PLODI

Pandemic Loan Outlier Detection and Indicators

MIDS W210 Capstone Presentation
12/14/2023

About Us

Sridhar Chadalavada

EXPERIENCE

10+ years in Healthcare

FOCUS

Healthcare technology, & management

Crystal Chen

EXPERIENCE

8 years in Finance/Accounting

FOCUS

Finance & Data Analytics

Roberto Saldivar

EXPERIENCE

3 years in Manufacturing

FOCUS

Energy and Sustainability, Catalyst Manufacturing

Mike Varner

EXPERIENCE

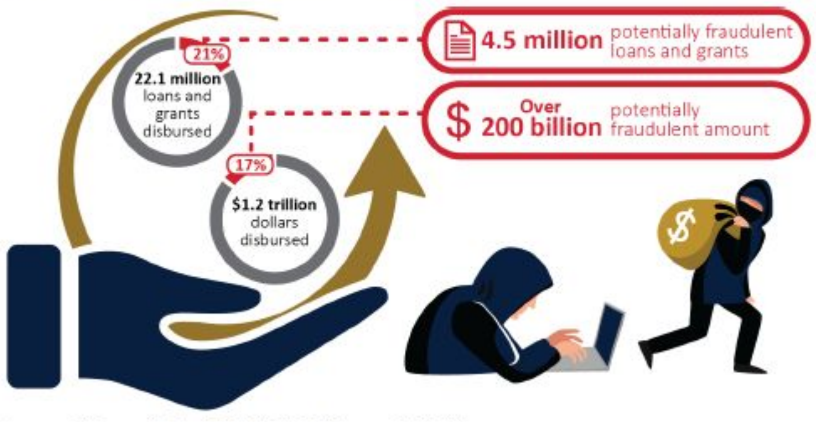
6 years in Finance/Consulting

FOCUS

Data Science & Analytics

The Problem

Figure 1: Potential fraud in SBA's pandemic loan programs



Source: [COVID-19 Pandemic EIDL and PPP Loan Fraud Landscape]

To support businesses during the COVID-19 pandemic, the US Small Business Administration disbursed **$1.2T** of **loans**.

Due to the rapid pace at which loans were processed, estimates of **fraudulent loans** range from **$100B-$200B**.

Our Goal

Mission

Identify PPP loan features for fraud risk using machine learning models and develop an open web dashboard ranking loans for investigation based on our risk assessment analysis.

