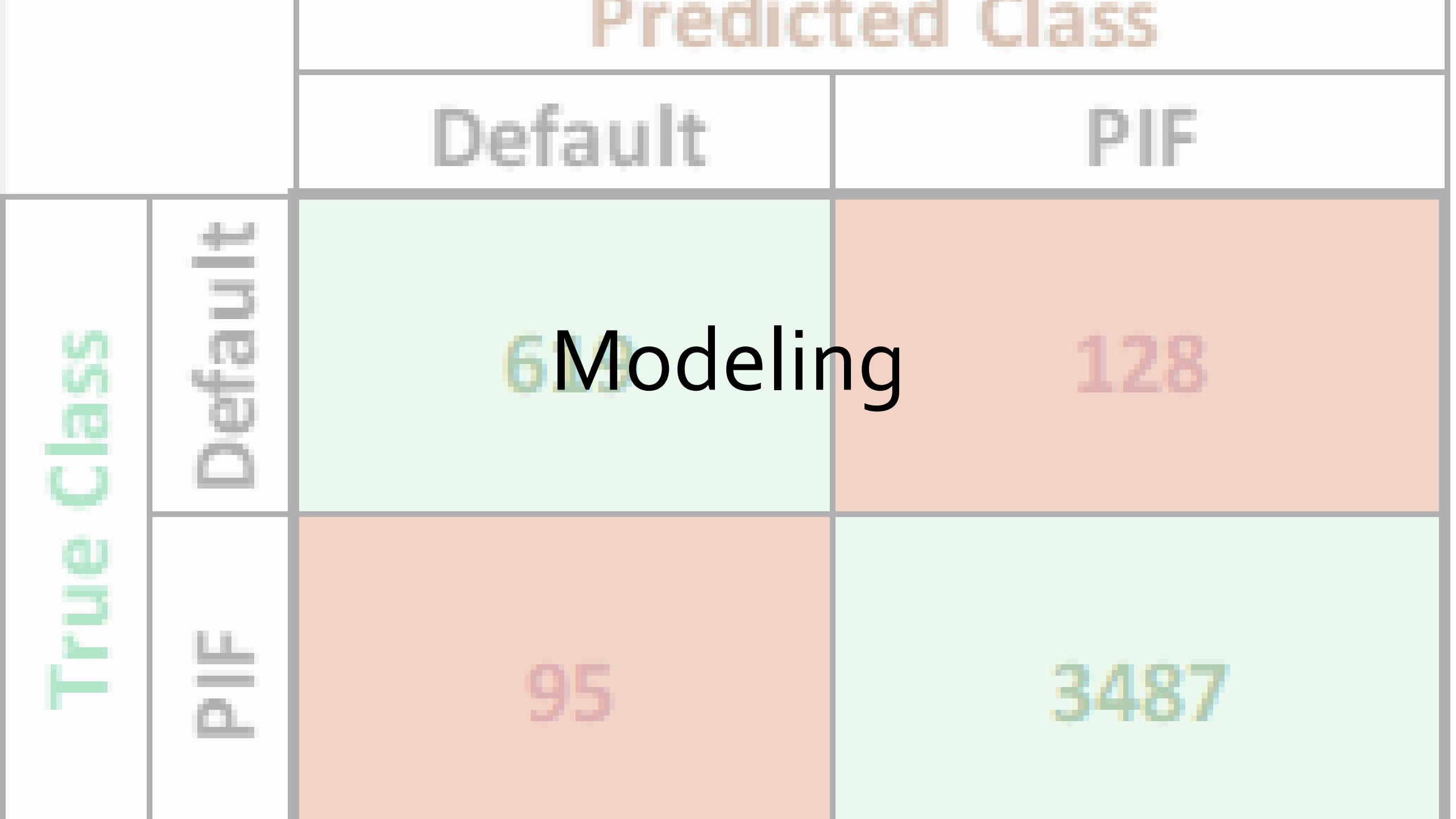
What can other dataset features tell us about default?

