TRAVELERS

Team 3W

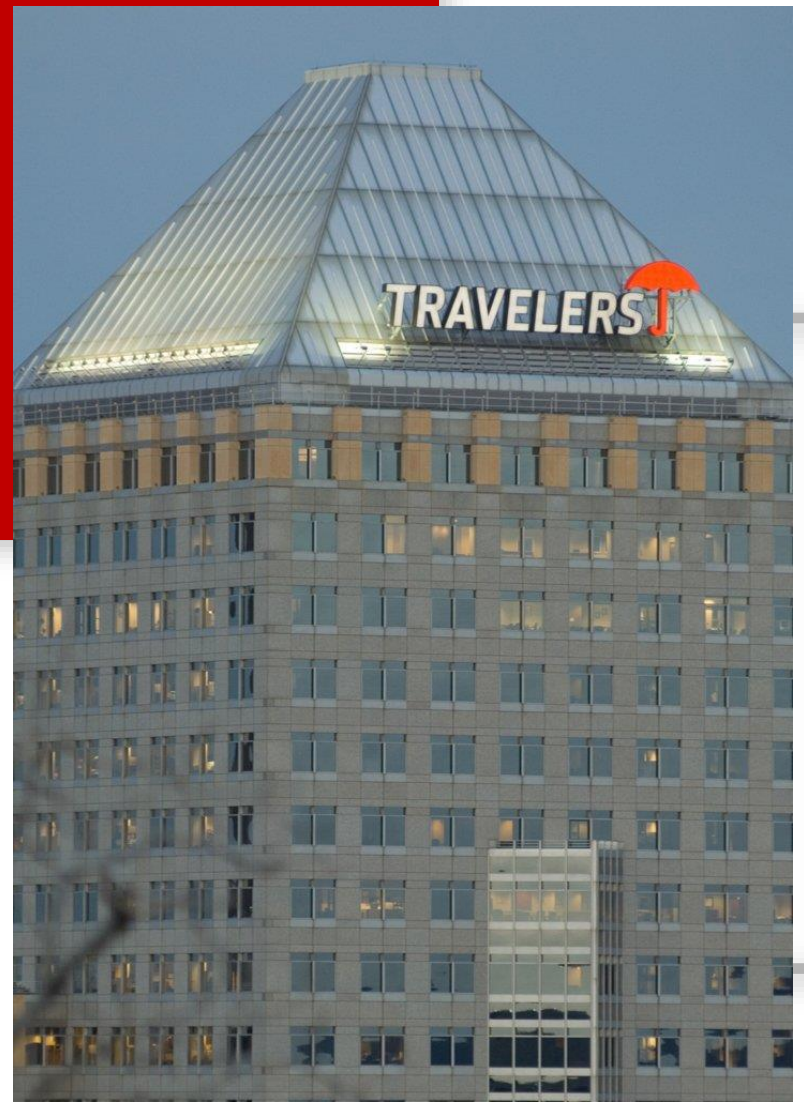
# Fraud Detection

## Case competition presentation

Team member: Wenye Qiu, Yaeeun Lee, Aoran Wang

Emory MSBA 2024

2021




# Agenda

01 . **Background**
What is the business problem?

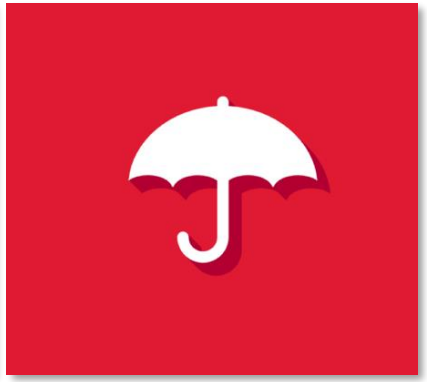
02 . **Data exploration**
Data visualizations and cleaning

03 . **Feature Engineering**
Feature selection and transformation

04 . **Models**
Model selection and comparison

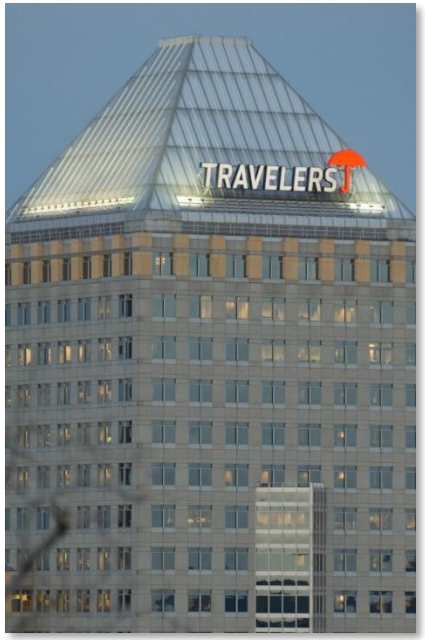
05 . **Implementation**
Cost/benefit Metrix and expected value

06 . **Implementation**
Risk and improvement



# 01

If Not For You
All people have is each other and that it's better under the umbrella




PART ONE

# Background

Insurance Fraud challenge
Fraud loss
Fraud detection

# Business Problem

## Insurance Fraud challenge

More than **7,000** companies
over **$1 trillion** in premiums each

Fraud accounted for about **10 percent** of the property/casualty (PC) insurance industry incurred losses and loss adjustment expenses each year.
**$40 billion** per year
average U.S. family between **$400 and $700** per year

Healthcare, workers compensation, and **auto insurance** are generally considered to be the sectors most affected.

Industry Facts

01

# Fraud Types and detection

## No-fault auto insurance

Unscrupulous medical providers, attorneys and others perpetrate fraud by padding costs associated with a legitimate claim

## Salvage fraud

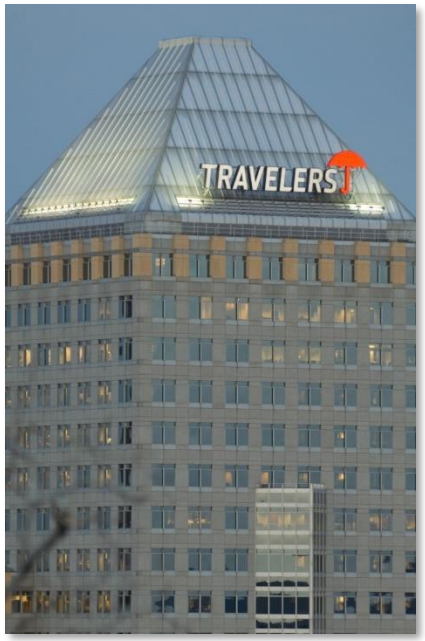
switch or clone manufacturers' serial number plates and put them on a flooded vehicle that has been repaired.


## Misrepresenting facts

## False reports of stolen vehicles

.....

