# Travelers
# Fraud Prediction Challenge

Group 2

King Yiu Suen, Sam Piehl, Somyi Baek, Xun Xian, Yu Yang

# Outline of Approach

- Data Cleaning and Feature Engineering
  - Patterns of data
  - Significant features
  - New features
- Model Selection and Parameter Tuning
  - Best model
  - Parameter optimization
- Insights
  - Model interpretation

# Data Cleaning & Feature Engineering

- Deleting values that are unrealistic and imputing appropriately
- Transforming categorical variables
  - Binary encoding
  - One-hot encoding
- EDA
  - Scatterplots
  - Kernel Density curves
  - Correlation Matrix

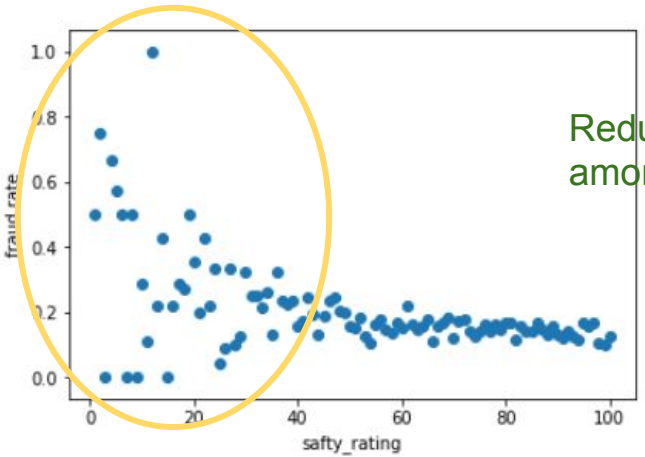
# Feature Engineering: Adding new features

- Three indices reflecting macroeconomic situation:
  1. Interest Rate 2. S&P 500 3. State Unemployment Rate
- Latitude and longitude
- Rating per Claim = safety rating / (past number of claims + 1)
- Count of missing values
- State

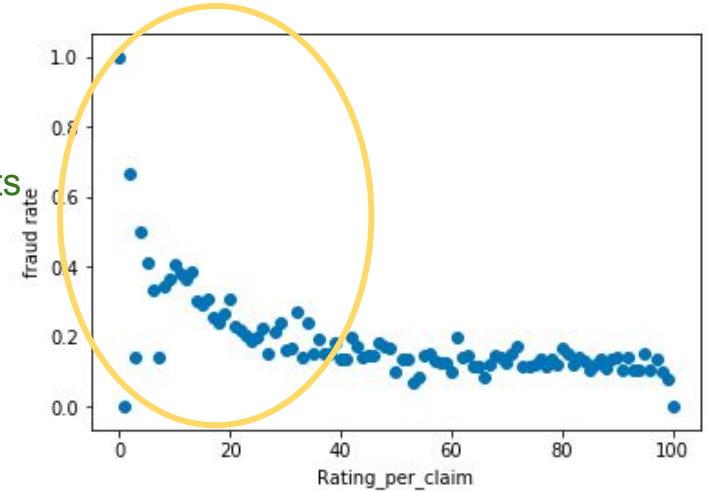
# Feature Engineering: Performance of new features

- Adding *latitude* and *longitude*, dropping *zip_code*
  - **0.001** boost in 5-fold CV AUC Score
- Adding *Rating Per Claim = safety_rating / (past_num_of_claims* + 1)
  - **0.0017** boost in 5-fold CV AUC Score
- Adding *interest rate*:
  - **0.0007** boost in 5-fold CV AUC Score
- Adding *count of missing values*:
  - **0.0004** boost in 5-fold CV AUC Score

