TRAVELERS
INSURANCE FRAUD MODEL

Emory MSBA
Saturday Working Club

MEET THE TEAM

Kyu Kim
MSBA Candidate

Steven Sun
MSBA Candidate

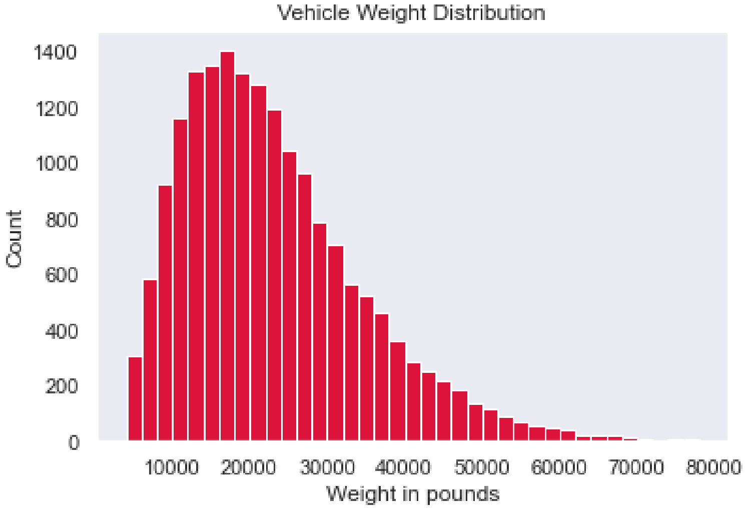
Kenneth Lee
MSBA Candidate

TABLE OF CONTENTS

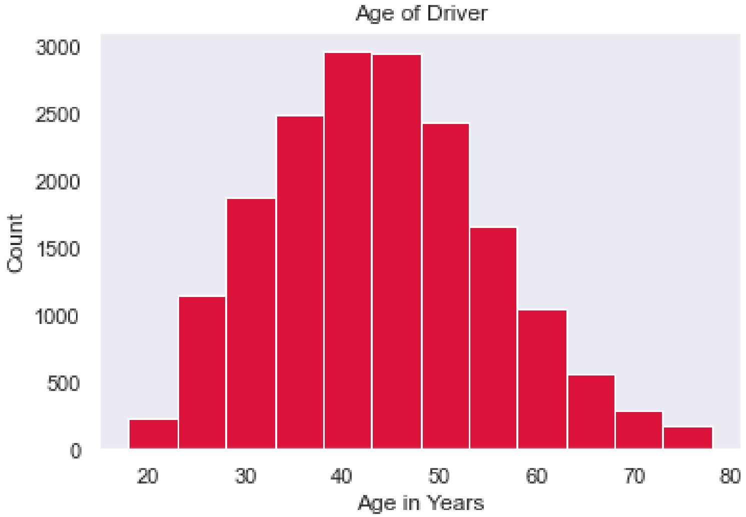
1

DATA UNDERSTANDING

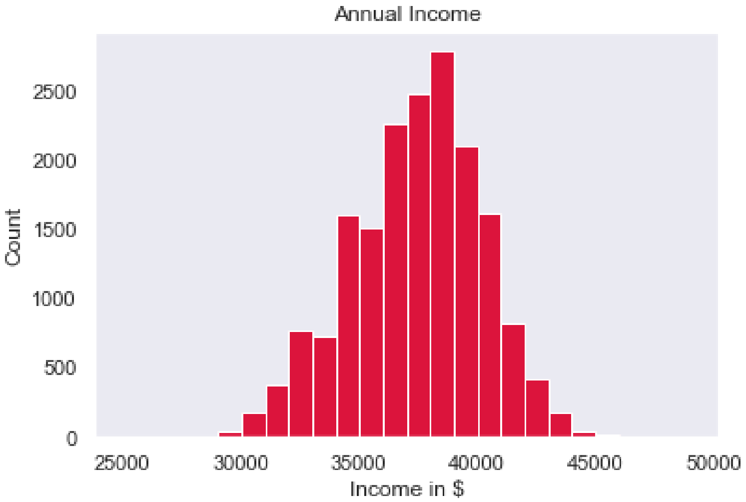
WHAT CLUES DOES THE DATA GIVE US?



