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
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

PPP Loan Data Sets → Train model in AWS/GCP → Rank loans by fraud risk → Output dashboard data to AWS S3

MVP Architecture

User accesses and interacts with web based dashboard

Mercury on AWS queries S3 and serves data to users

Dashboard Data in S3
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
Assumptions & Methodology

We assume a true fraud rate of 8% as estimates range from $70B-$200B 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.

### Model Summary

#### Models Used

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

* 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

<table>
<thead>
<tr>
<th>Family</th>
<th>Model</th>
<th>Sensitivity (Recall)</th>
<th>Negative Predictive Value</th>
<th>Specificity</th>
<th>Positive Predictive Value (Precision)</th>
<th>F1</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guess</td>
<td>Assume All Are Not Suspect</td>
<td>0.00</td>
<td>92.02</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
<td>92.02</td>
</tr>
<tr>
<td>Linear</td>
<td>Logistic Regression</td>
<td>14.40</td>
<td>93.08</td>
<td>99.86</td>
<td>90.00</td>
<td>24.83</td>
<td>93.04</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>XGBoost</td>
<td>40.80</td>
<td>95.06</td>
<td>98.89</td>
<td>76.12</td>
<td>53.12</td>
<td>94.25</td>
</tr>
<tr>
<td>Neural Network</td>
<td>MLP (3x100)</td>
<td>29.60</td>
<td>93.83</td>
<td>92.85</td>
<td>26.43</td>
<td>27.92</td>
<td>87.80</td>
</tr>
<tr>
<td>Non-Parametric</td>
<td>KNN (N = 75)</td>
<td>20.80</td>
<td>93.56</td>
<td>99.79</td>
<td>89.66</td>
<td>33.77</td>
<td>93.49</td>
</tr>
<tr>
<td>Ensemble</td>
<td>XGBOD*</td>
<td>21.60</td>
<td>93.62</td>
<td>99.86</td>
<td>93.10</td>
<td>35.06</td>
<td>93.61</td>
</tr>
<tr>
<td>Ensemble</td>
<td>Co-Training*</td>
<td>36.80</td>
<td>93.06</td>
<td>10.57</td>
<td>73.49</td>
<td>82.12</td>
<td>70.56</td>
</tr>
</tbody>
</table>

**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.
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 XGBoost Model Results

Pay Ratio - Not Suspect
Pay Ratio - Suspect
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.

<table>
<thead>
<tr>
<th>Top N</th>
<th>Top N %</th>
<th>Sensitivity (Recall)</th>
<th>Negative Predictive Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.5%</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>95</td>
<td>5.1%</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>100</td>
<td>5.3%</td>
<td><strong>100.00</strong></td>
<td>100.00</td>
</tr>
<tr>
<td>300</td>
<td>15.9%</td>
<td>89.47</td>
<td>97.42</td>
</tr>
<tr>
<td>500</td>
<td>26.6%</td>
<td>69.86</td>
<td>94.92</td>
</tr>
<tr>
<td>1000</td>
<td>53.2%</td>
<td>54.26</td>
<td>95.39</td>
</tr>
</tbody>
</table>
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
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
References (Models)

- PYOD library [https://github.com/yzhao062/pyod](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](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](https://github.com/dreamquark-ai/tabnet)
- Benchmark evaluation and model selection
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