# How we came up with the *Rating Per Claim*




Reduced variance among these points



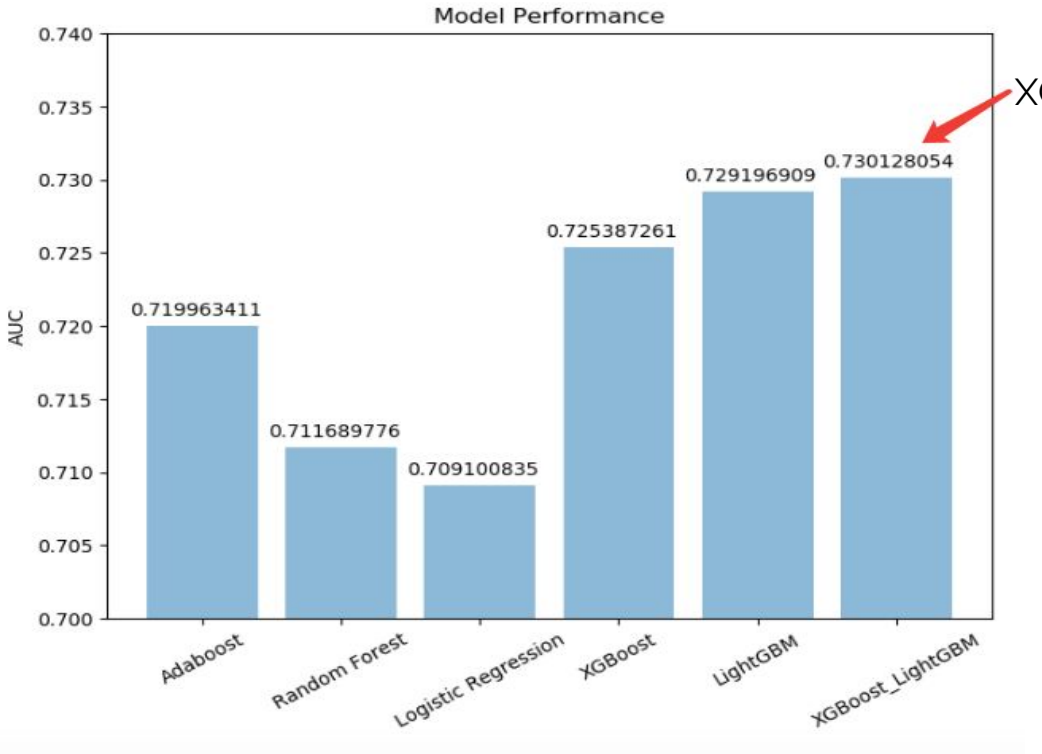

Original ‘Safety rating’

Safety rating / (no. of claims +1)

# Models and Implementations

- LightGBM, XGBoost, Adaboost, Random Forest, Logistic Regression
  - (LightGBM + XGBoost) > LightGBM > XGBoost > Adaboost > Random Forest > Logistic
- Parameter tuning:  Manual tuning, Bayesian tuning
  - Manual tuning > Bayesian tuning
- Cross-validation for verification
  - 5-fold CV to evaluate the performance

# Model Performance

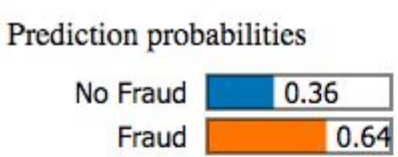


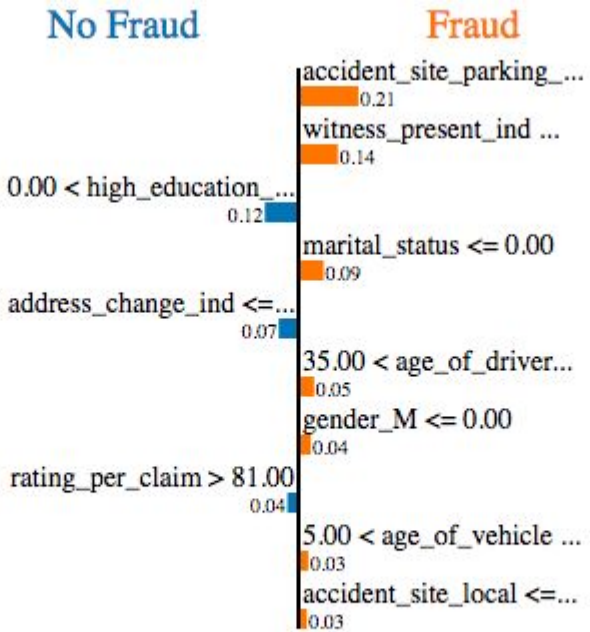

Winner!
XGBoost + LightGBM

# Final Model

- Chosen model:
  - 0.6 × LightGBM + 0.4 × XGBoost  (AUC = 0.74961 on Public Leaderboard)
- Existing features that we didn't use:
  - *claim_date*, *vehicle_color*, *zip_code*
- New features that we didn't use:
  - *claim_day*, *claim_month*, *claim_year*, *weekday*, *SP_Index*, *umemployment_rate*, *state*
- New features that we kept:
  - *latitude*, *longitude*, *interest_rate*, *rating_per_claim*