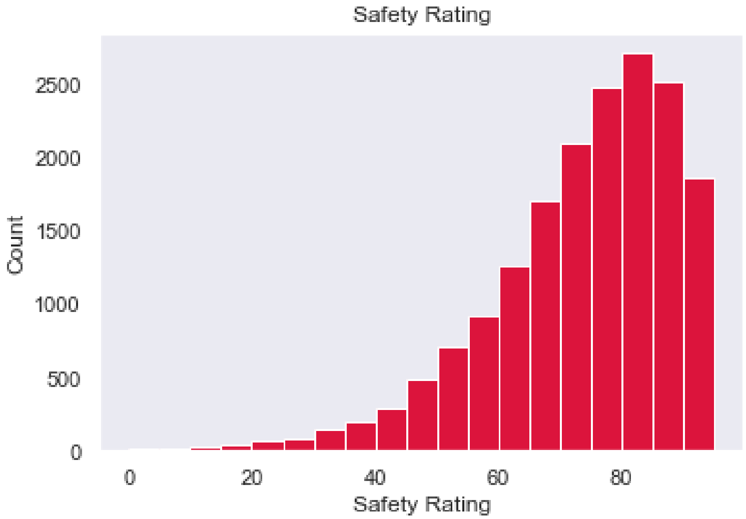
Median weight of all vehicles: 20,838 pounds



Median age of drivers: 43 years old



Median salary: $37,610



Median safety ratings: 76 points

Source: Kaggle Traveler's Dataset

SETTING OUR FOCUS ON THE COMMERCIAL TRUCKING INDUSTRY

Online research on the characteristics led us to believe we are dealing with the trucking industry

Individual identification information allowed us to profile the typical claimant

Led us to conclude that the dataset we were provided was from Northland Insurance, a Travelers Company Subsidiary

Source: Kaggle Traveler's Dataset

2

SWC INSURANCE FRAUD MODEL

GATHERING WEB DATA

Zip Code Information

- Source: uszipcode 0.2.6 Python Package
- 276 Unique Zip Codes
- 11 Features
  - Zip code type, Latitude, Longitude, Population density, Land area, Water area, Housing units, Occupied housing units, Median home value, Median household income

2016 Offenses Known to Law Enforcement

- Source: FBI Uniform Crime Reporting
- 1287 Cities
- 10 Features
  - Violent crime, Murder, Rape, Robbery, Aggravated assault, Property crime, Burglary, Larceny theft, Motor vehicle theft, Arson

CONDUCTING FEATURE ENGINEERING

Driver Safety Ratio

$$\frac{\text{Driver safty rating}}{\text{Past \# of claims}}$$

Payout Ratio

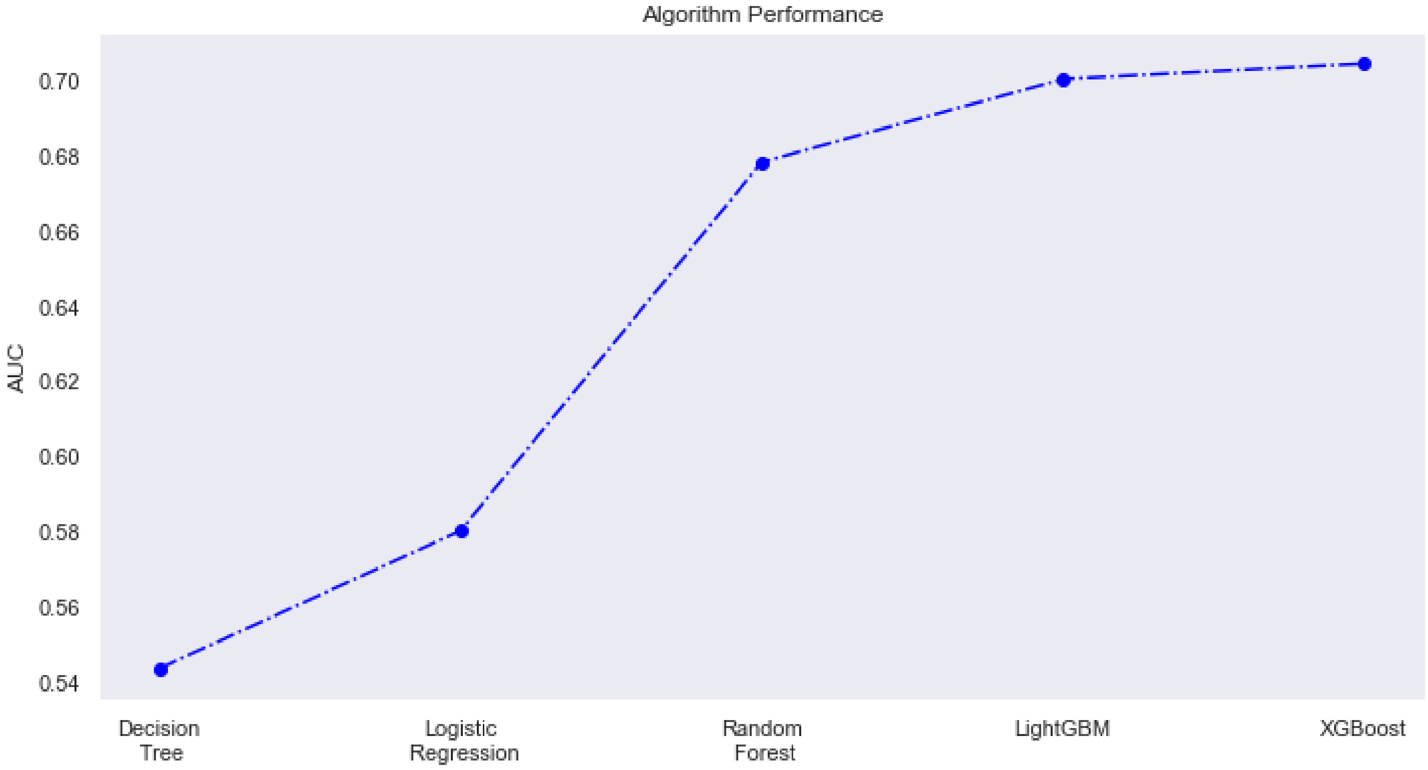
$$\frac{\text{Claim estimate payout}}{\text{Annual income}}$$

