Claim Fraud Detection

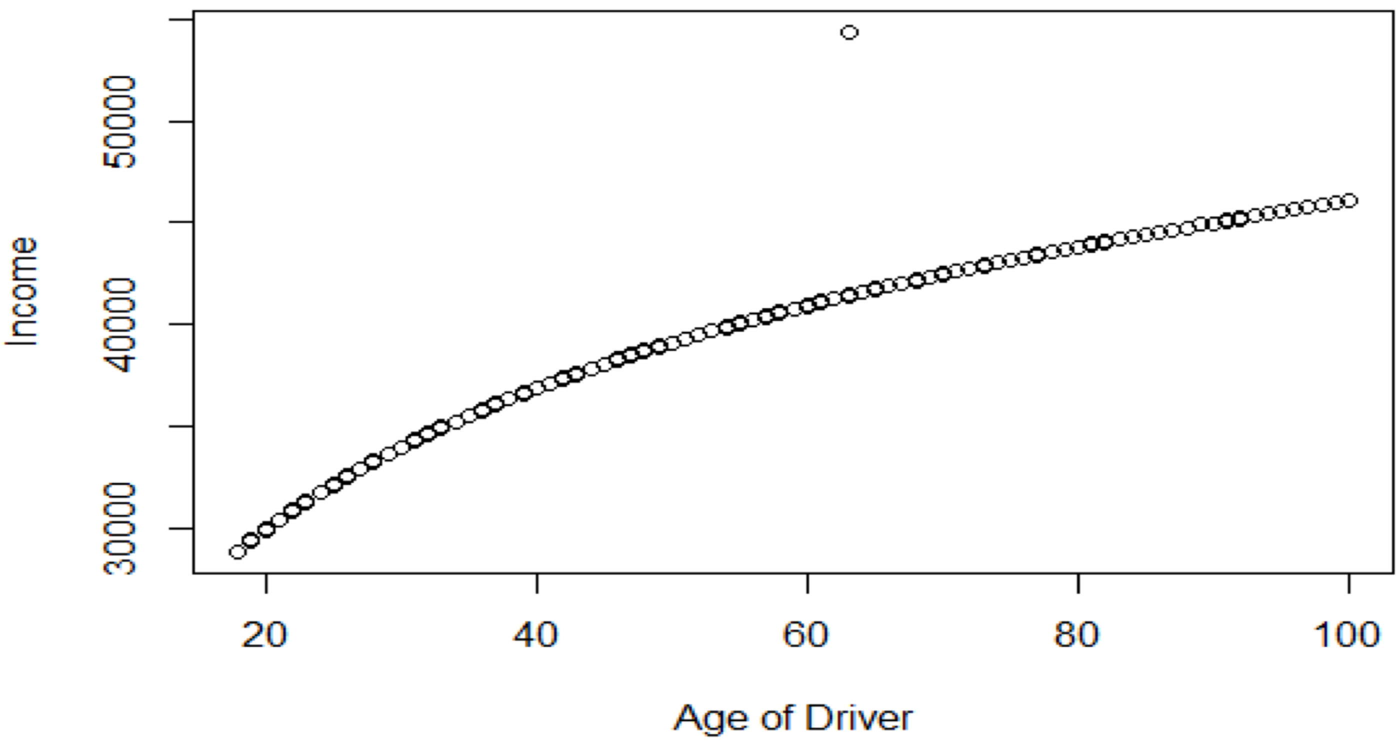
The Magnificent Six

Ruoyu He, Tate Jacobson, Yinuo Zeng, Yifan Yan, Yuchen Yao

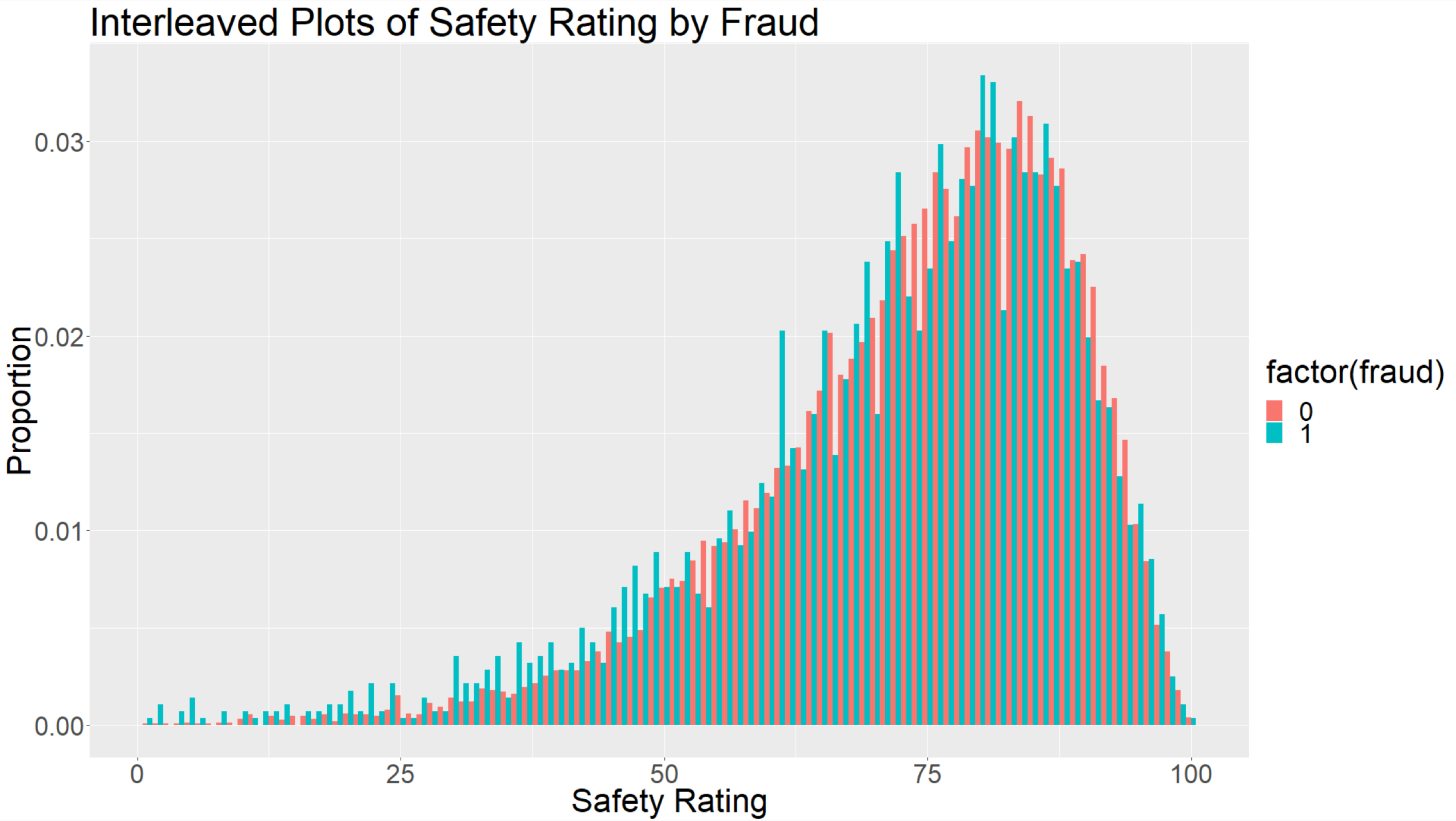
Outline

- **Introduction**
- **Preprocessing**
  - Data cleaning
  - Feature engineering
  - Feature selection
- **Prediction**
  - Stacking
  - Modeling
- **Conclusion**