# Explanation of model using LIME (correct classification)

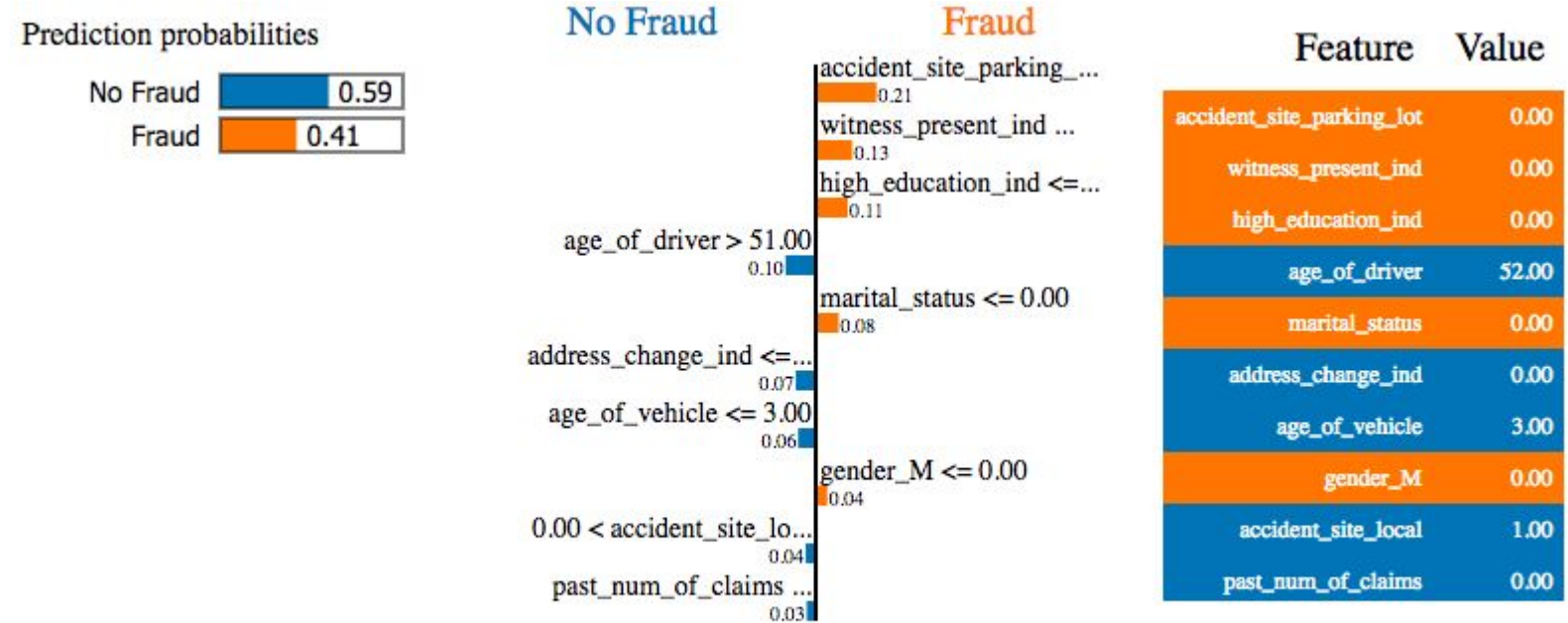



| Feature | Value |
|---|---|
| accident_site_parking_lot | 0.00 |
| witness_present_ind | 0.00 |
| high_education_ind | 1.00 |
| marital_status | 0.00 |
| address_change_ind | 0.00 |
| age_of_driver | 40.00 |
| gender_M | 0.00 |
| rating_per_claim | 87.00 |
| age_of_vehicle | 6.00 |
| accident_site_local | 0.00 |

```
Predicted value by Model Averaging is 1
Actual value is 1
```

# Explanation of model using LIME (incorrect classification)




```
Predicted value by Model Averaging is 0
Actual value is 1
```