**Travelers has a nationwide staff of experts who investigate a wide array of insurance fraud schemes using in-house forensic resources and other technological tools. This staff also has specialized expertise in fire scene examinations, medical provider fraud schemes, law firm fraud schemes and data mining. Travelers also dedicates investigative resources to ensure that violations of law are reported to and prosecuted by law enforcement agencies.**




# 02

If Not For You
All people have is each other and that it's better under the umbrella

PART TWO

# Data exploration

Data visualization dashboard
Data cleaning

02

# Data Description and cleaning

## Data Description

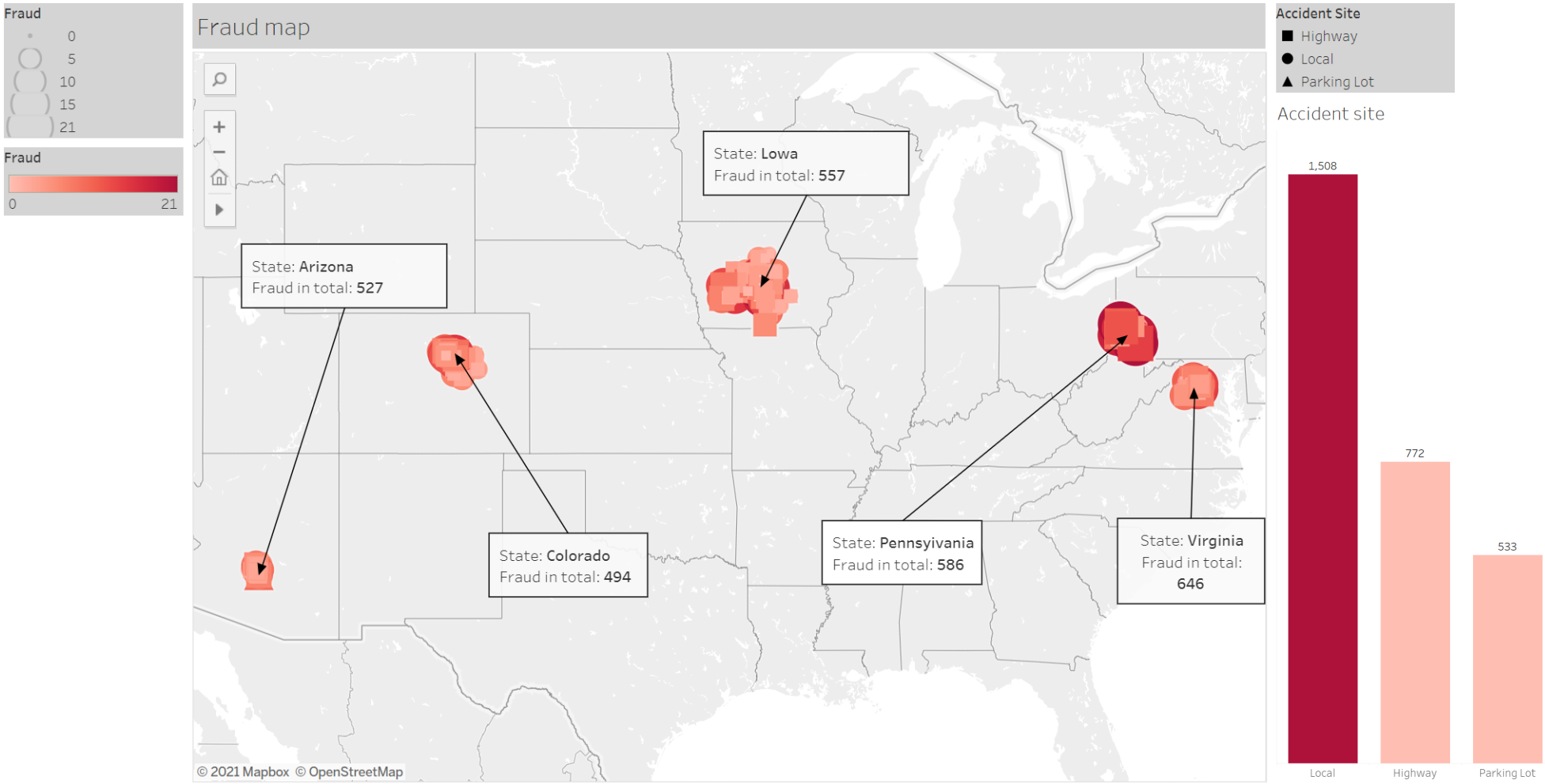
- This training dataset consists 17,998 observations.
- There are 269 missing values, 0.9% of two datasets.
- The original data contains 24 columns. After dropping meaningless attributes like client number, it contains 19 columns now for our data exploration.

## Data cleaning

- Replace missing values
- Replace false values
- No specific outliers need to be considered

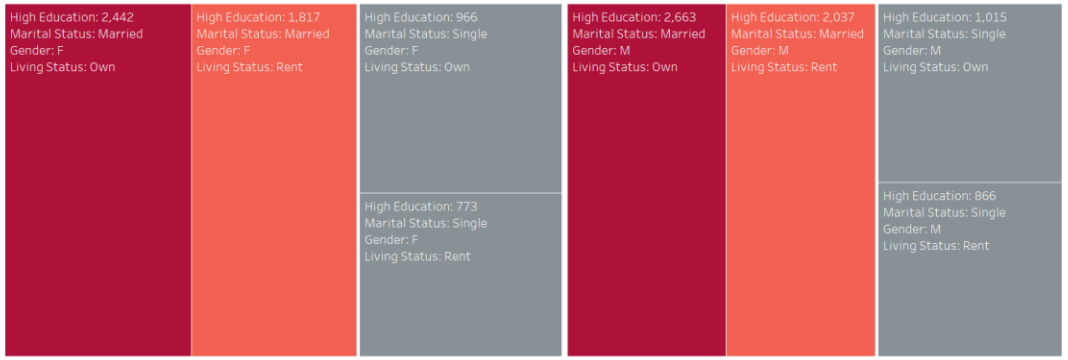
| | claim_number | age_of_driver | gender | marital_status | safty_rating | annual_income | high_education_ind | address_change_ind | living_status | zip_code | past_num_of_claims | witness_present_ind | liab_prct | policy_report_filed_ind | claim_est_payout | age_of_vehicle | vehicle_price | vehicle_weight | fraud |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 | 17998.000000 |
| mean | 14970.601622 | 43.695466 | 0.523058 | 0.712524 | 73.562951 | 37367.655684 | 0.699189 | 0.577286 | 0.446105 | 49875.595955 | 0.505001 | 0.230970 | 49.423269 | 0.600678 | 4975.792083 | 5.008060 | 23089.123114 | 23031.322385 | 0.156295 |
| std | 8659.940765 | 11.959819 | 0.499482 | 0.452598 | 15.346807 | 2957.297249 | 0.458623 | 0.494004 | 0.497101 | 29214.655149 | 0.955504 | 0.421465 | 33.678470 | 0.489773 | 2214.659783 | 2.257889 | 11988.429767 | 12052.385584 | 0.363604 |
| min | 1.000000 | 18.000000 | 0.000000 | 0.000000 | 1.000000 | -1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 282.639432 | 0.000000 | 2457.329316 | 2429.429302 | -1.000000 |
| 25% | 7479.250000 | 35.000000 | 0.000000 | 0.000000 | 65.000000 | 35554.000000 | 0.000000 | 0.000000 | 0.000000 | 20111.000000 | 0.000000 | 0.000000 | 17.000000 | 0.000000 | 3339.205052 | 3.000000 | 14279.574850 | 14164.122133 | 0.000000 |
| 50% | 14965.500000 | 43.000000 | 1.000000 | 1.000000 | 76.000000 | 37610.000000 | 1.000000 | 1.000000 | 0.000000 | 50028.000000 | 0.000000 | 0.000000 | 50.000000 | 1.000000 | 4671.827763 | 5.000000 | 20948.879250 | 20838.150260 | 0.000000 |
| 75% | 22467.750000 | 51.000000 | 1.000000 | 1.000000 | 85.000000 | 39318.000000 | 1.000000 | 1.000000 | 1.000000 | 80038.000000 | 1.000000 | 0.000000 | 81.000000 | 1.000000 | 6254.708103 | 6.000000 | 29562.232780 | 29430.446292 | 0.000000 |
| max | 30000.000000 | 229.000000 | 1.000000 | 1.000000 | 100.000000 | 54333.000000 | 1.000000 | 1.000000 | 1.000000 | 85083.000000 | 6.000000 | 1.000000 | 100.000000 | 1.000000 | 17218.345010 | 16.000000 | 127063.506000 | 123016.650400 | 1.000000 |

# 02 Data Visualization



# 02 Data Visualization

Demofraphic and fraud




Highly educated married female?