- Some binary features are significant
  - Low Documentation
  - Revisable Line of Credit

Feature	0 (PIF%)	1 (PIF%)	Z-Stat
NewExist (0 = Exist, 1 = New)	82.8%	81.7%	-1.64
RevLineCr (0 = No, 1 = Yes)	87.2%	67.6%	-30.21
LowDoc (0 = No, 1 = Yes)	81.5%	90.7%	11.04
FranchiseCode (0 = No, 1 = Yes)	82.5%	83.0%	0.47

- Loan Amount influences default rate

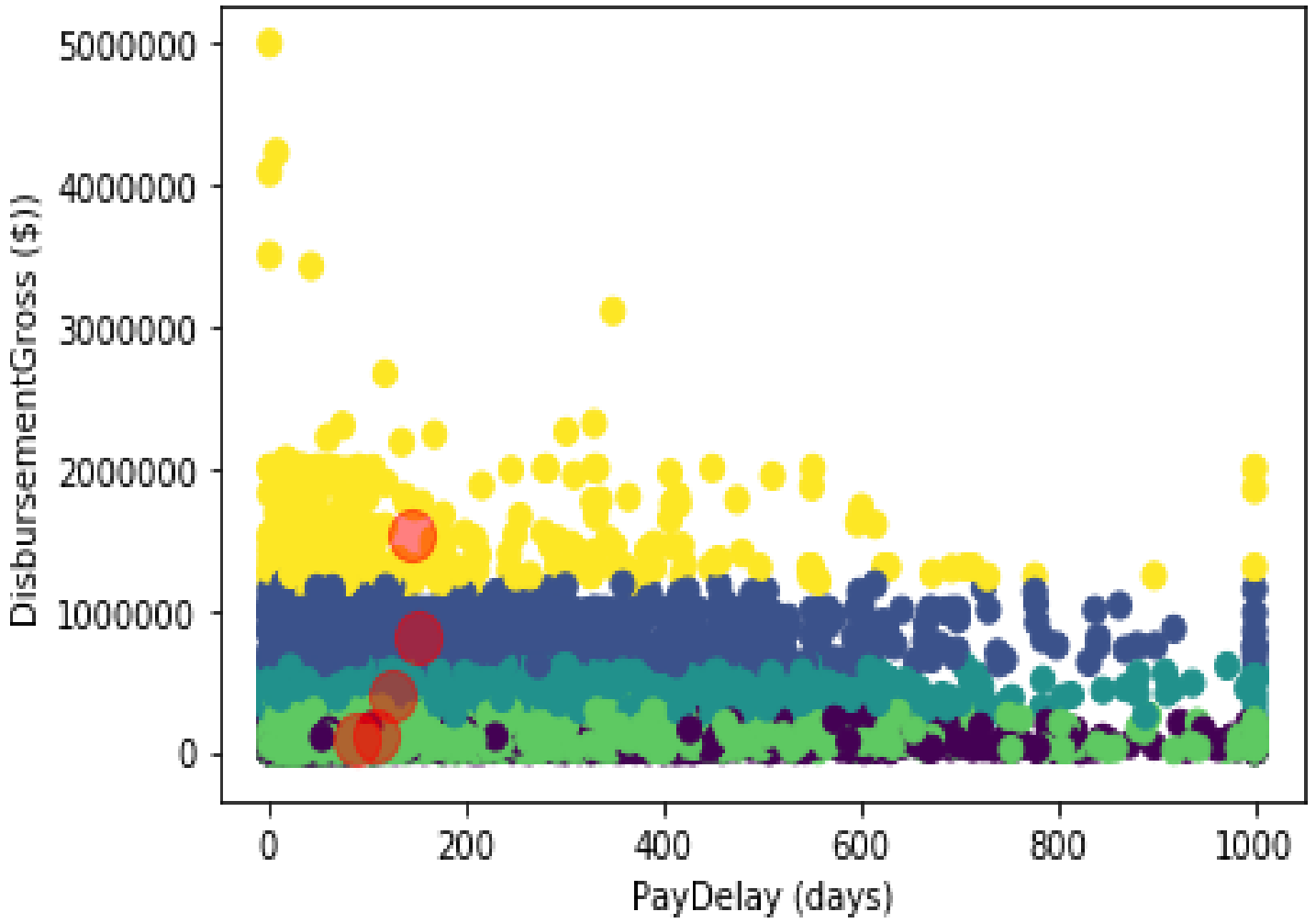



Predicted Class
Default
PIF
True Class
Default
PIF
610
128
95
3487
Modeling

Unsupervised Learning

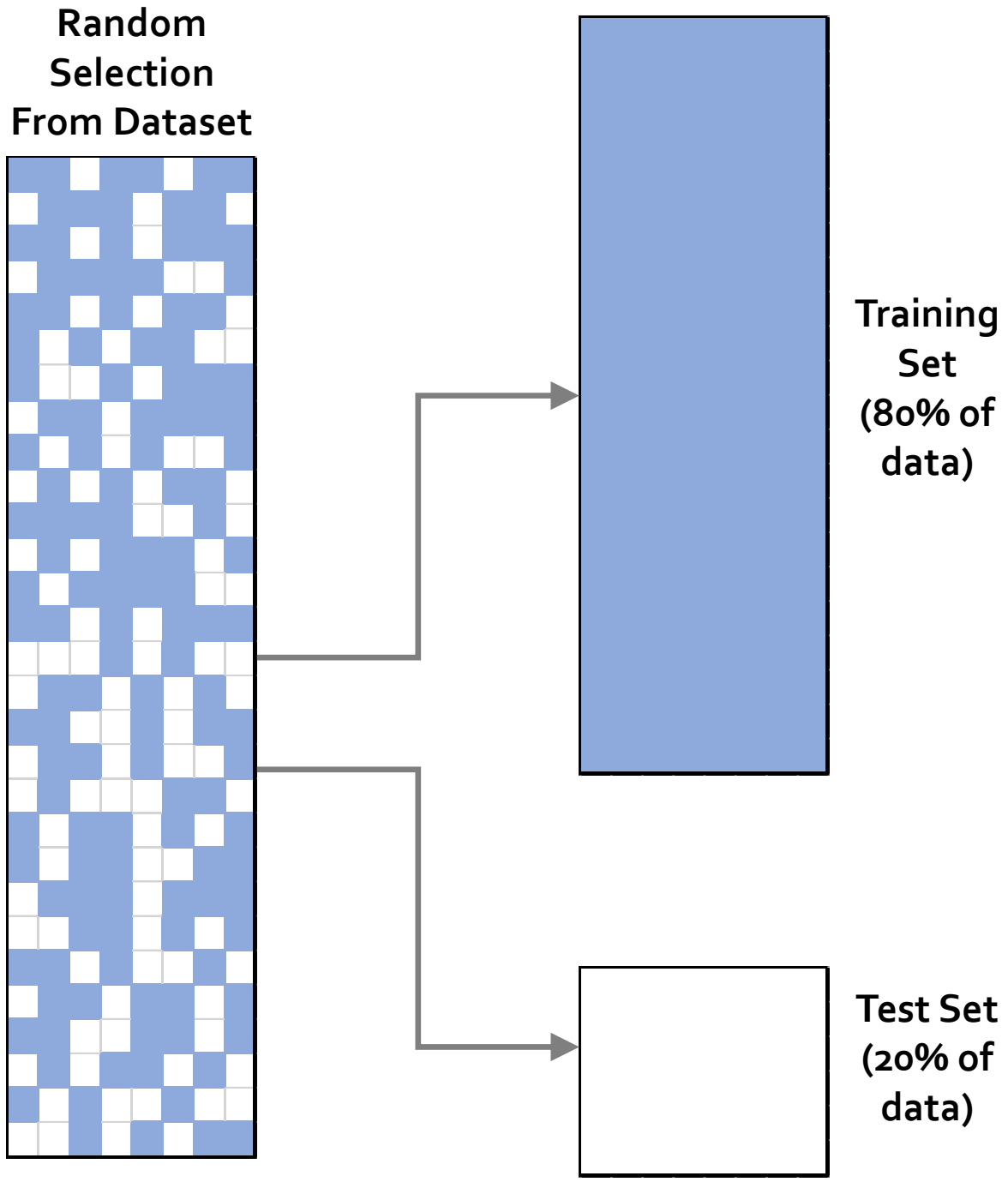
- Before training classification models
- K-Means clustering on data
- 5 clusters

(Red dots: Centroids of each cluster)