Vacancy Ratio

$$\frac{\text{\# of occupied housing units}}{\text{\# of housing units}}$$

Source: Emory MSBA - Travelers Modeling Competition

SELECTING ALGORITHM(S)



Gradient boosting models has best performance

XGBoost

- Higher out-of-box ROC AUC
- Level-wise growth
- High memory usage

LightGBM

- Similar performance
- Histogram binning
- Leaf-wise growth
- Low memory usage
- Faster tuning speed

Source: Emory MSBA - Travelers Modeling Competition

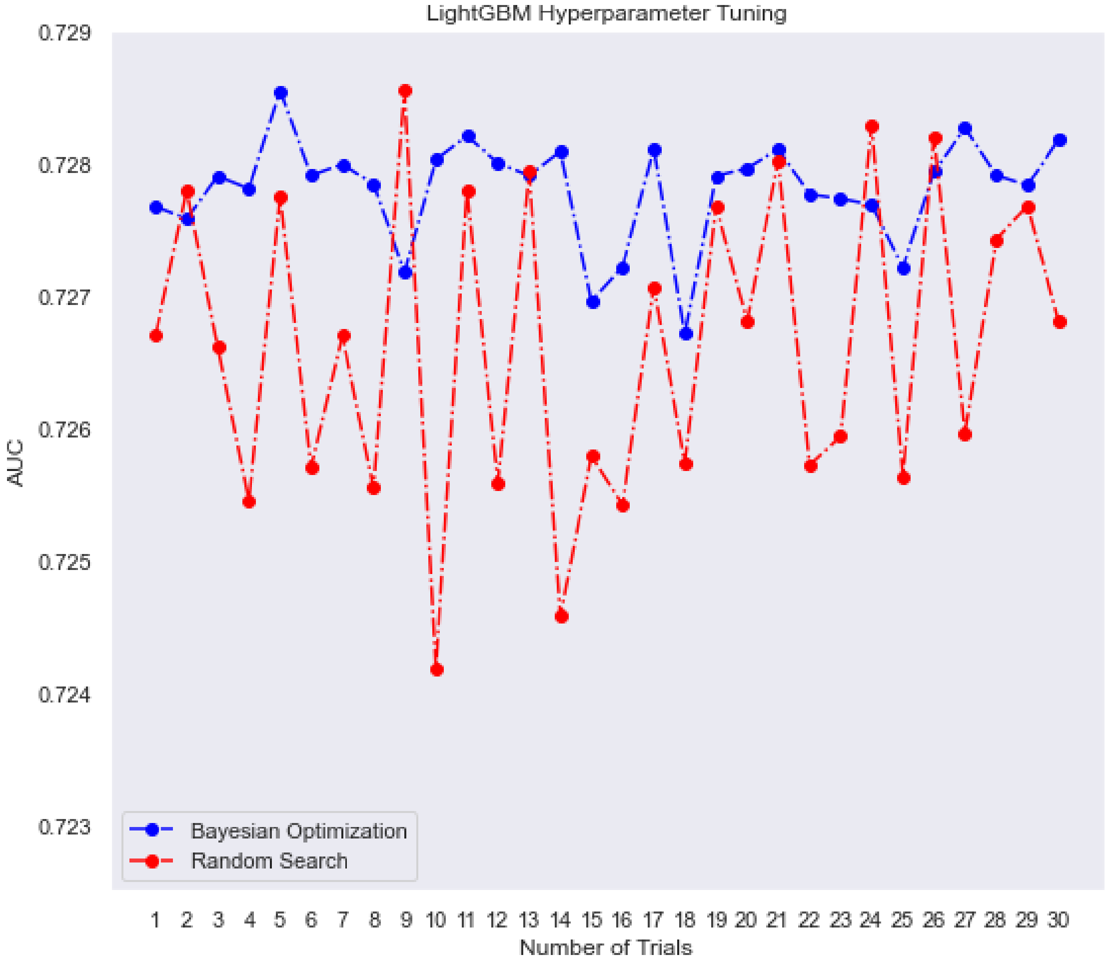
TUNING HYPERPARAMETERS

Parameters

- Learning rate
- Maximum depths
- Number of Leaves
- Lambda L2
- Feature fraction
- Bagging fraction
- Class weight

Optimizers

- Random Search
  - Fastest tuning method
  - Search within provided list
  - Used in preliminary tuning
- Bayesian Optimization
  - Prioritize parameters more promising from past results
  - Search in range of distribution
  - Used in fine tuning