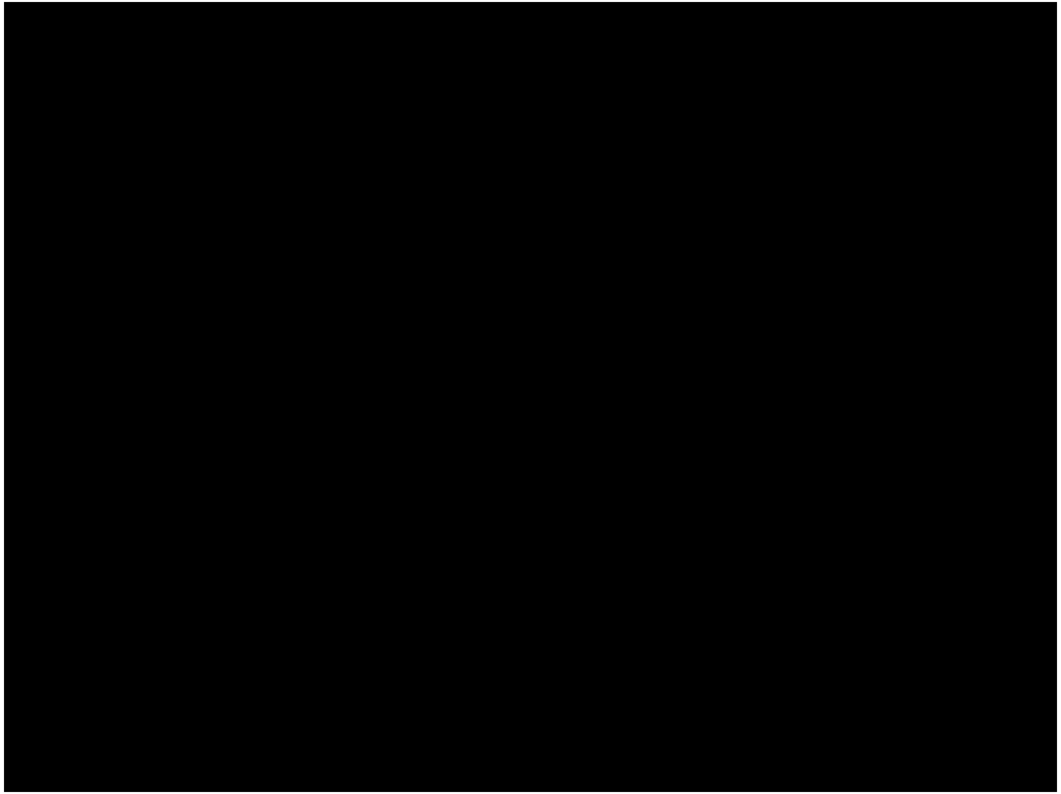
Impacts

Identification of fraud indicators and investigation priority for regulatory agencies

Guidance for loan screening for future loan programs

Open access to government spending for journalists and interested public individuals