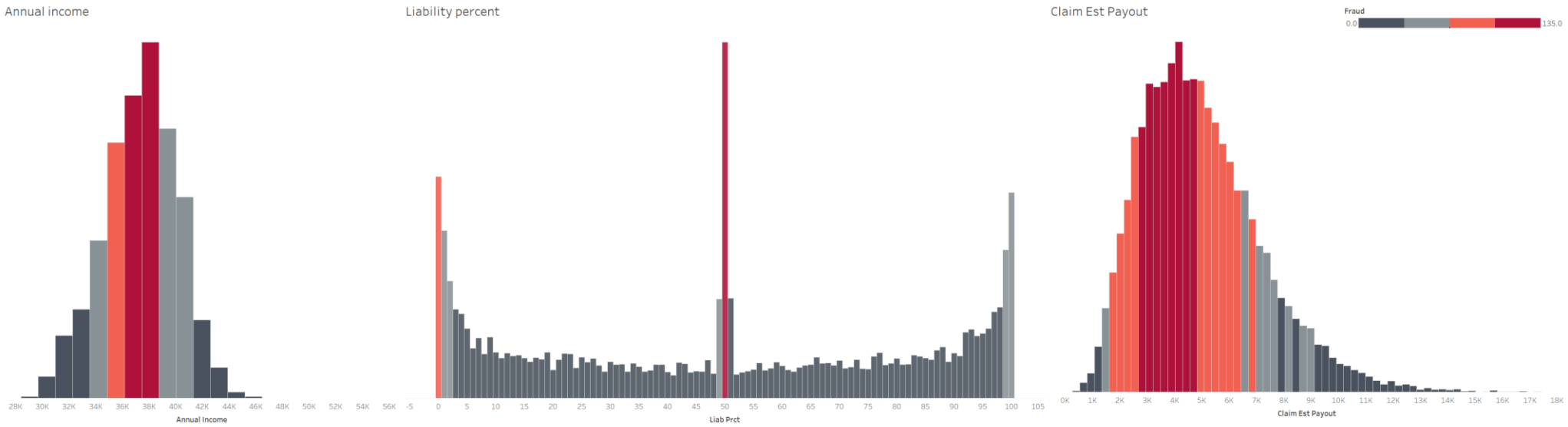
High Education
0 2,700

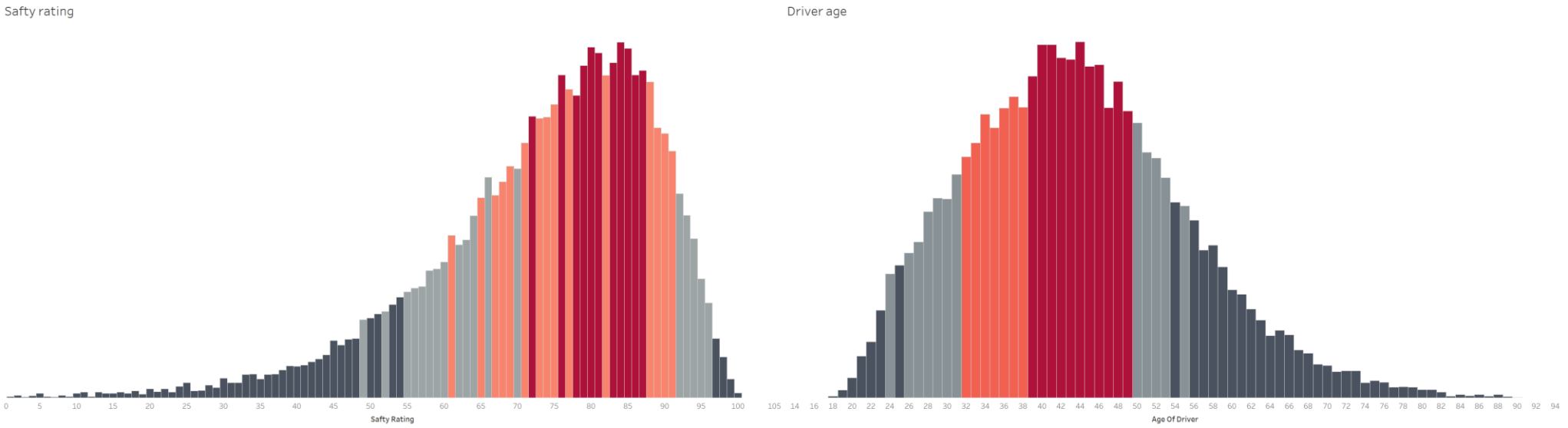
Claim facts


Fraud
0.0 722.0

02

# Data Visualization



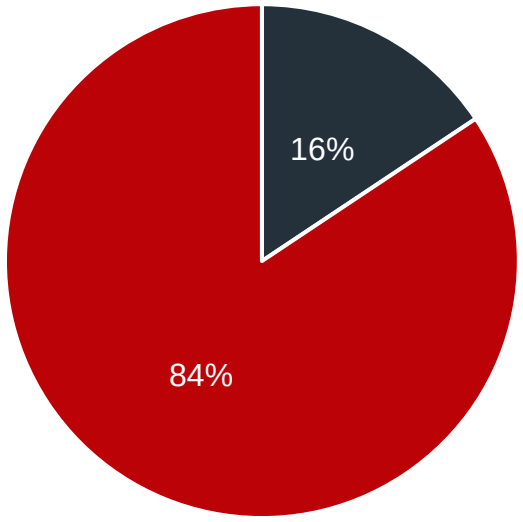



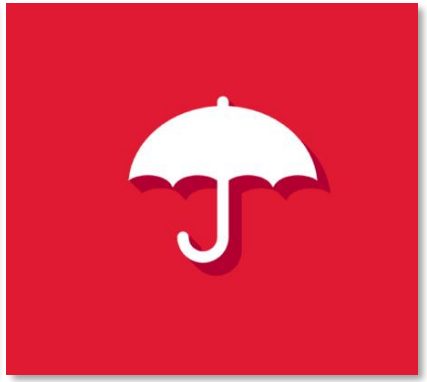
# 02 Data visualizing

## Correlation matrix

- No significantly high correlation
- The highest correlation is with past claim number

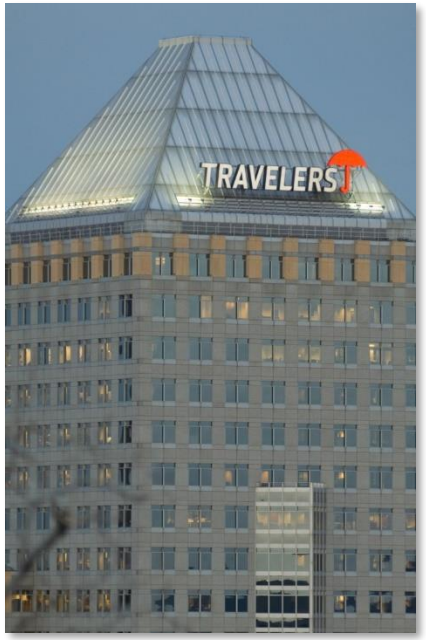
## Class Imbalance



# 03

If Not For You
All people have is each other and that it's better under the umbrella