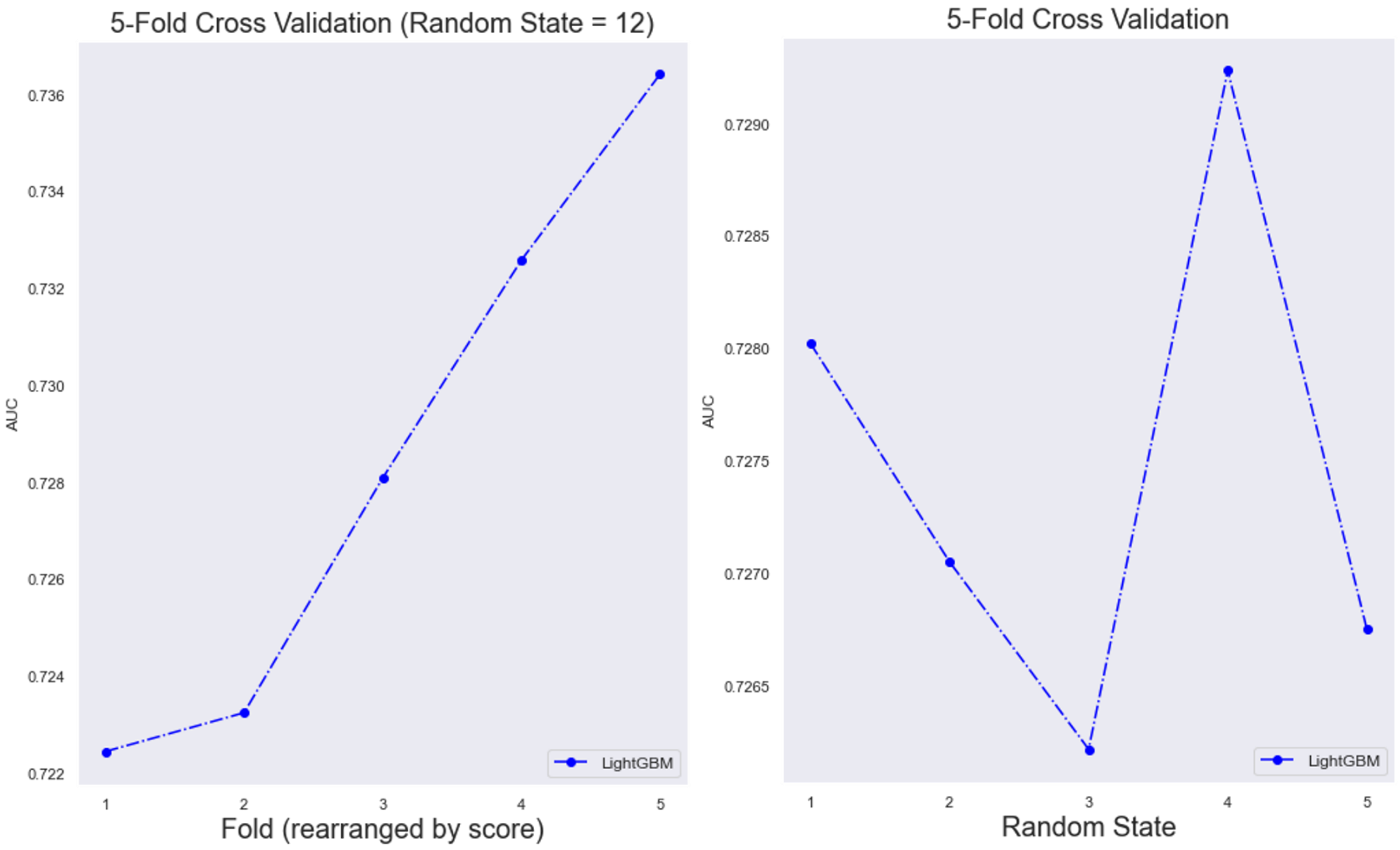


Source: Emory MSBA - Travelers Modeling Competition

EVALUATING PERFORMANCE

Good Fold, Bad Fold

- Holdout testing and cross validation heavily relies on selection of random seeds
  - Seed 12, fold 5 AUC: 0.736
  - Seed 4, 5-fold CV average AUC: 0.7291
- **Sticking to one random seed can be biased**
- Validation across more folds reflects more accurate performance estimation
- What is model selection best practice?



Source: Emory MSBA - Travelers Modeling Competition

TRAVELERS

EVALUATING PERFORMANCE (CONT.)

Model Comparison

- Repeated 100-time 5-fold cross validation
- Why 100 different random seeds?
  - Lower error in mean performance estimation
  - Lower standard deviation than single run CV
- Why 5 folds?
  - Test data size
    - 5-fold cross validation test size: ≈ 3600
    - Public leaderboard test size: ≈ 4800
    - Private leaderboard test size: ≈ 7200

Computational Expense Overview

- LightGBM 100x 5-fold CV run time: 21.1 minutes / 500 fits
- 30 iterations of Baysian optimization fine tuning:
  - 21.1 minutes * 30 ≈ 10.55 hours

Source: Emory MSBA - Travelers Modeling Competition

EXTRACTING MODELING INSIGHTS

Mismatch between Training and Test Score

- 100 times 5-fold cross validation
  - AUC: 0.72745
- Public leaderboard
  - AUC : 0.75826
- Private leaderboard
  - AUC: 0.73795
- Leaderboard AUC is consistently higher
  - Test dataset is more predictable
  - Sampling bias during data collection

Performance Ceiling

"If we can improve cross-validation AUC from 0.72 to 0.73, why can't we improve from 0.73 to 0.74?"

- Kenneth Lee

- Predictive modeling is pattern recognition
  - Pattern recognition suffers from irregularity