Minimum Viable Product Demo

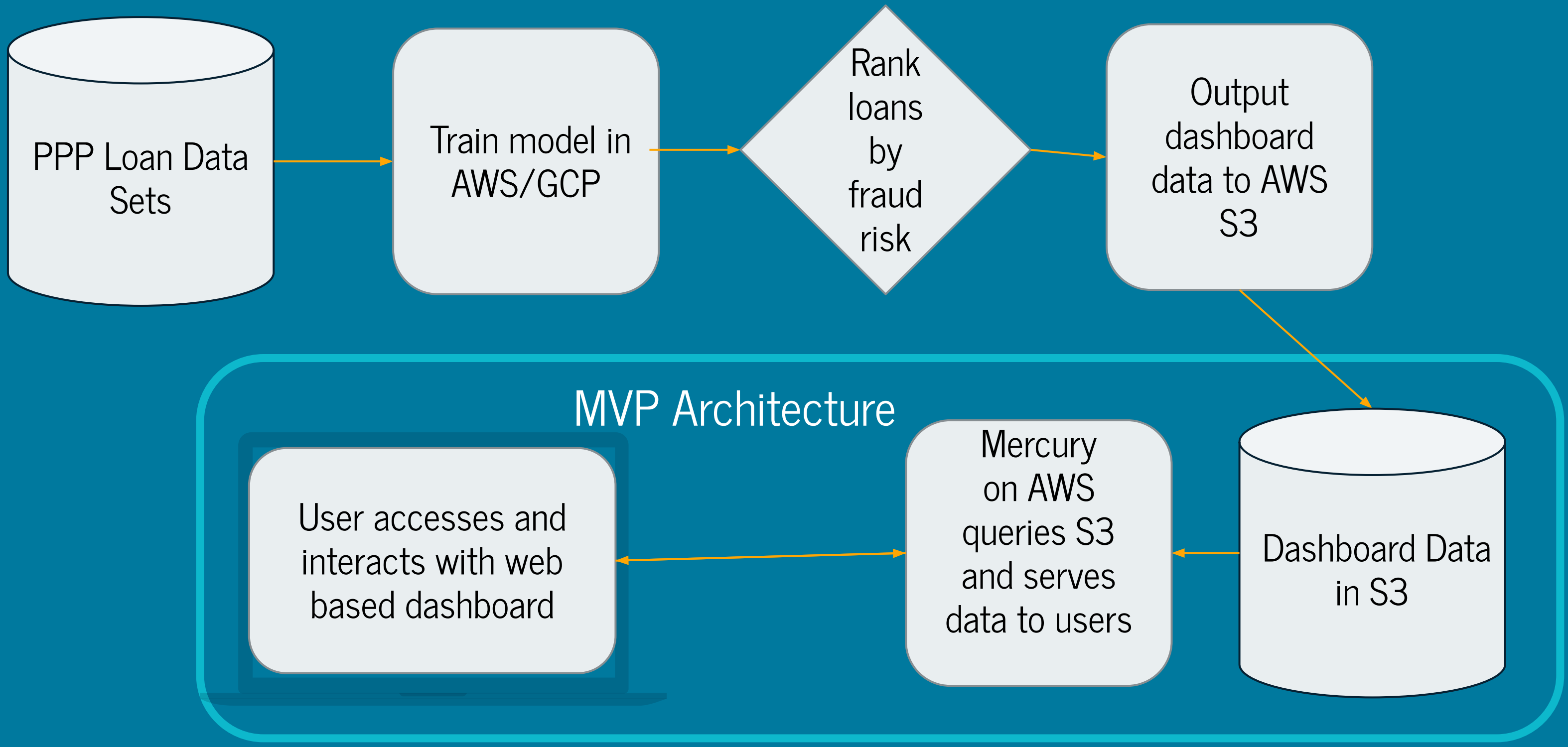


[MVP Link Here]

[Project Description Here]

Data and Modelling Pipeline

Data analysis / dashboard is user accessible



Data Used

Primary Loan Data
PPP Loan Data from SBA
9.1M individual approved loans with names, address, loan amount, term, jobs reported.

Supplemental Data

USPS Address Verification
Validate applicant's address under the theory invalid addresses are suspicions **(feature engineering)**

NAICs Codes & CBSA Data
Census data by region and industry to determine normalized implied pay **(feature engineering)**

Case Data
We reviewed and labelled 108 adjudicated DOJ cases. 108 cases. 614 individual and company names yielding 752 unique loans. Assume that a loan is "suspect" if it's associated with one of the cases **(data labelling)**

Final Modelling Data
30 features and 1 label

Model Summary

Assumptions & Methodology

We assume a true fraud rate of 8% as estimates range from $70B-$200B[1] of the $1.2T disbursed

All models are trained and tested on 9.4k loans, via downsampling of the non-case related loans, and assumed to be non-suspect.

Train:Test split of 80%:20% respectively.

Model Evaluation

Prosecuted cases are positive loan labels but remaining loans are unknown status.

Weigh Recall (Sensitivity) and Negative Predictive Value as primary measures for MVP model selection.

1. COVID-19 Pandemic EIDL and PPP Loan Fraud Landscape and Griffin et al. Did FinTech Lenders Facilitate PPP Fraud? (August 15, 2022).

Model Summary

Models Used

Model	Notes	Model	Notes
Baseline	Assume no fraud given 8% true rate	MLP (Neural Network)	3 layers with 100 nodes each
Logistic Regression	No Regularization	K-Nearest Neighbors	75 neighbors
XGBoost - tree based ensembling approach	Decision tree based ensembling	TabNet	Deep Neural Network framework with encoder/decoder architecture
Co-Training*	Iterative ensembling approach with majority voting mechanism	XGBOD*	Ensemble of KNN, K-Median, AvgKNN, LOF, LoOP, One-Class SVM, Isolation Forests for unsupervised scoring & XGBoost for supervised classification from scores & original features.

* Semi-Supervised learning models

1. COVID-19 Pandemic EIDL and PPP Loan Fraud Landscape and Griffin et al. Did FinTech Lenders Facilitate PPP Fraud? (August 15, 2022).

Model Results

Key Metrics

Family	Model	Sensitivity (Recall)	Negative Predictive Value	Specificity	Positive Predictive Value (Precision)	F1	Accuracy
Guess	Assume All Are Not Suspect	0.00	92.02	100.00	0.00	0.00	92.02
Linear	Logistic Regression	14.40	93.08	99.86	90.00	24.83	93.04
Decision Tree	**XGBoost**	**40.80**	**95.06**	98.89	76.12	53.12	94.25
Neural Network	MLP (3x100)	29.60	93.83	92.85	26.43	27.92	87.80
	TabNet (DNN, Google)	22.40	93.67	99.58	82.35	35.22	93.42
Non-Parametric	KNN (N = 75)	20.80	93.56	99.79	89.66	33.77	93.49
Ensemble	XGBOD*	21.60	93.62	99.86	93.10	35.06	93.61
Ensemble	Co-Training*[1]	36.80	93.06	10.57	73.49	82.12	70.56