# Insights: Significant Features

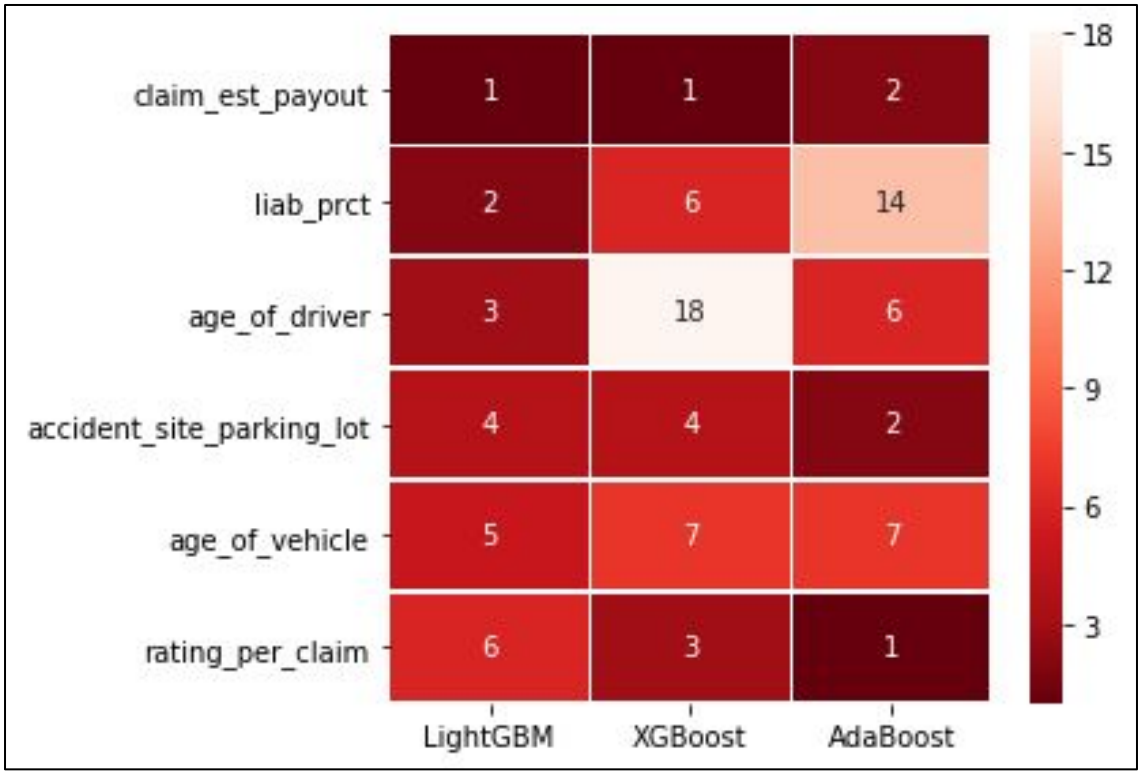



Table shows ranked feature importance for each model.

# Insights

- When the economy is in recession, there is a higher chance for people to commit fraud. Among the three macroeconomic indices, the interest rate seems to predict fraud most accurately.
- Localized geographic location improves prediction ability.
- Credit level is associated with fraud.
- Further macroeconomic, geographic, and credit level data could be collected to improve future prediction accuracy.