PART THREE

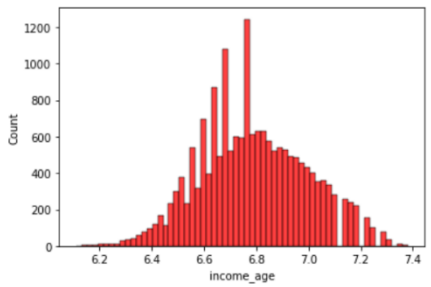
# Feature Engineering

New variables
Transformation
Dropped variables

# 03 Feature Engineering

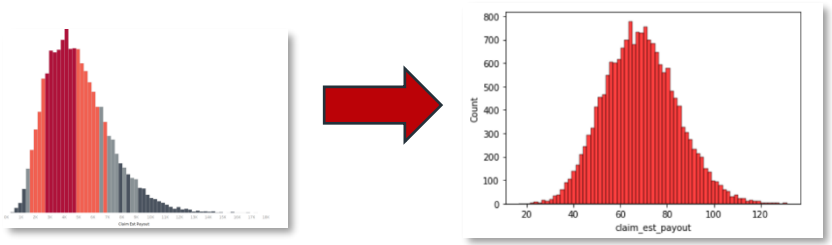
## New Variable

Annual Income/Age of driver




## Transformation

Log, cube root, square root



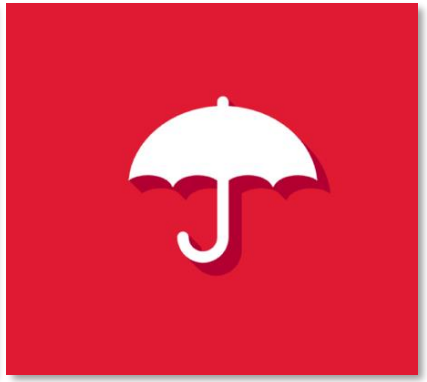

## Dropped Variables

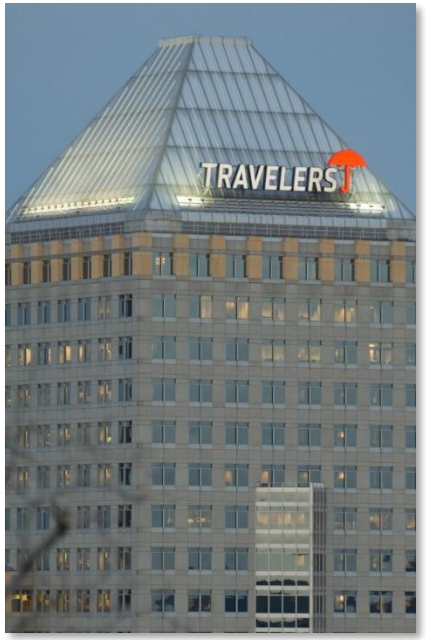
claim number

claim date

zip code

vehicle color






# 04

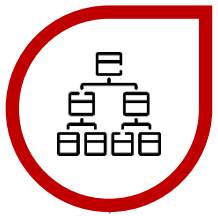
If Not For You
All people have is each other and that it's better under the umbrella

PART FOUR

# Modeling

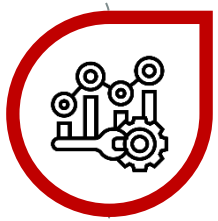
Selection process
Comparison
CatBoost

# 04 Selection Process



## Build model with different algorithms

- Decision Tree, Random Forest, Support Vector Machine, XGBoost, CatBoost



## Optimize each model to have the best AUC

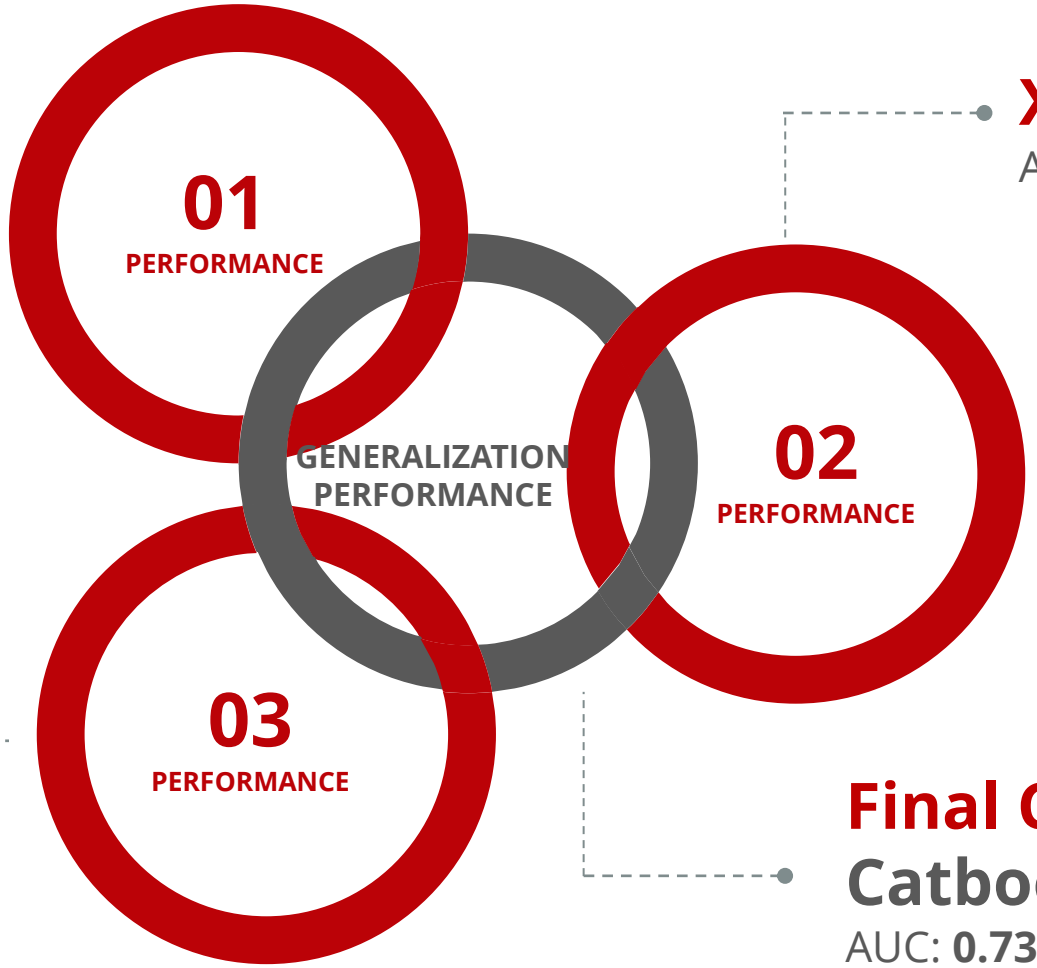
- Grid Search
- 3-fold Cross-Validation

## Select the model with the highest AUC

# 04 Model Comparison

**Catboost**
AUC: **0.73**

**Xgboost**
AUC: **0.70**

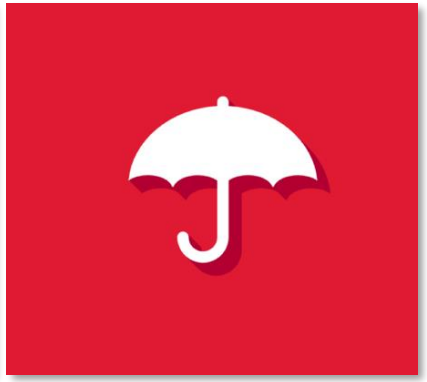



**Logistic Regression**
AUC: **0.68**

**Final Choice:**
**Catboost**
AUC: **0.73**

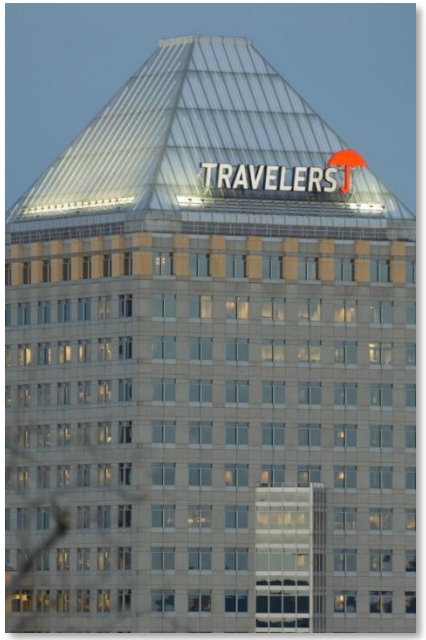
# 04 Catboost

- No one-hot-encodings/sparse data frame
- Keeps original format of data frame, making collaboration easier as well
- Training is faster
- Categorical features are more important
- Model is more accurate
- You can now work with features that you could not before like ID’s, or categorical features with high unique counts