XGBoost outperformed all models across Recall and Negative Predictive Value and serves as our **Champion Model**

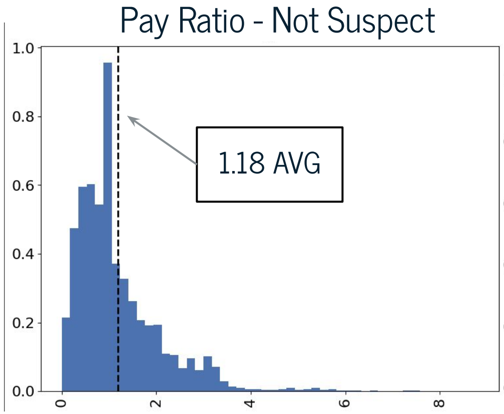
* Semi-Supervised learning models.

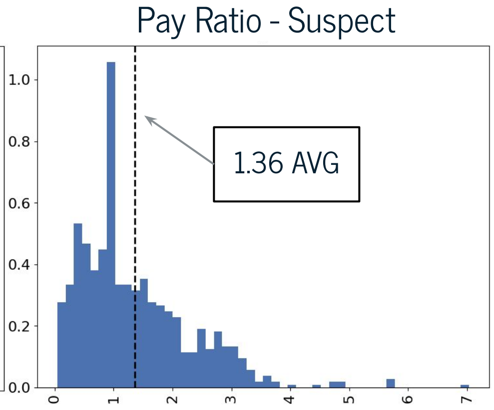
1. Co-Training results are included as representative performance, but model run on final test/train set used for other models is pending.

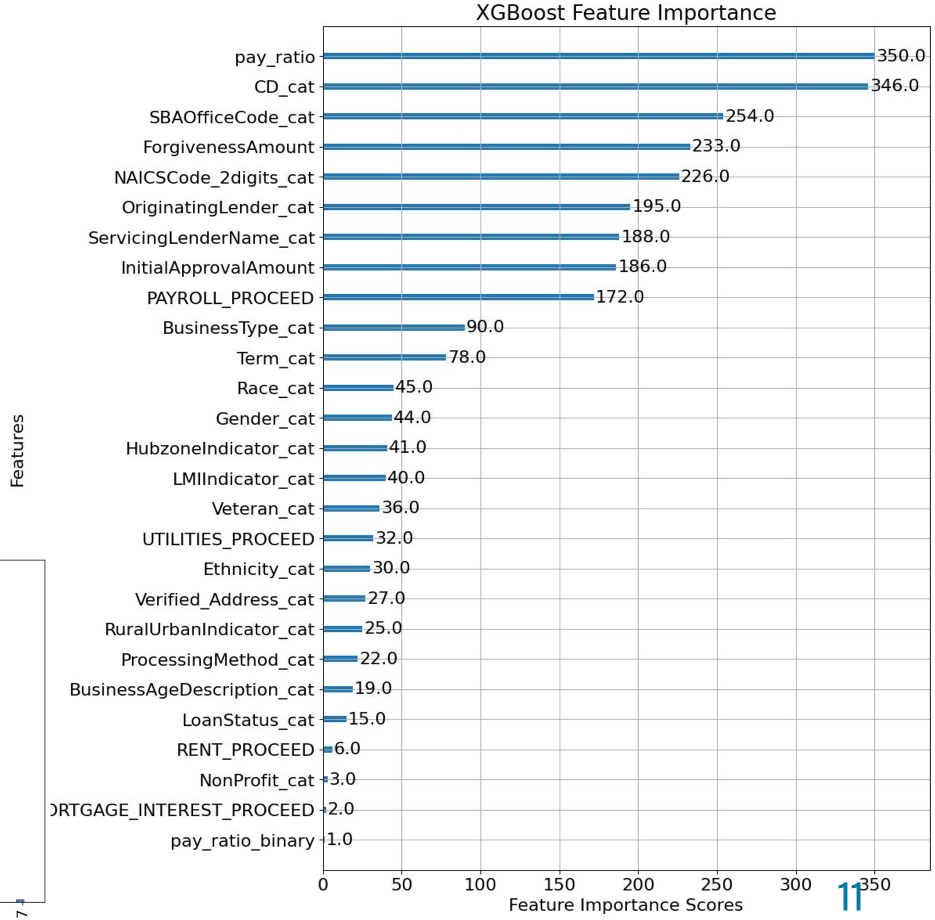
Champion XGBoost Model Results

Feature Importance

- Implied employee pay measures are the most important with the amount forgiven
- Geographic information, CD congressional district and SBA office code, are also highly important
- Suspicious loans tend to be clustered by geography, due to how cases were prosecuted, so this result may not generalize