# References

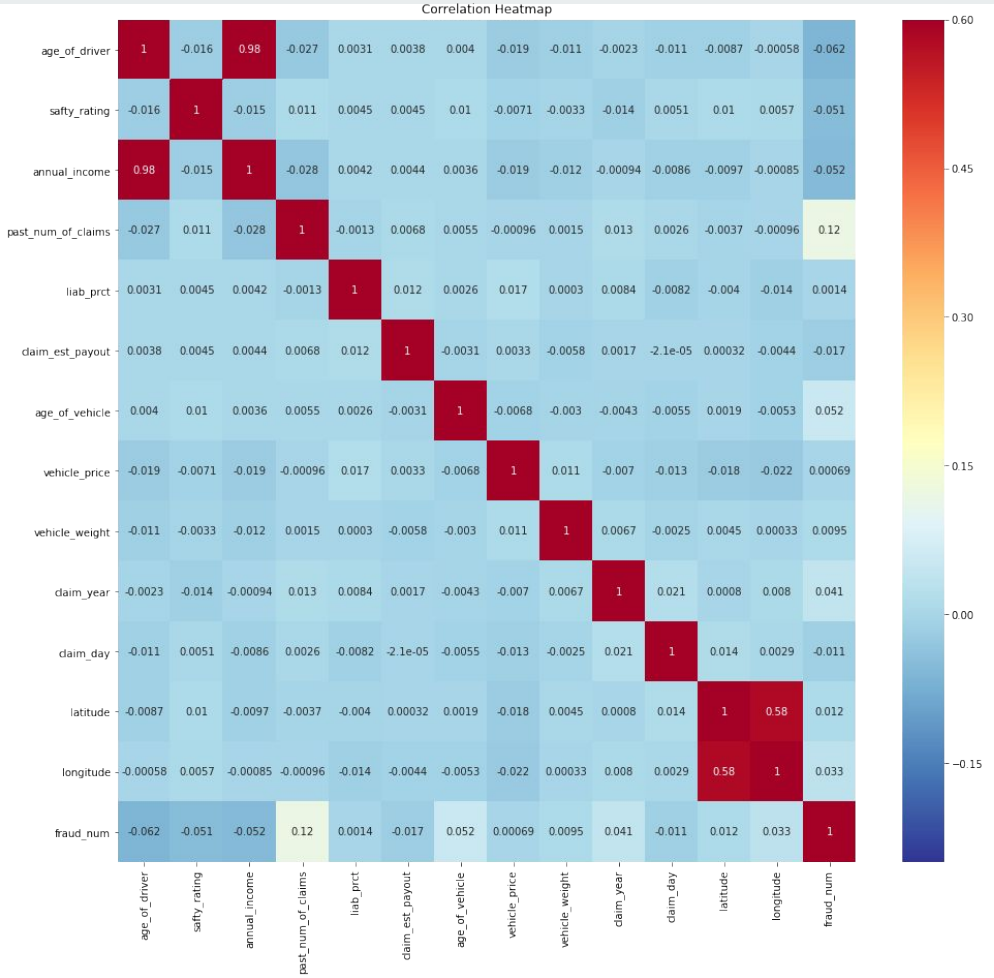

1. M. Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining, 2016
2. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016.
3. Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems*. 2017.
4. Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." *Journal-Japanese Society For Artificial Intelligence* 14.771-780 (1999): 1612.
5. Data Sources:
    a. https://finance.yahoo.com/quote/%5EGSPC/ for Interest rate and S&P 500 Index
    b. https://www.zipcodedatabase.org/zip-code-database/ for latitude/longitude paired with zip code
    c. Bureau of Labor Statistics (https://www.bls.gov/home.htm) for unemployment rate

# Q & A

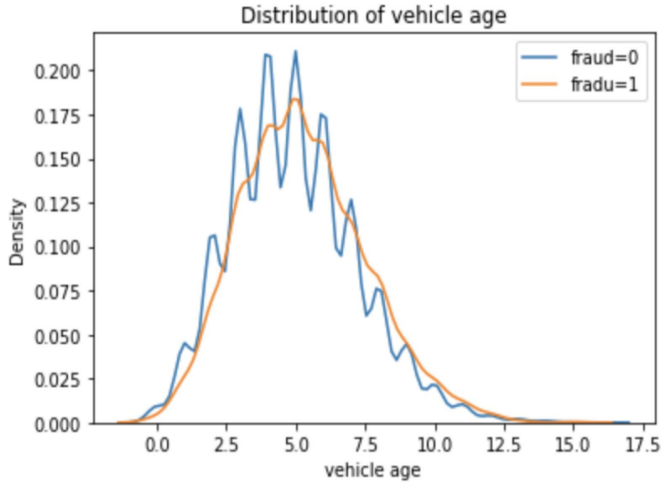
Thank you!