  - Pattern recognition cannot find scientific formula
- How to improve?
  - Collect larger dataset for more generalizable patterns
  - Collaborate with actuarial scientists on technical analysis

Source: Emory MSBA - Travelers Modeling Competition

3

MODEL SUPPLEMENTS

IDENTIFYING IMPORTANT DRIVERS OF FRAUDULENT REPORTING

GEOGRAPHIC LOCATION

Housing-Units | Population Density | Accident Site

PERSONAL INFORMATION

Annual Income | Median Household Income | Age of Driver | Safety Rating

CLAIM CHARACTERISTICS

Claim Estimated Payment | Past Number of Claims

Identifying the key drivers of fraudulent reporting will not only be able to help us improve our model accuracy, but also give us a starting point for further investigations

IMPROVING THE MODEL & PERFORMANCE

Type of Claim

- > 80% of automobile insurance claims are bodily injury claims
- US Tax Code states that any medical insurance benefits are non-taxable

Social Security Number

- The SSN is a very good general ID number that can be later used as a primary key for future data addition
- Can be useful in procuring other demographic, socio-economical data in the future

Collaborating with Experts

- Collaborate with experts, actuarial scientists and loss claim specialists who can provide insights into model improvement that generalists can overlook

Adding features and improving surrounding operation can increase both accuracy and efficiency of the model in detecting fraudulent claims

Source: Kaggle Traveler's Dataset, Verisk Insurance Solutions, Insurance Information Institute, Missouri Department of Insurance

4

BUSINESS IMPACTS & ACTION ITEMS

TRAVELERS

WHAT IS THE BUSINESS IMPACT OF IMPLEMENTING OUR MODEL?