# 05

If Not For You
All people have is each other and that it's better under the umbrella



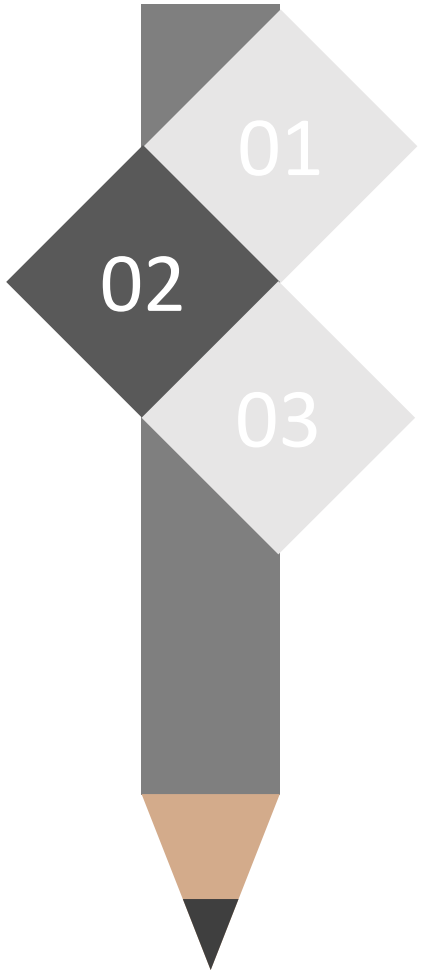

PART FIVE

# Implementation

Expected Value
Timeline

# 04 Expected Value

| Matrix of Probability | | |
|---|---|---|
| | Positive | Negative |
| Positive | 84.33% | 0.02% |
| Negative | 15.17% | 0.48% |




| Confusion Matrix | | |
|---|---|---|
| | Positive | Negative |
| Positive | 15175 | 4 |
| Negative | 2729 | 87 |

| Cost & Benefit Matrix | | |
|---|---|---|
| | Positive | Negative |
| Positive | Intervention Cost - estimated payout | Intervention Cost+ Customer |
| Negative | estimated payout | No cost & benefit |

# 04 Timeline

## Utilize the Model

Input data into the model to predict whether a claim is fraud or not

02

## Investigation

Start investigate on claims that are highly likely to be fraud

04

01

03

05

## Collect & Clean Data

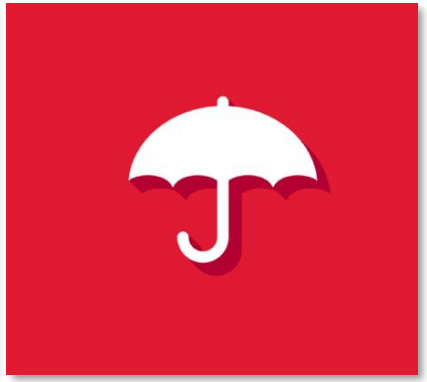
Collect data coming in from users and conduct data cleaning

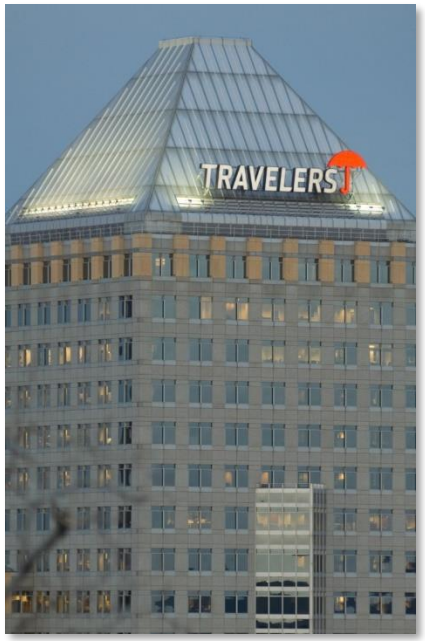
## Confirm Results with Experts

Confirm the prediction with industry experts that have domain knowledge. Combining machine learning and experts' knowledge achieves higher certainty

## Model Improvement

Use the investigation results as past data, keep updating the model to yield better performance in fraud detection






# 06

If Not For You
All people have is each other and that it's better under the umbrella