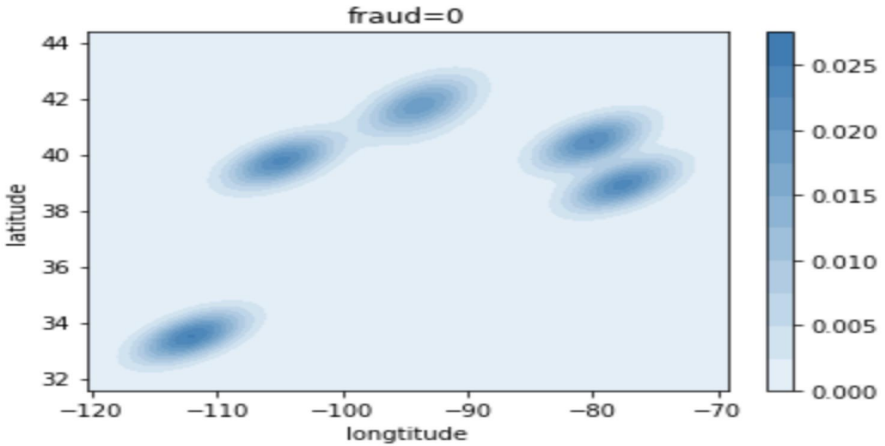
# Backup Slides

# Correlation Matrix

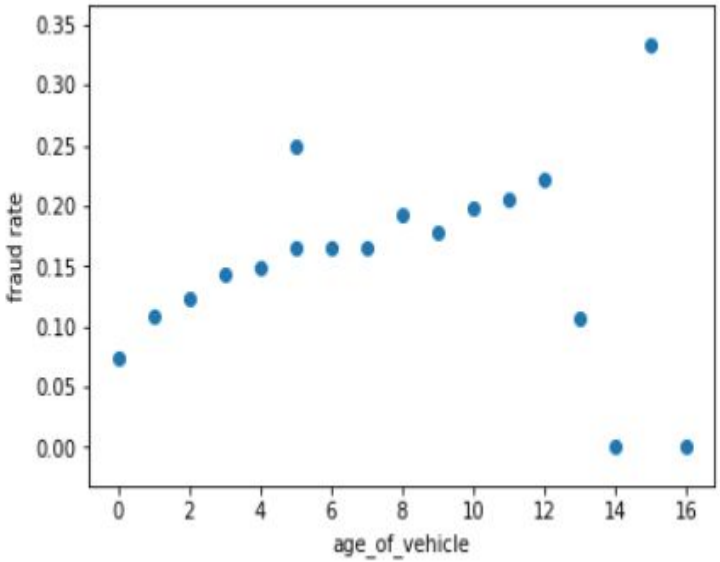


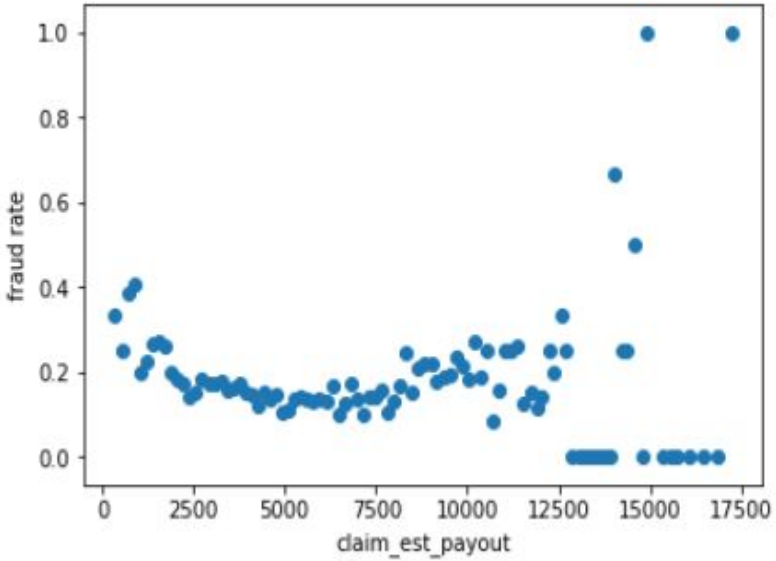
# Kernel Density Curve



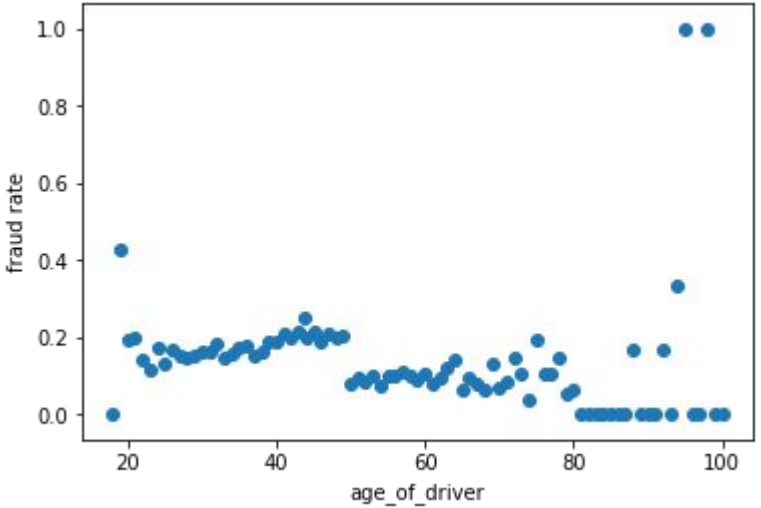