$28.7 m

Annual Financial Impact

$11 m

REDUCTION

in “Claims & Claim Adjustment Expenses” on Income Statement

$177 m

MORE

in additional revenue generation due to decreased premium fees

1 %

MORE

in bottom line growth just through the implementation of the model

Source: SEC, Traveler's Company, Versik, FBI, DNB, Northland Insurance

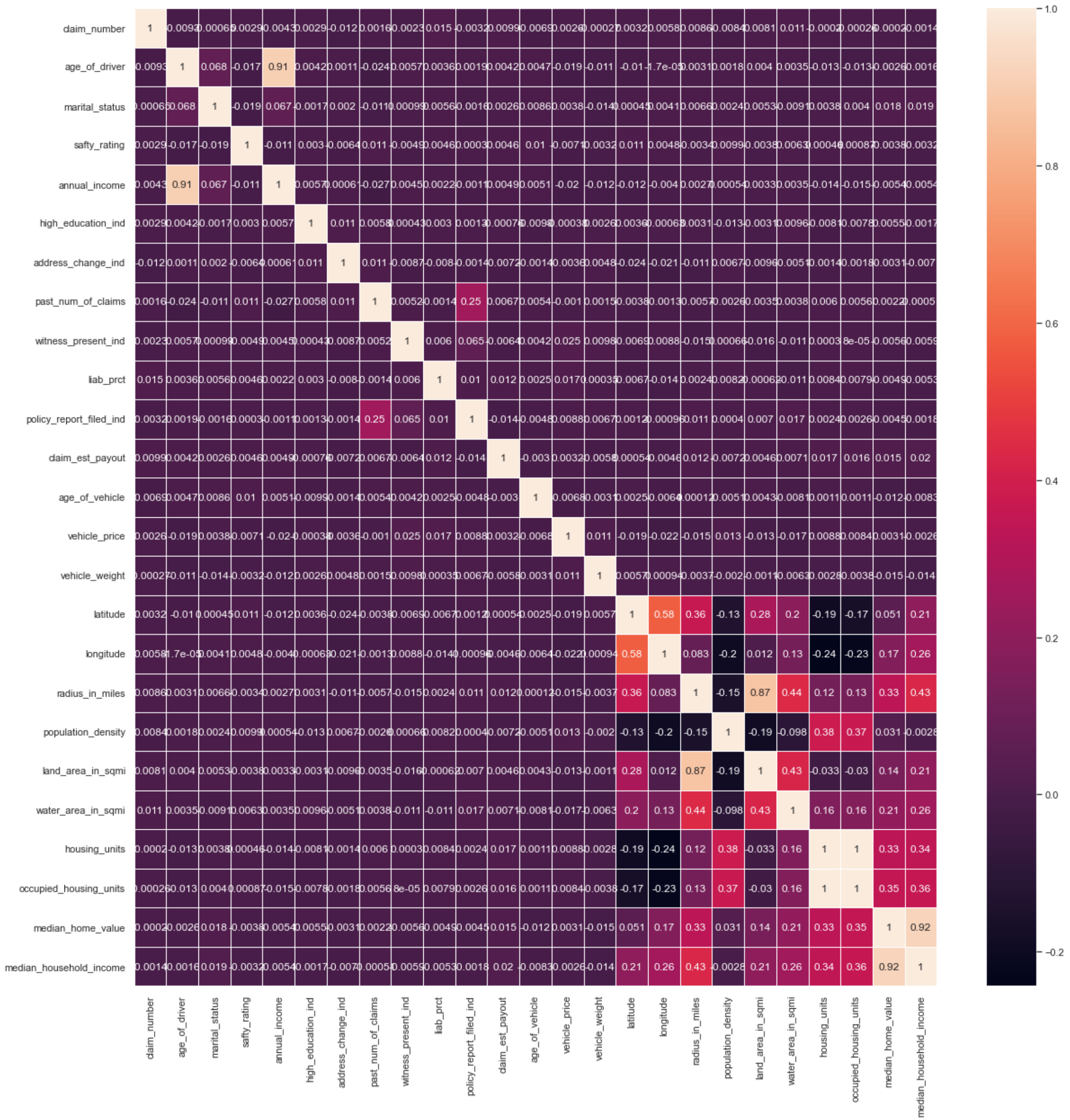
THANK YOU!

ANY QUESTIONS?