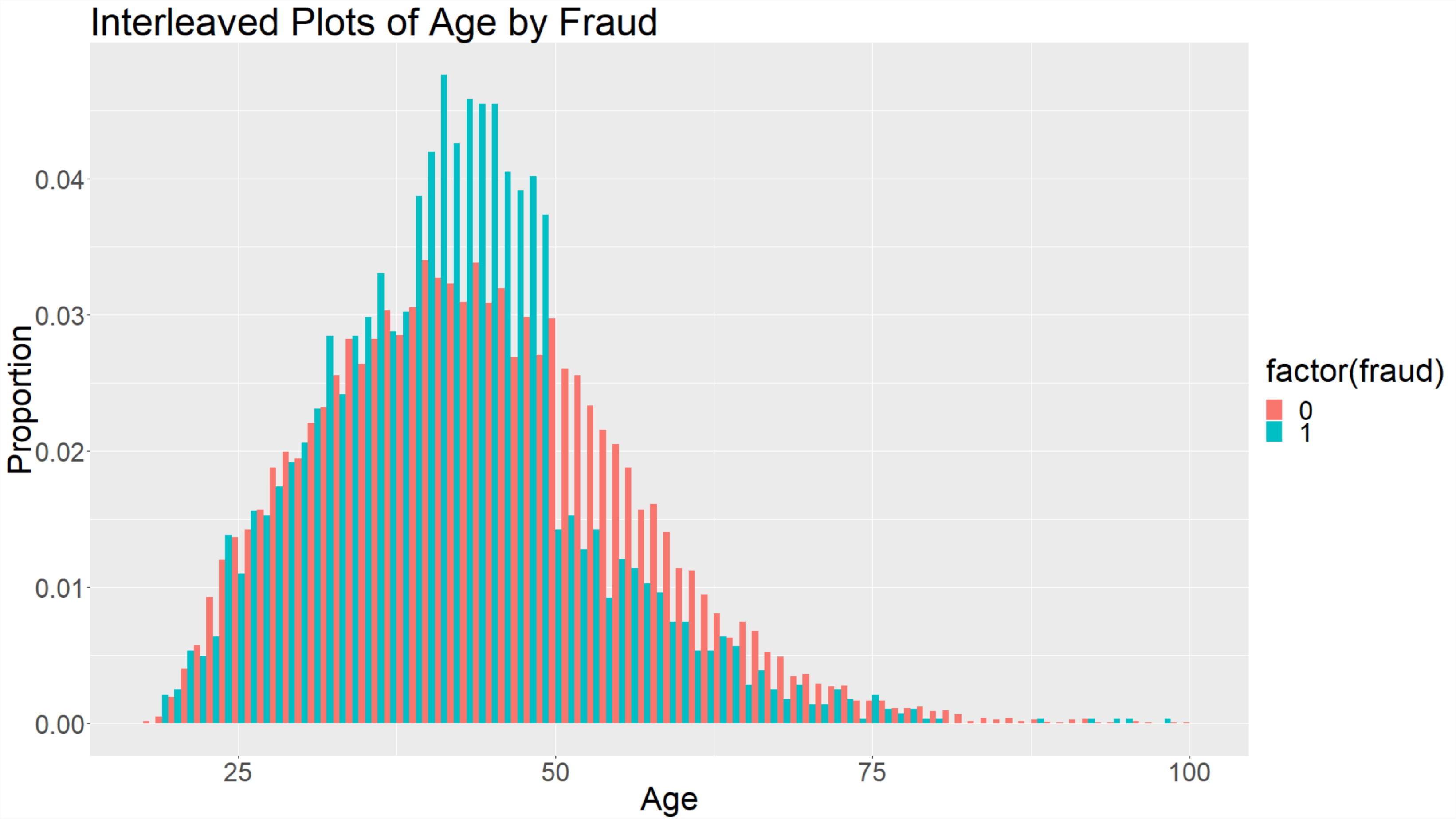
Introduction



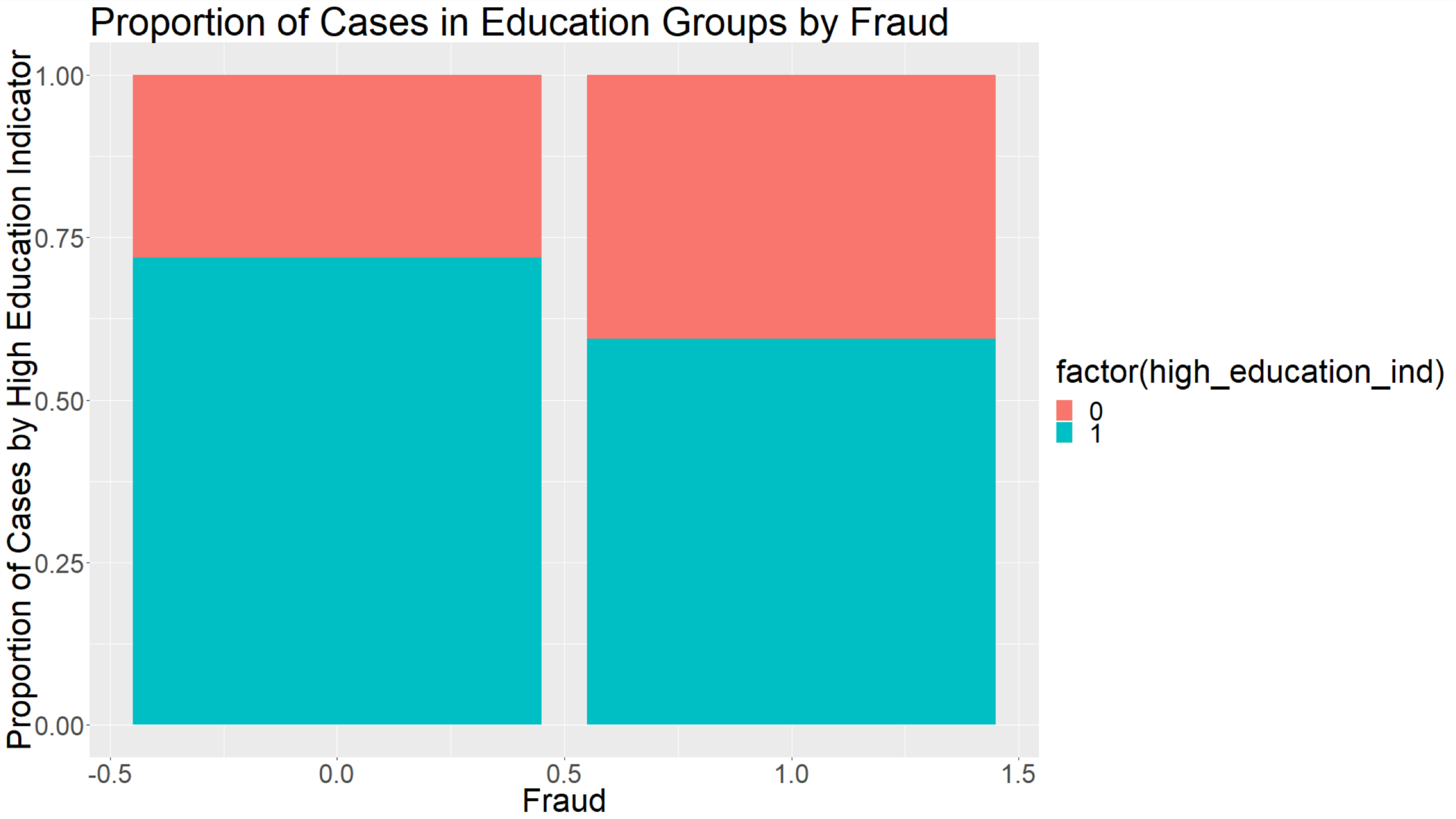
Highly correlated factors



Interleaved plots of safety rating by Fraud

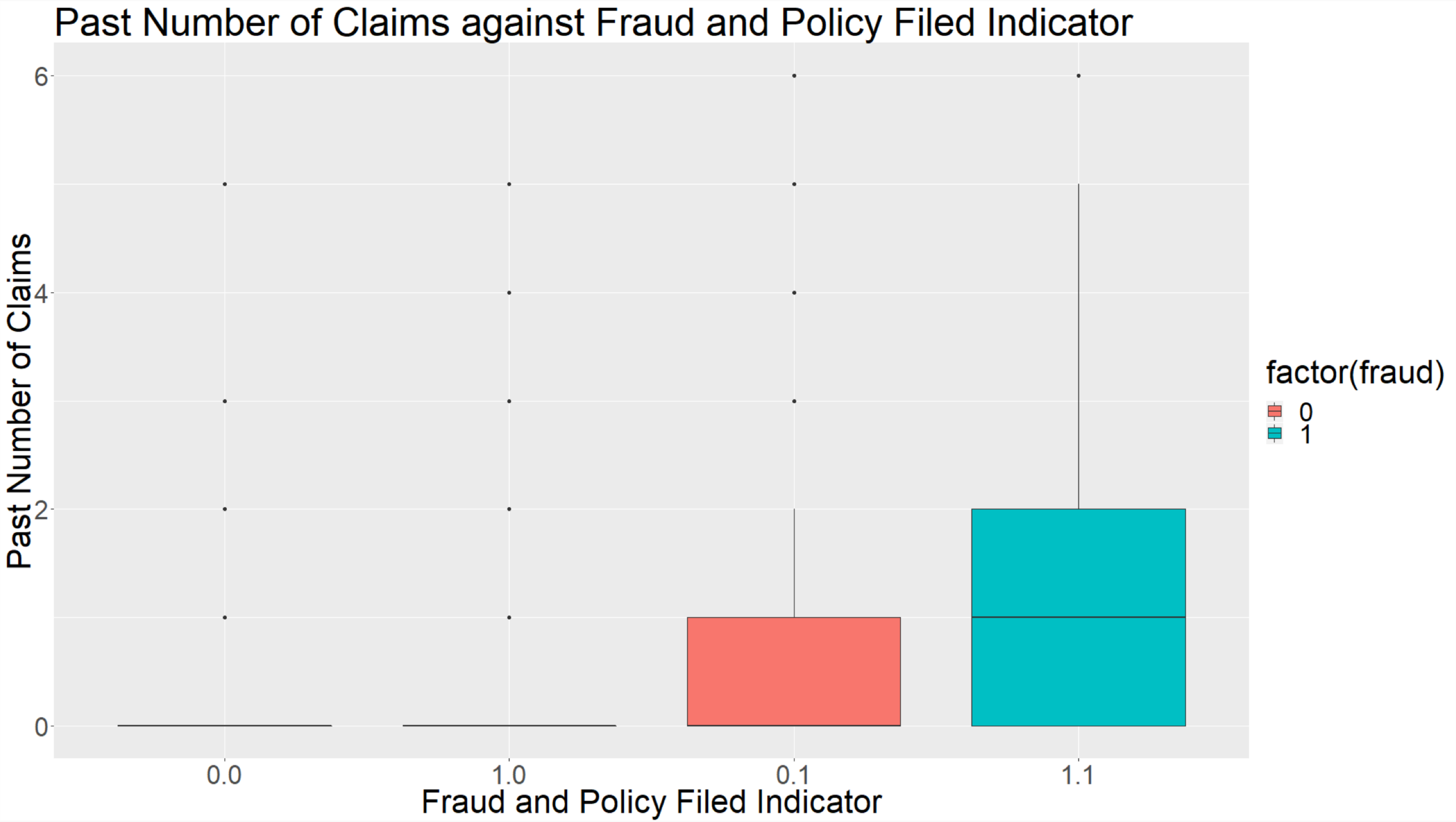


Interleaved plots of Age by Fraud



Proportion of cases in Education Groups by Fraud

Proportion of cases per Accident Site by Fraud



Interaction between Past Number of Claims and Policy Filed Indicator

Preprocessing

- Data Cleaning
- Feature Engineering
- Feature Selection

Preprocessing - Data Cleaning

- **NAs**
  - Train data (fraud=0): delete
  - Train data (fraud=1) & Test data: impute (Random Forest)
- **Special variables: Zip code**
  - NAs: impute (kNN)
  - Add city, state, longitude and latitude
- **Highly correlated variables**
  - Age and Income (0.98)