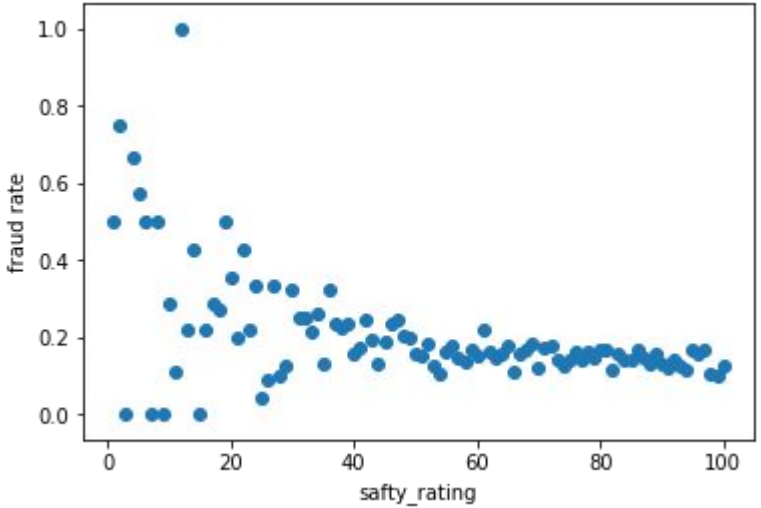
# Scatterplots



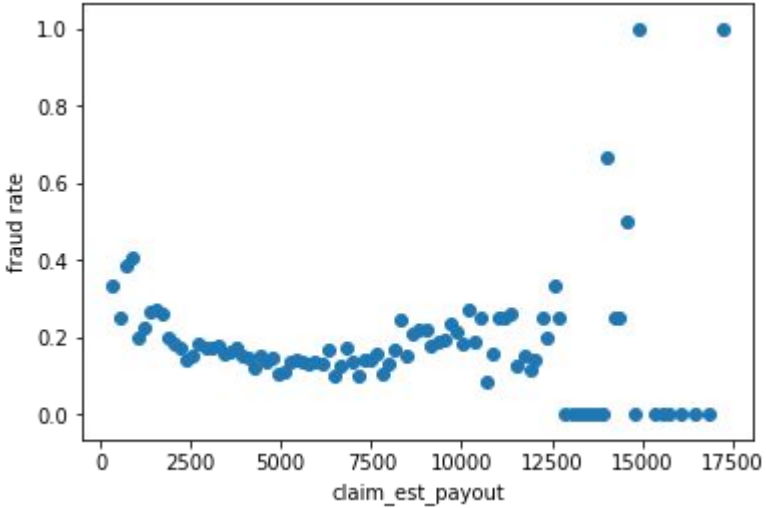



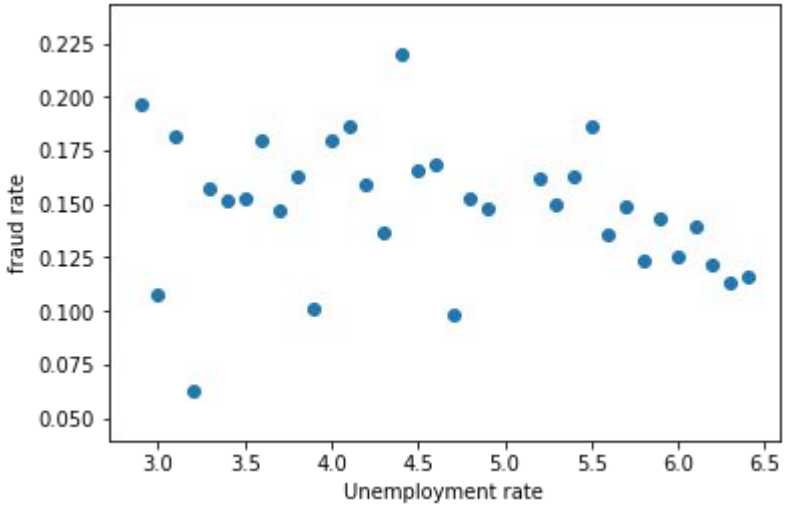
# Scatterplots



# Scatterplots



# Scatterplots