PART SIX

# Conclusion

Risk and Challenge
Future improvement

06

# Risk and Challenge

01 Binary VS Multi-class Classification (different types of fraud)

02 Supervised VS Unsupervised Learning

03 Noise VS Anomaly

04 Limited model implementation

06

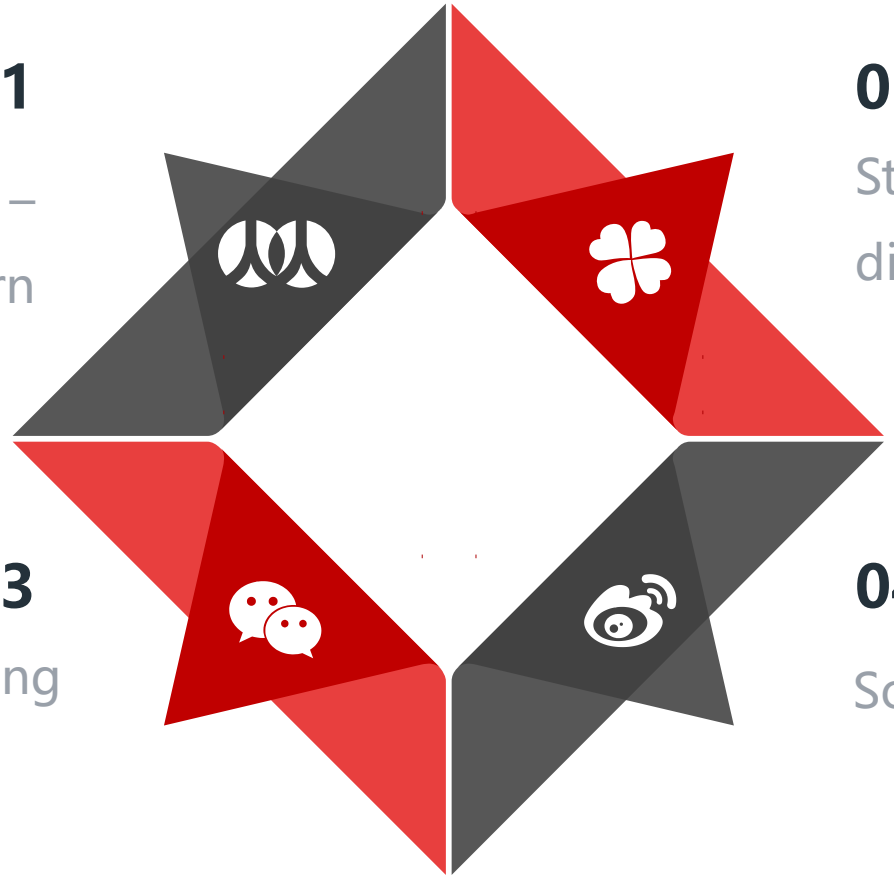
# Future Improvement

**01**

Time Series Analysis – detect pattern

**02**

Statistical Test on significant differences



**03**

Unsupervised Learning for unlabeled data

**04**

Social networking analysis

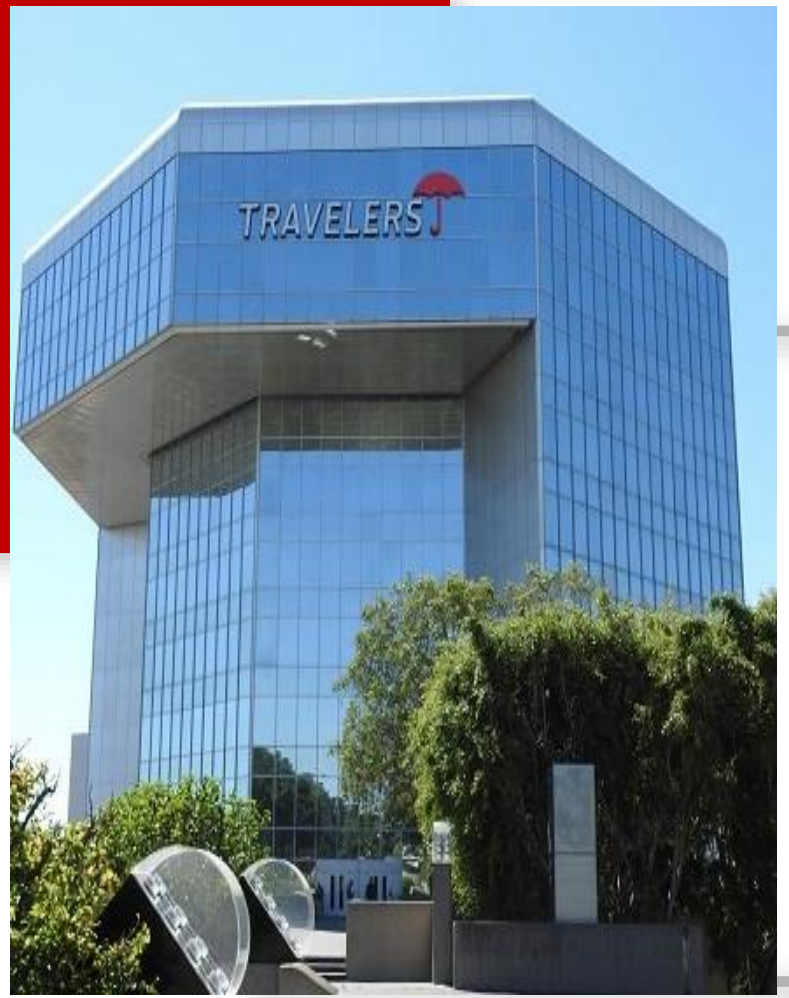
TRAVELERS

# THANK YOU

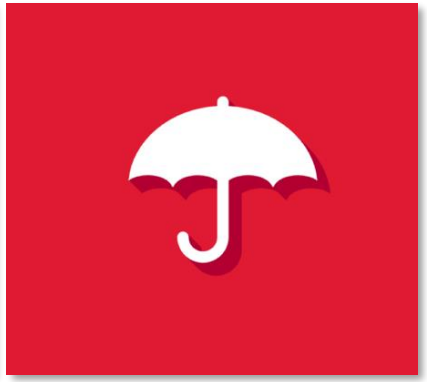
# Q&A

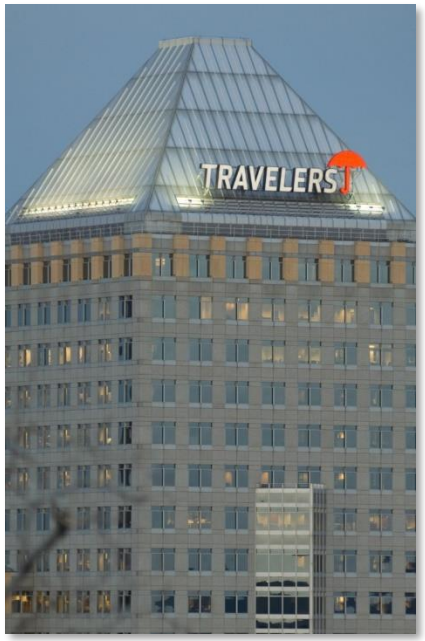
**Team 3W**

Emory MSBA 2022




2021





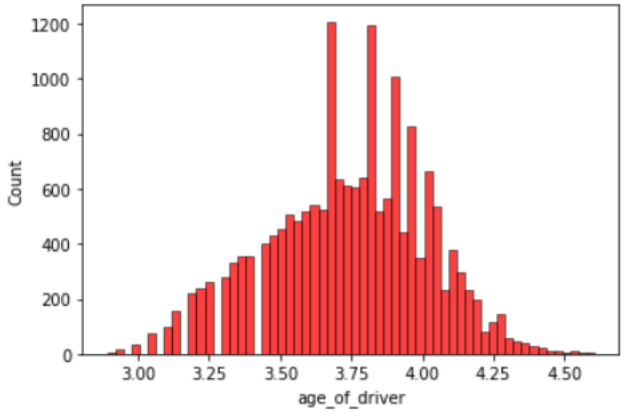

# 07

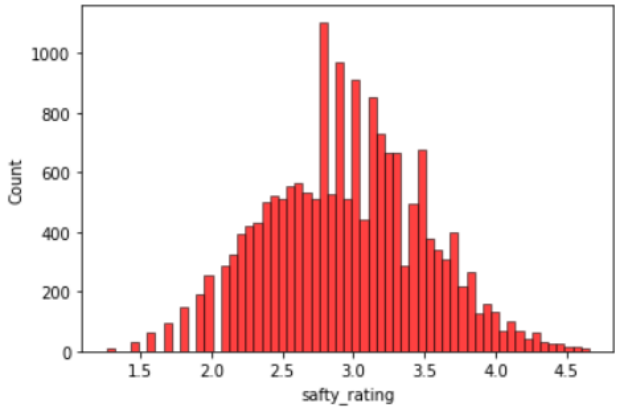
If Not For You
All people have is each other and that it's better under the umbrella