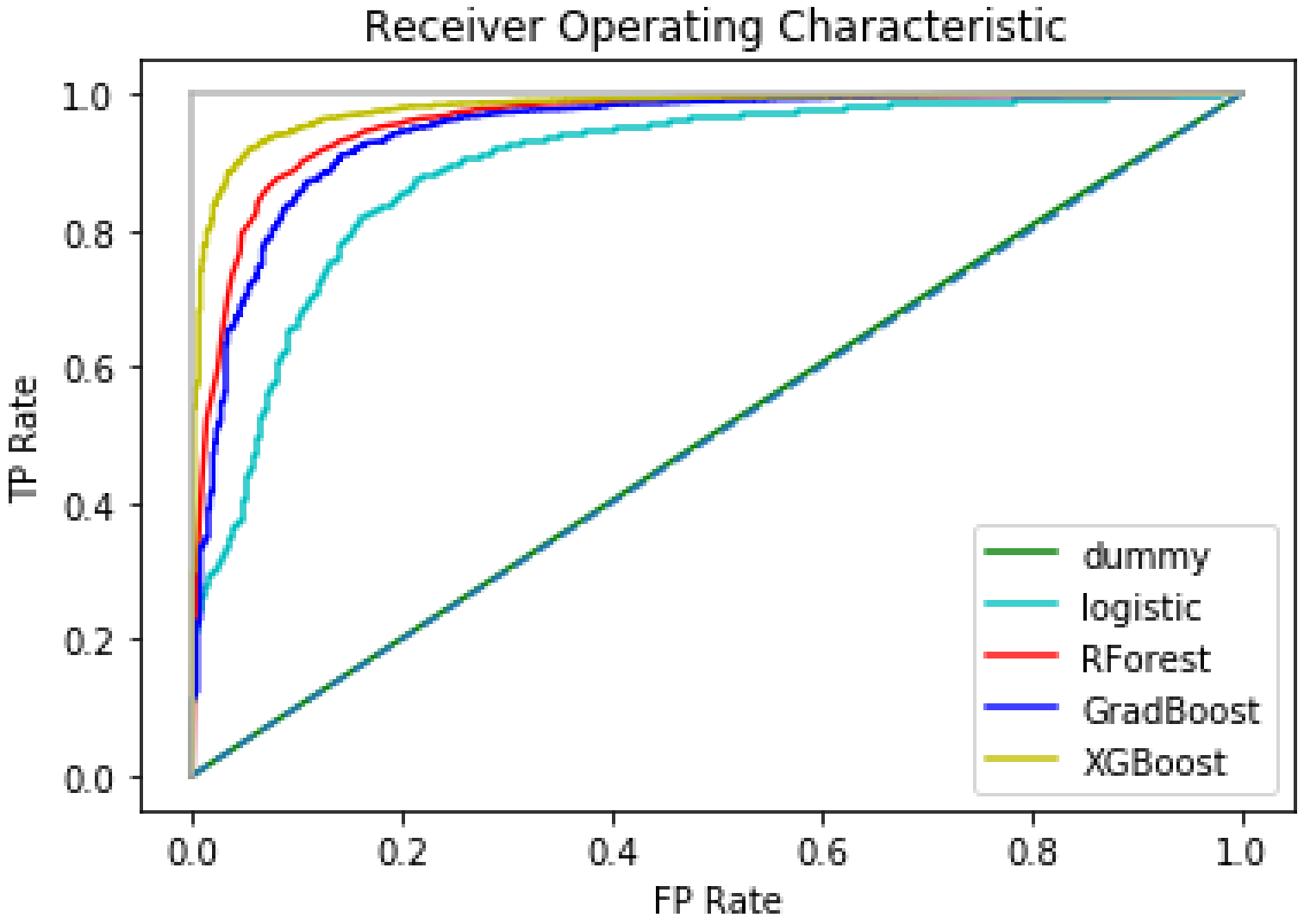
Classifier Models

- 20,000 rows, 40 columns in dataset
- Train/test split
- Try several different models on same split
- Tune & optimize hyperparameters