Preprocessing - Feature Engineering

- **Categorical variables**
  - Count
  - Supervised ratio

$$SR_i = \frac{P_i}{N_i + P_i}$$

  - WOE

$$WOE_i = \log\left(\frac{P_i/TP}{N_i/TN}\right)$$

- **Continuous variables**
  - Transform into categorical variables

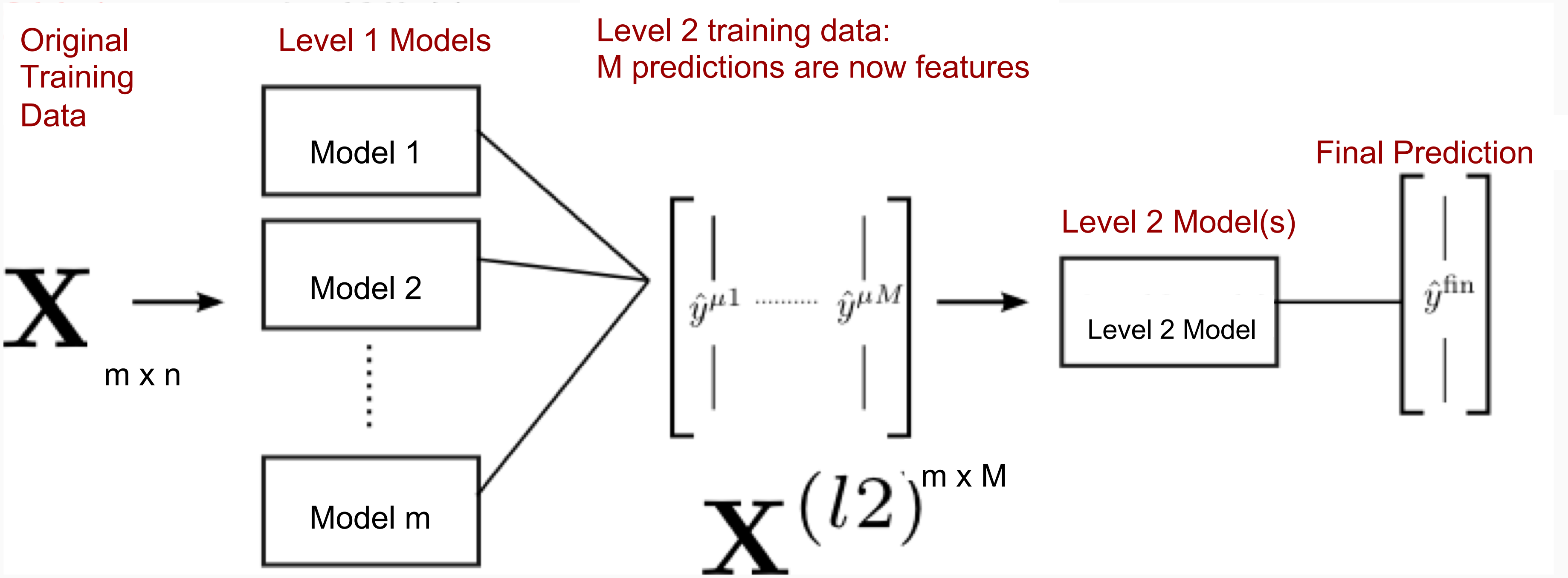
Preprocessing - Feature Selection

- **Variable Importance**
  - Random Forest
  - Chi-squared Test
  - Kruskal Test