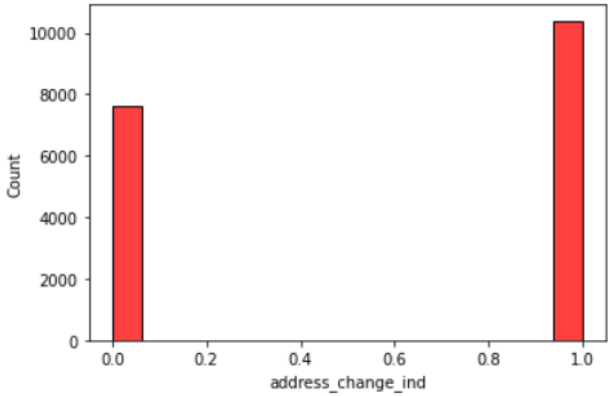
PART SEVEN

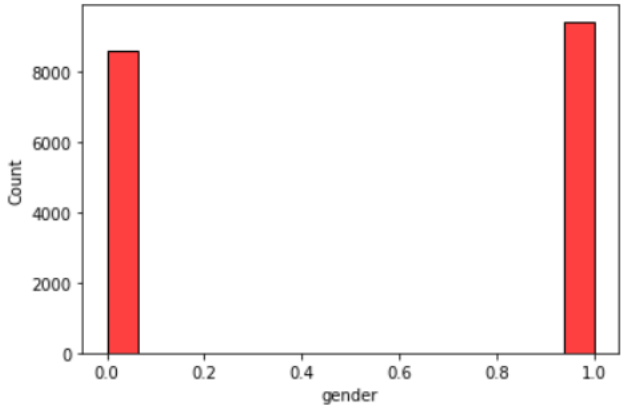
# APPENDIX

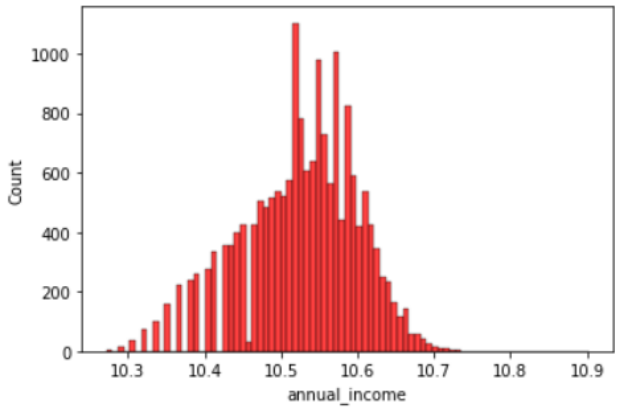
# Histogram of All Variables

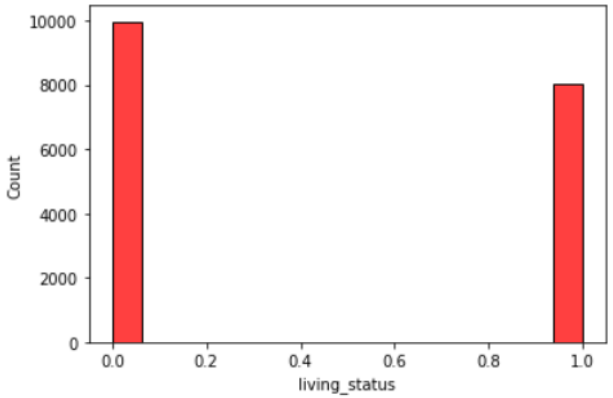


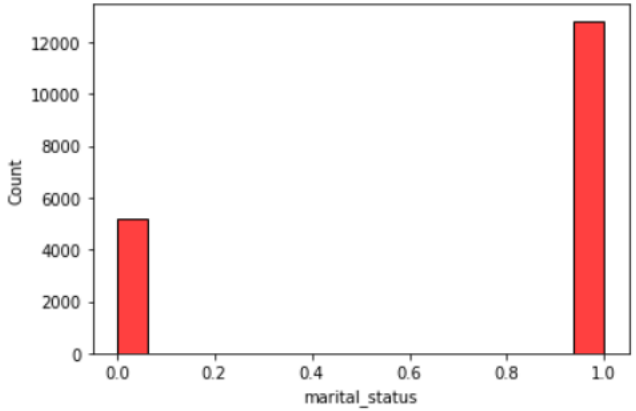



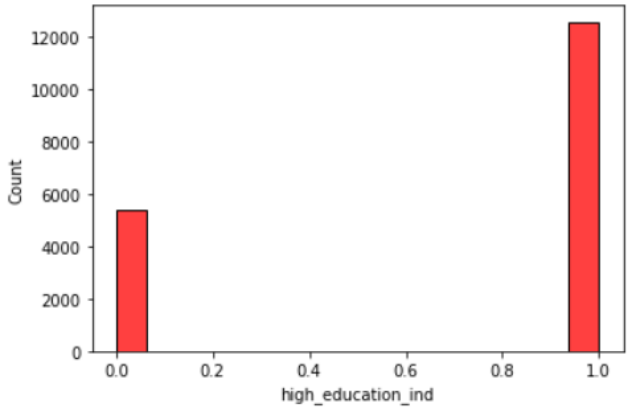



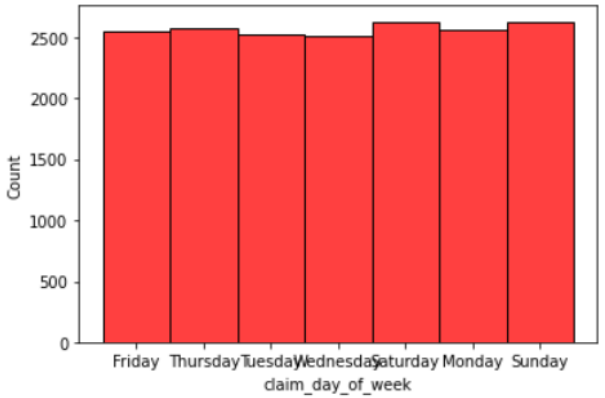



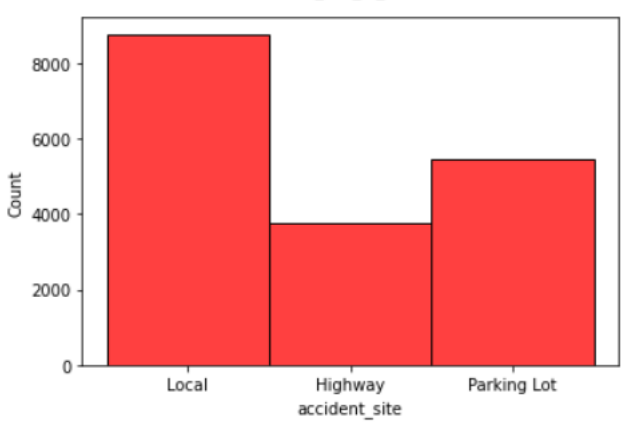



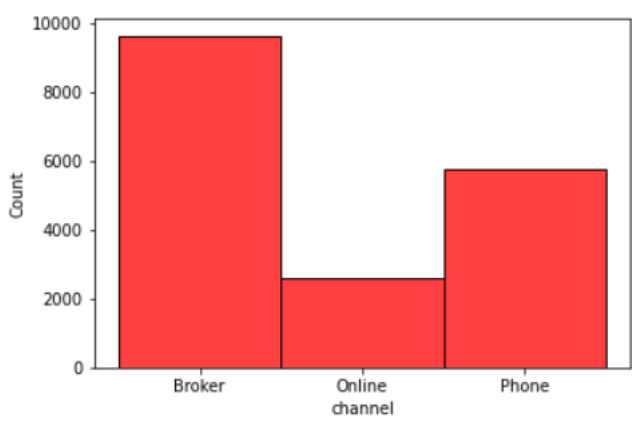



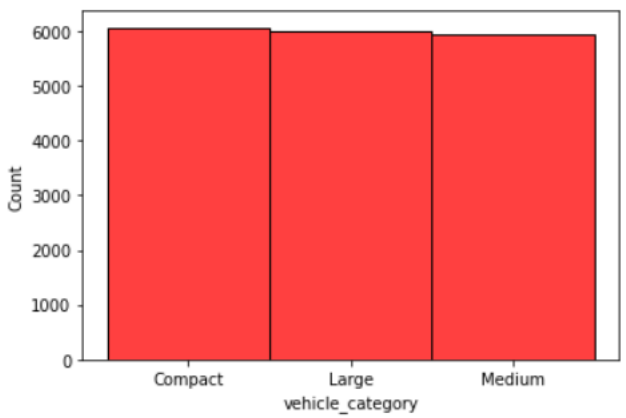



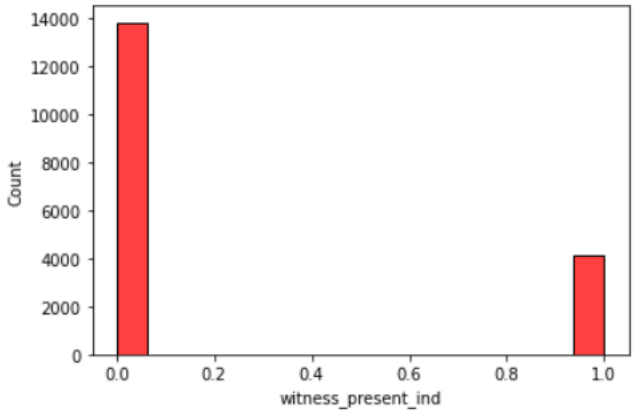



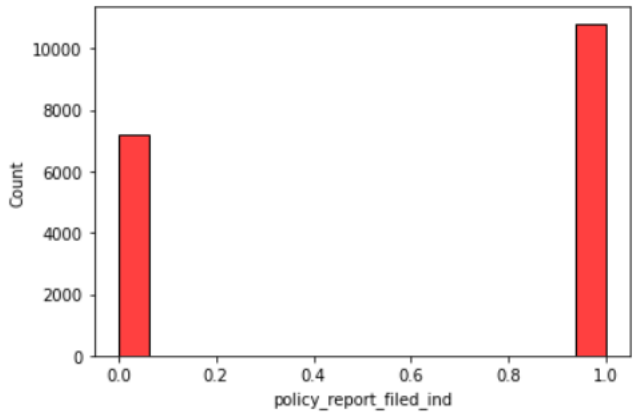
# Histogram of All Variables

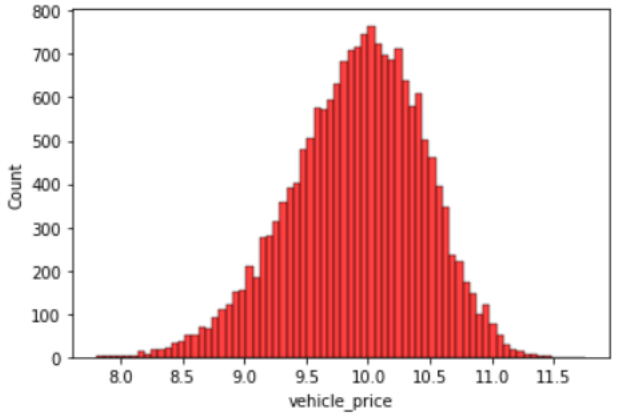


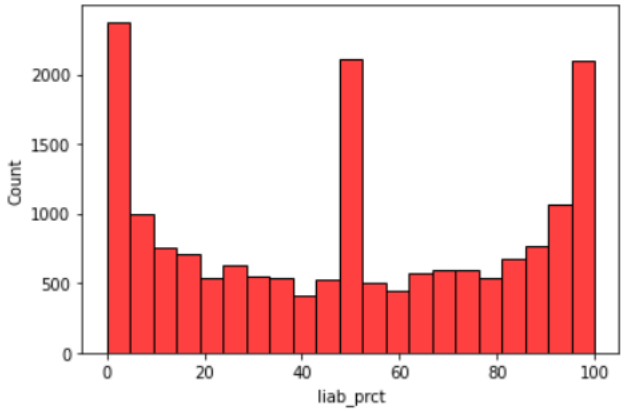



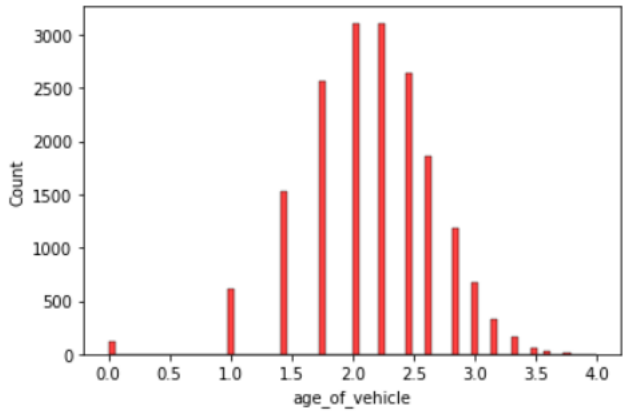



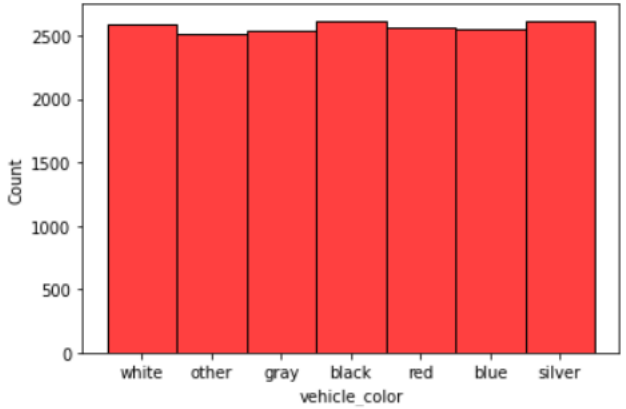



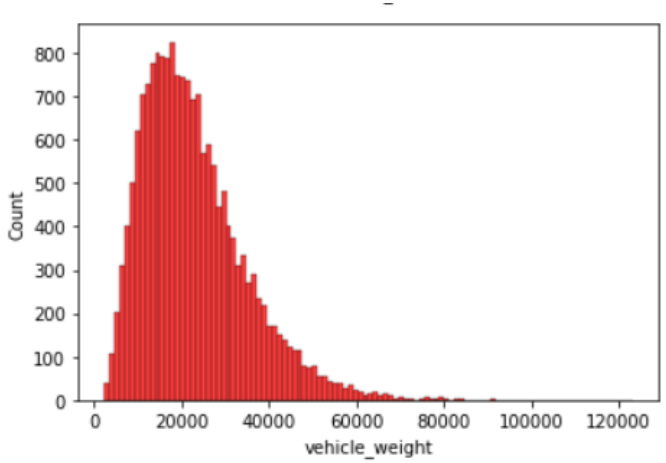