ROC curves

- Best AUC: XGBoost model at 89.8%



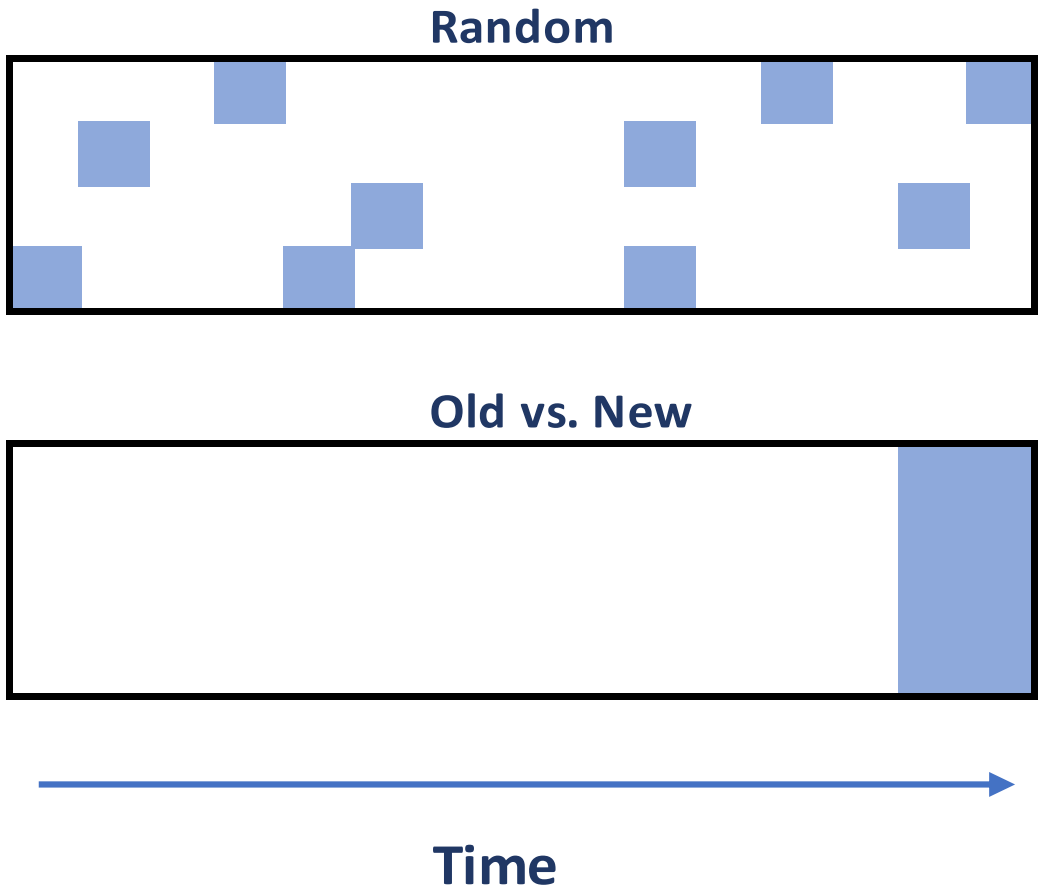
Models agree, disagree on feature importance

Top Ten most important features for each model

Rank	Random Forest	Gradient Boost	XG Boost
1	Loan Term	Loan Term	Loan Term
2	Approval FY	Delay in Payment	Approval FY
3	Delay in Payment	Approval FY	Revisable Line of Credit
4	Loan Size	Revisable Line of Credit	Industry - Wholesale
5	Median Income	Loan Size	Industry - Health
6	No. Employees	Industry - Unknown	Delay in Payment
7	Employed in County	Median Income	K-Means Cluster #2
8	Unemployment % in County	Unemployed in County	Industry - Finance
9	Unemployed in County	Employed in County	Population Change in County
10	Revisable Line of Credit	PIF Rate in County	Low Documentation

What's the best way to train?

- Leakage of feature information into test set
- Artificially boosts predictive power of model
- Reality: train on past, test on future



What's the best way compare models?

- False Positive: Predicting a loan will be paid, when it won't.
- False Negative: Predicting a loan won't be paid, when it will.

<u>Both are bad!</u>

- Matthews Correlation Coefficient penalizes a model for both:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Model Performance Metrics

	True Negative	False positive	False Negative	True Positive	Precision	Recall	Matthews Correlation Coefficient	Predicted Default Rate if Used
Dummy Reg. (Stratified	129	618	612	2970	0.828	0.829	0.143	14.3%
Logistic Regression	355	392	109	3473	0.899	0.970	0.563	9.1%
Random Forest	524	223	73	3509	0.940	0.980	0.753	5.2%
Gradient Boost	517	230	110	3472	0.938	0.969	0.720	5.3%
XGBoost	619	128	95	3487	0.965	0.973	0.821	3.0%
XGBoost old v new	66	15	37	765	0.981	0.954	0.698	1.7%

- Train/test split based on time degrades performance of XGBoost from 0.82 to 0.70 (based on MCC)

Conclusions

- Features with most predictive power come from SBA dataset
- But, all models use Colorado-specific data to improve performance

- XG Boost model could **reduce default rate to 15% to 2%**, even when we are predicting the future based on the present.

Thank You!

more info at [github repository]

Springboard