Prediction

- Stacking
- Model

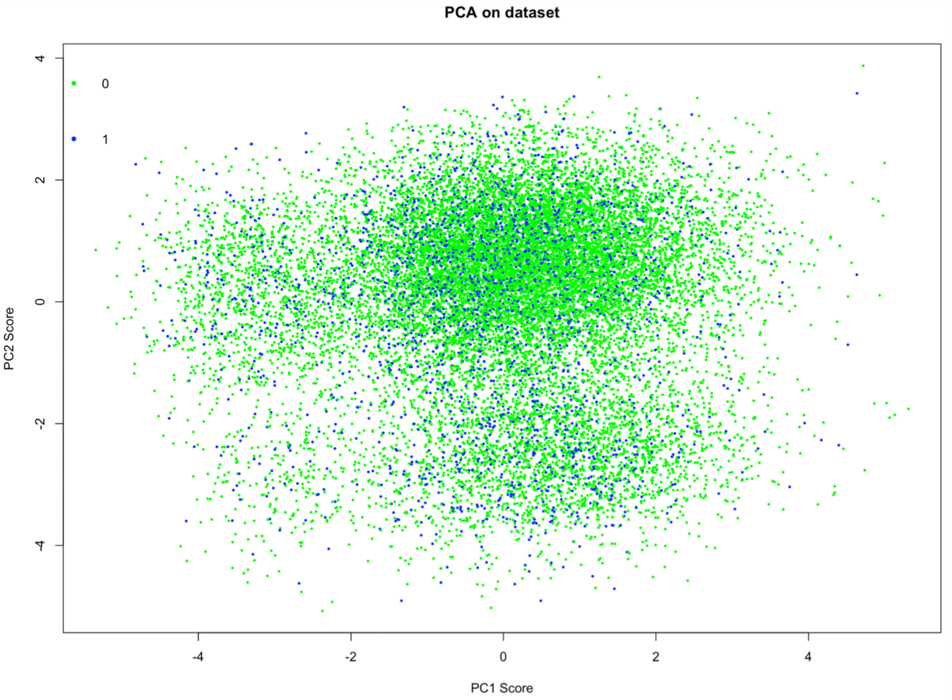
Prediction - Stacking (General Outline)