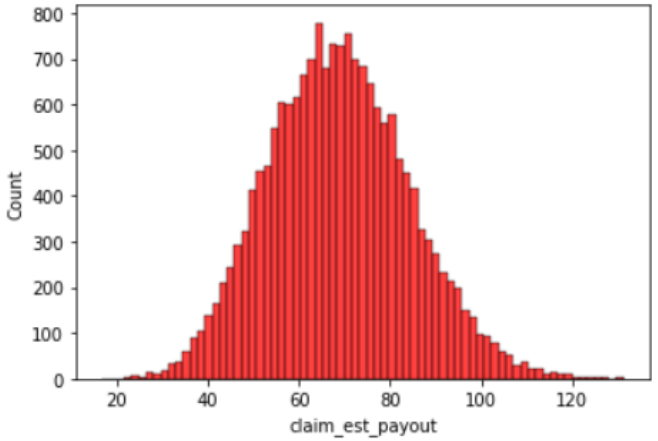



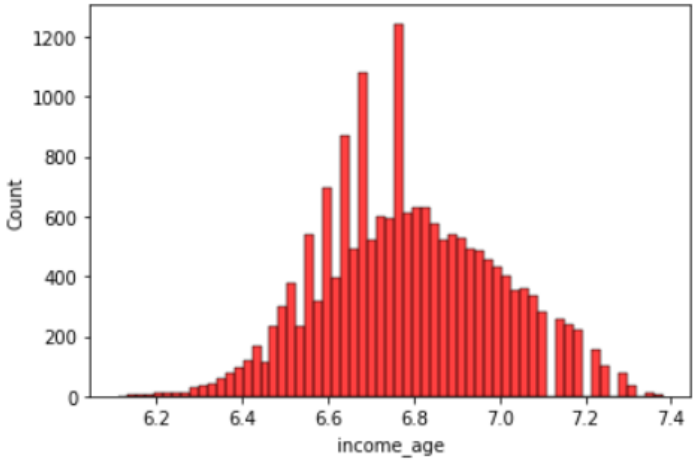



# Histogram of All Variables



# Feature Selection - L1 based feature selection

```
lr.coef_
```

```
array([[ 2.90183474e-06, -4.46563557e-01, -2.67243669e-01,
        -4.59833964e-01,  2.56639253e-01, -8.52156822e-02,
        -5.86010920e-01,  4.42227198e-01,  1.19875632e-01,
        -7.07094737e-01,  3.31403837e-04,  2.04306785e-01,
        -4.36238118e-03,  3.16016087e-01,  3.23693734e-02,
         1.87837462e-06,  9.10032325e-02, -8.15908285e-04,
         4.57435241e-02, -2.12998572e-02,  1.40872939e-01,
         1.23276551e-01, -2.25517366e-01, -1.00749261e+00,
        -2.02254637e-01,  2.81175374e-02, -9.37593224e-02,
         1.16212646e-02]])
```

# Feature Selection - Sequential feature selection algorithms