APPENDIX

MODELING METHODOLOGY

Feature Engineering → Algorithm Selection → Hyperparameter Tuning → Performance Evaluation → Model Insights



Correlation Map

Summary Statistics

	claim_nun	age_of_dr	marital_st	safty_ratir	annual_in	high_educ	address_c	past_num	witness_p	liab_prct	policy_rep	claim_est_	age_of_ve
count	17998	17998	17993	17998	17998	17998	17998	17998	17866	17998	17998	17981	17990
mean	14970.6	43.6955	0.71272	73.563	37367.7	0.69919	0.57729	0.505	0.23268	49.4233	0.60068	4975.79	5.00806
std	8659.94	11.9598	0.45251	15.3468	2957.3	0.45862	0.494	0.9555	0.42255	33.6785	0.48977	2215.71	2.25839
min	1	18	0	1	-1	0	0	0	0	0	0	282.639	0
25%	7479.25	35	0	65	35554	0	0	0	0	17	0	3337.03	3
50%	14965.5	43	1	76	37610	1	1	0	0	50	1	4668.8	5
75%	22467.8	51	1	85	39318	1	1	1	0	81	1	6255.9	6
max	30000	229	1	100	54333	1	1	6	1	100	1	17218.3	16

	vehicle_pr	vehicle_w	latitude	longitude	radius_in_	populatior	land_area	water_are	housing_u	occupied_	median_h	median_
count	17998	17998	17644	17644	17998	17998	17998	17998	17998	17998	17998	179
mean	23089.1	23031.3	38.9036	-93.4354	4.07096	1396.48	27.5324	0.22191	5222.67	4795.03	149219	484
std	11988.4	12052.4	2.8301	13.4455	4.0796	2766.41	48.975	0.47655	6309.49	5786.03	151614	393
min	2457.33	2429.43	33.3022	-112.247	0	0	0	0	0	0	0	
25%	14279.6	14164.1	38.8318	-105.077	0.39773	9	0.09	0	29	28	0	
50%	20948.9	20838.2	39.7945	-93.4921	3	412	11.37	0.03	1630	1520	112600	482
75%	29562.2	29430.4	40.6252	-79.8568	6	1819	28.9	0.27	9170	8666	200000	650
max	127064	123017	42.633	-77.2888	26	28779	498.67	6.41	29594	25991	813700	1804

Appendix

MODELING DETAILS

Feature Importance

Feature Name	Chi-Squared Feature Importance
housing_units	29320.428918
population_density	13795.514388
vehicle_weight	10068.531405
annual_income	8813.913495
claim_est_payout	4705.814218
safety_rating_percentage	2207.531805
claim_income_percentage	2153.529149
median_household_income	462.358119
past_num_of_claims	459.064911
age_of_driver	208.096947
safty_rating	148.076097
accident_site_Parking Lot	145.293145
witness_present_ind	75.122551
longitude	61.221502
high_education_ind	52.712519
age_of_vehicle	49.760848
address_change_ind	39.769575
marital_status	31.392122
vehicle_price	19.394472
gender_M	18.214358
accident_site_Local	16.605227
state_code_VA	11.095430
state_code_CO	7.100361
living_status_Rent	6.881231
policy_report_filed_ind	6.278338
zipcode_type_Unique	4.820228
vehicle_category_Large	2.743725

Optimal Parameters

```
Optimal Parameter:  {'reg_lambda': 3.4000000000000004, 'num_leaves': 3, 'max_depth': 45, 'learning_rate': 0.015, 'f
eature_fraction': 0.2, 'bagging_fraction': 0.9000000000000004}
Optimal Estimator:  LGBMClassifier(bagging_fraction=0.9000000000000004, drop_rate=0.1,
               feature_fraction=0.2, is_unbalance=True, learning_rate=0.015,
               max_depth=45, metric='auc', min_child_weight=150,
               min_split_gain=0, num_iterations=1500, num_leaves=3,
               reg_lambda=3.4000000000000004, subsample=0.9, verbose=-1)
```