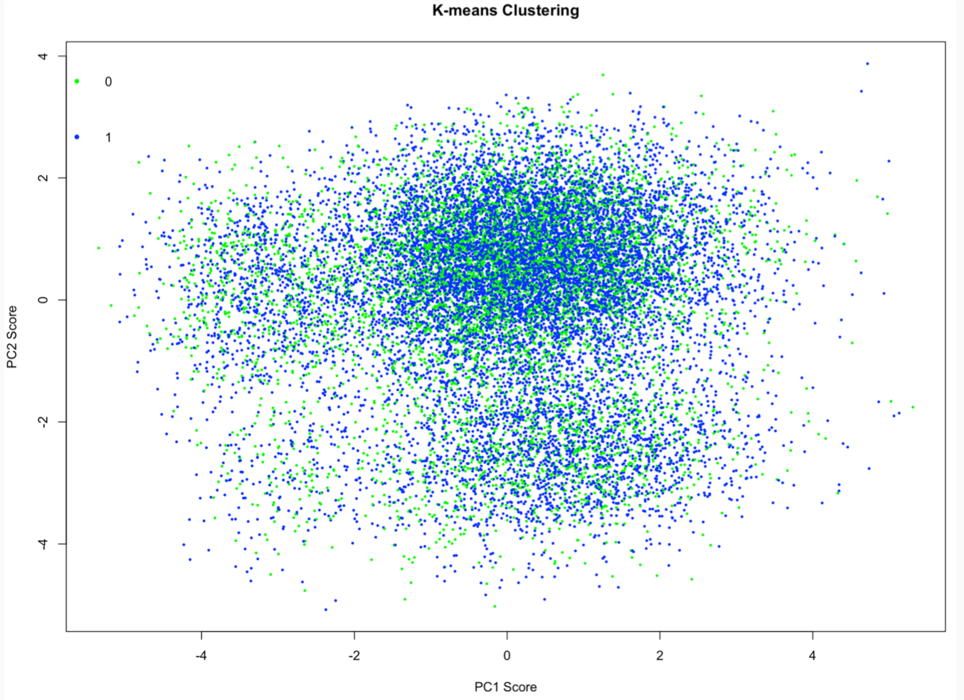


Prediction - Model Considered

Penalized Logistic Regression (LASSO, SCAD, MCP)
Discriminant Analysis (LDA, QDA, RDA)
Tree-Based Method (Random Forest, XGBoost)
K-Nearest-Neighbors (KNN)
Generalized Additive Model (GAM)
Clustering Methods
Neural Network (NN)

Visualization: Principal Component Analysis





Prediction - Model Considered

Prediction - Model Considered

Penalized Logistic Regression (LASSO, SCAD, MCP)
Discriminant Analysis (LDA, QDA, RDA)
Tree-Based Method (Random Forest, XGboost)
K-Nearest-Neighbors (KNN)
Generalized Additive Model (GAM)
Clustering Methods
Neural Network (NN)

Prediction - Modeling

We use Stacking to build a 3-layer architectures

First Layer
Second Layer
Third Layer

Prediction - Modeling

First Layer	Penalized Logistic Regression
	Naive Bayes Classifier
	Linear Discriminant Analysis (LDA)
Second Layer	
Third Layer	

0.680-0.700

Prediction - Modeling

First Layer	
Second Layer	Generalized Additive Model (GAM)
	Regularized Discriminant Analysis (RDA)
Third Layer	

0.720-0.730

Prediction - Modeling

First Layer

Second Layer

Third Layer

XGBoost

ANN

0.722-0.728

Final Prediction AUC

Our group's final prediction on the Kaggle Public Leaderboard had an AUC of 0.74781.

Conclusion

- Business Insights
- Takeaway
- Suggestions

Conclusion

Our findings from this project fall into two categories:

1. **What the model tells us**
2. **What the modeling process taught us**

Conclusion - Business Insights

- **What the model tells us:**
  - We found that the following variables are important:
    - Accident Site, Level of Education, Safety Rating, and Past Number of Claims
    - We used the XGBoost variable importance measure to find this
  - Fraud was more likely if the accident occurred in a parking lot
  - Drivers with lower safety ratings were more likely to commit fraud

Conclusion - Takeaway

- **What the modeling process taught us:**
  - Cast a wide net in researching different methods
    - Our research led us to gradient boosting, a key method in our model.
    - We added several methods to our "toolbox" for future data analysis, even if we did not use them for this project

Conclusion - Takeaway

- **What the modeling process taught us:**
  - Consider a variety of methods to test in cross validation
    - Every method can make predictions, so throw them into the mix for CV
    - Your preferred method may not work for every problem (e.g. SCAD and MCP performed poorly)

Conclusion - Takeaway

- **What the modeling process taught us:**
  - Prediction vs other goals
    - Machine Learning and nonparametric methods perform well, but do not give us easy-to-interpret models.
    - In other settings, we might focus more on variable importance and finding the true model.

Conclusion - Takeaway

- **What the modeling process taught us:**
  - Exploratory data analysis pointed us in the right direction.
    - Summary statistics often do not tell the whole story.
      - Remember Anscombe’s Quartet
    - Trust your intuition, but verify.
      - As we might expect, age and income are highly correlated.
      - Safety Rating is related to Past Number of Claims.

Conclusion - Suggestions

Variables which might be useful

- Driver's Criminal History
- Weather on the day of the accident
    - Bad weather might lead to more legitimate accidents
- Time of day of the accident
    - More legitimate accidents may happen during rush hour.
- Whether the accident occurred in the policyholder's home city
    - Perhaps a policyholder is more likely to commit fraud in a situation they can control

Thank you!

The Magnificent Six