Champion Model Results on Most Suspect Loans

Model performance dramatically improves, when looking at the most suspect loans

Given the scale of the PPP loan program and resourcing constraints, machine learning could guide expert review by providing a ranking.

XGBoost Performance On Most Suspect Loans (from test set)			
Top N	**Top N %**	**Sensitivity** (Recall)	**Negative Predictive Value**
10	0.5%	100.00	100.00
95	5.1%	100.00	100.00
100	5.3%	**100.00**	100.00
300	15.9%	**89.47**	97.42
500	26.6%	69.86	94.92
1000	53.2%	54.26	95.39

Ethical / Privacy Considerations

- Loan, case, and secondary data sources are publicly available and contain PII and other identifiers

  Mitigation:

  - **Privacy**: Removal of general PII including but not limited to loan ID, names, address, and company names from MVP.
  - **Defamation**: Qualify modeling results and resulting ranking
  - **Bias**: Removal of sensitive variables such as gender or race from MVP published data.

Future Opportunities

- Increase the size of our labelled loans to utilize more of the data
- Engage regulators in our ranking approach for our non-public data
- Transfer learning for other loan programs
- Privacy: Increase feature availability in MVP and reduce individual identification
- Model explainability and evaluating impact of labelled prosecutorial case discretion

Summary

Leverage public data for visibility into the PPP loan program to provide insights into disbursed loans, to help regulators prioritize investigations, and to improve future programs.

Thanks!

Any questions?

- Mike Varner mike_varner@ischool.berkeley.edu
- Crystal Chen crystalqianchen@ischool.berkeley.edu
- Roberto Saldivar roberto_saldivar@ischool.berkeley.edu
- Sridhar Chadalavada sridhar@ischool.berkeley.edu

Appendix

Acknowledgements

- We'd like to thank Daniel Aranki and Puya Vahabi, our course instructors, for their excellent guidance and feedback throughout the semester. We also want to thank Dakota Sky Potere-Ramos for working with us to identify and mitigate data privacy and ethics risks. Lastly, the authors of Did FinTech Lenders Facilitate PPP Fraud?John M. Griffin, Samuel Kruger, and Prateek Mahajan as many of our engineered features take inspiration from their work.

References

- Slide 1 logo: https://designs.ai/en/logomaker
- U.S. Small Business Administration OIG. (2023, June 17). Covid-19 pandemic EIDL and PPP loan fraud landscape. COVID-19 Pandemic EIDL and PPP Loan Fraud Landscape. https://www.sba.gov/sites/sbagov/files/2023-06/SBA%20OIG%20Report%2023-09.pdf

References (Models)

- PYOD library https://github.com/yzhao062/pyod
  - COPOD (Copula-Based Outlier Detection)
  - XGBOD (Extreme Boosting Based Outlier Detection)
- Scikit-learn https://github.com/scikit-learn/scikit-learn
  - Unsupervised Models (PCA, TSNE, IsolationForest)
  - KNN
  - Logistic Regression
  - MLP Neural Network
- Tabnet https://github.com/dreamquark-ai/tabnet
- Benchmark evaluation and nodel selection
  - Songqiao Han and Xiyang Hu and Hailiang Huang and Mingqi Jiang and Yue Zhao. ADBench: Anomaly Detection Benchmark. Neural Information Processing Systems (NeurIPS) (2022) https://github.com/Minqi824/ADBench

Presentation Template Attribution

- Presentation template created by SlidesCarnival
  - http://www.slidescarnival.com
- Template Photographs created by Unsplash
  - https://unsplash.com/
- Typographies:
  - Titles: Oswald
  - Body copy: News Cycle