iter	target	baggin...	featur...	learni...	max_depth	num_le...	reg_la...
1	0.728	0.9252	0.6534	0.02029	47.67	3.0	3.317
2	0.7278	0.7946	0.6354	0.0215	7.493	3.0	3.463
3	0.728	0.9032	0.5508	0.02746	35.56	3.0	3.137
4	0.7277	0.846	0.6787	0.02494	18.25	3.0	3.153
5	0.7278	0.6588	0.407	0.04133	48.86	3.0	3.271
6	0.7281	0.5579	0.2591	0.03666	26.47	3.0	3.463
7	0.7278	0.6466	0.2045	0.03578	9.592	3.0	3.238
8	0.7279	0.676	0.5693	0.03382	43.76	3.0	3.284
9	0.7273	0.92	0.4557	0.01253	43.39	3.0	3.253
10	0.7278	0.7358	0.5811	0.03227	21.17	3.0	3.246
11	0.7268	0.5545	0.6445	0.01062	41.64	3.0	3.113
12	0.7279	0.9459	0.4521	0.03436	15.99	3.0	3.184
13	0.7279	0.7112	0.5999	0.03893	37.24	3.0	3.425
14	0.7279	0.8227	0.3814	0.03031	31.82	3.0	3.073
15	0.7274	0.7308	0.4925	0.02593	45.11	3.0	3.037
16	0.7261	0.5831	0.472	0.01009	29.67	3.0	3.485
17	0.7281	0.6664	0.3614	0.02113	25.88	3.0	3.122
18	0.7274	0.9197	0.3984	0.03303	43.58	3.0	3.312
19	0.728	0.7488	0.5382	0.03545	23.01	3.0	3.087
20	0.7278	0.5947	0.4142	0.02532	48.94	3.0	3.03
21	0.7277	0.8746	0.5218	0.03885	43.38	3.0	3.396
22	0.7265	0.8982	0.2063	0.01092	32.63	3.0	3.201
23	0.7275	0.935	0.484	0.01308	40.11	3.0	3.485
24	0.728	0.7317	0.6708	0.03211	58.2	3.0	3.065
25	0.7285	0.7468	0.3075	0.03969	49.65	3.0	3.108
26	0.7281	0.5518	0.2973	0.0313	46.79	3.0	3.416
27	0.7266	0.9246	0.5942	0.01066	32.69	3.0	3.28
28	0.7279	0.6029	0.6423	0.01629	42.01	3.0	3.171
29	0.7284	0.8624	0.4614	0.0369	8.836	3.0	3.066
30	0.7277	0.7213	0.5559	0.01292	37.21	3.0	3.214

Bayesian Optimization

FINANCIAL IMPACT CALCULATIONS

Claim Expense Cost Reduction = $11 million

- 29 billion annual cost incurred for automobile related fraud
- 15% of automobile insurance claims are trucking related insurance claims
- Northland Insurance has a 1.3% market share in the trucking insurance industry
- Assumes our model would be 20% efficient

$29 bn * 0.15 * 0.013 * 0.2 = $11.0 mn

Increase in revenue = $177 mn

- Northland Insurance has an estimated $231 mn sales (private company does not disclose, but is estimate)
- Northland Insurance has a 1.3% market share in the trucking insurance industry
- Total market for truck insurance is $17 bn
- Assume that premium fees can be reduced by around $500~$1,000 (FBI Data) due to the reduction in costs related to insurance fraud, which allows Northland Insurance to lower price and increase market share by a modest 1% (1.3% --> 2.3%)

($231 mn / 0.013 * 0.023) – $231 mn = $177 mn

Increase in bottom line = 1% OR 28.7 mn

- Traveler's Company has a cost of revenue of 90%
- Traveler's Company 2020 Net Income was $2,697 mn

$11 mn + ($177 mn * 0.1) = $28.7 mn in net income
$28.7 mn / $2,697 mn = 1%

BIBLIOGRAPHY

BIBLIOGRAPHY

- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00380-z
- https://www.kaggle.com/threnjen/2019-airline-delays-and-cancellations
- https://www.travelinc.com/impact-of-flight-disruptions/
- https://digital.hbs.edu/platform-rctom/submission/delta-airlines-flying-high-in-a-competitive-industry/
- https://engineering.berkeley.edu/news/2010/11/flight-delays-cost-more-than-just-time/
- https://www.allthingsontimeperformance.com/flight-delays-in-numbers-not-only-painful-for-passengers/
- https://www.bankrate.com/finance/taxes/taxes-on-insurance-benefits.aspx
- https://www.bankrate.com/insurance/car/auto-insurance-statistics/
- https://www.iii.org/fact-statistic/facts-statistics-auto-insurance
- https://www.insurance-research.org/search/node/Auto%20Insurance%20Fraud
- https://www.verisk.com/insurance/visualize/auto-insurance-premium-leakage-a-29b-problem-for-the-industry/
- https://insurancefraud.org/fraud-stats/
- https://www.nicb.org/
- https://truckinsurancequotes.com/one-of-the-reasons-your-trucking-insurance-is-high-fraud/
- https://www.iii.org/fact-statistic/facts-statistics-insurance-company-rankings
- https://insurance.mo.gov/CompanyAgentSearch/CompanySearch/compSearchDetails.php?id=125906&t=CP&n=NORTHLAND+INSURANCE+COMPANY&c=
- https://www.dnb.com/business-directory/company-profiles.northland_insurance_company.2182475f3924d1c5da0962c14382a098.html