STOR 320: Introduction to Data Science

Spring 2025

Project Proposal Group 6 (Ex: Project Proposal Group 12)

Project Roles

- **Creator:** Branda Sisoutham
- **Interpreter(s):** Riley Little
- **Orator(s):** Yueen Ma, Jacob Dang
- **Deliverer:** Eleni Kafexhiu

Datasets

https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data

- We plan to download this dataset. This dataset contains auto insurance claims data from 2015, including both numerical and categorical variables. We found this dataset on Kaggle and some potential limitations to note would be there are some missing data values and the exact origin of this dataset is unknown. The data also only contains data from policyholders in Ohio, Indiana, and Illinois, and accidents in South Carolina, Virginia, New York, Ohio, West Virginia, North Carolina, and Pennsylvania, so we have to be careful about extrapolating conclusions. The dataset includes 40 total variables. But since the last variable _c39 has all null values, we only consider the first 39 variables that contain data values. We plan to focus on the variables policy_state, incident_state, age, policy_deductible, policy_annual_premium, incident_severity, fraud_reported, insured_education_level, insured_occupation, insured_hobbies, incident_type, incident_severity, incident_hour_of_the_day, bodily_injuries, witnesses, total_claim_amount, vehicle_claim, injury_claim, property_claim, months_as_customer, auto_make, and auto_year. Policy_state is the state where the insurance policy was issued and is a categorical variable. Incident_state is the state where the incident occurred and is a categorical variable. Age is the age of the insured individual and is a numerical variable. Policy_deductible is the amount the policyholder must pay before insurance coverage applies and is a numerical variable. Policy_annual_premium is the total amount of money a policyholder pays for an insurance policy in one year and is a numerical variable. Incident_severity is the level of severity of the reported incident and is a categorical variable. Fraud_reported is whether fraud was reported for the claim and is a binary variable. Insured_education_level is the highest education level attained

by the insured and is a categorical variable. Insured_occupation is the occupation of the policyholder and is a categorical variable. Insured_hobbies indicates the hobbies of the policyholders and is a categorical variable. Incident_type is the type of incident that prompted the claim and is a categorical variable. Incident_severity indicates the damage that occurred in the incident and is a categorical variable. Incident_hour_of_the_day indicates the hour the incident occurred based on a 24-hour clock, and is a numerical variable. Bodily_injuries indicates the number of people harmed in the incident, and is a numerical variable. Witnesses indicates the number of people who saw the incident, and is a numerical variable.Total_claim_amount is the total monetary amount of the insurance claim and is a numerical variable. Vehicle_claim represents the monetary amount requested for damages to the insured vehicle as part of an insurance claim and is a numerical variable. Injury_claim is the portion of the claim related to injury compensation and is a numerical variable. Property_claim represents the amount of money claimed as compensation for damages to the vehicle in the incident and it is a numerical variable. Months_as_customer is how many months the customer has been with the insurance company and it is a numerical variable. Auto_make is the type of make of the car the customer has insured and it is a categorical variable. Auto_year is the year of the vehicle involved in the insurance claim and is a numerical variable.

https://www.fhwa.dot.gov/policyinformation/statistics/2015/mv1.cfm

- We plan to scrape this dataset. This dataset includes state-level motor vehicle registration statistics from 2015. We found this on the Federal Highway Administration website. A possible limitation of this dataset is that it is from 2015 therefore it is older and may not be a good representation of vehicle registration today. However, we chose the year 2015 because our other dataset is from the year 2015 and we thought it would be best to match the years accordingly. Another limitation would be many states do not maintain records on publicly owned motorcycles therefore the total may not represent an accurate count of the total number of publicly-owned motorcycles. The Federal Highway Administration had to estimate those values. Furthermore, Alaska and Massachusetts did report State data but estimations were still made by the Federal Highway Administration. Indiana did not report any state data and the Federal Highway Administration had to estimate all their values.The variables include the number of registered vehicles in each state, categorized by automobiles, buses, motorcycles, and trucks. It also differentiates between private and publicly owned vehicles. We plan to use the variables: total number of registered motorcycles, total number of registered automobiles, total number of registered vehicles which are all numerical variables. We also plan to join the state column of this data set onto either the policy_state or incident_state column of the other data set depending on the question.

Initial Questions

Question 1. How do vehicle types and time of day impact the severity of auto insurance claims? (auto_make, incident_hour_of_the_day, incident_severity)

Question 2. How do policyholder tenure and deductible amount impact the total claim amount? (month_as_customer, policy_deductable, total_claim_amount)

Question 3. How does the time of day of reported incidents differ between states with high vs. low vehicle registration? (25th and 75th percentile) (variables: incident_hour_of_the_day, incident_state, all_motor_vehicles_2015)

Question 4. Do states with a higher percentage of young drivers (16-25) have higher auto insurance claim rates per person? (variables: age, vehicle_claim, policy_state, incident_state)

Question 5. Are policyholders with lower deductibles more likely to be involved in more frequent low-severity incidents compared to those with higher deductibles? (variables: policy_deductable, incident_severity)

Question 6. Are policyholders with lower deductibles more likely to submit a fraudulent claim rather than higher deductibles policyholders? (variables: policy_deductable, fraud_reported)

Question 7. Are older customers willing to pay a higher annual premium, or do they tend to choose lower annual premiums with higher deductibles? (variables: age, policy_annual_premium, policy_deductable)

Question 8. Is there a correlation between the number of publicly-owned automobiles in a state and the frequency or severity of vehicle claims filed by policyholders? (variables: publicly owned automobiles, incident_severity)

Question 9. Is there a relationship between the year an automobile is made and average injury claim amounts? (variables: auto_year, injury_claim)

Question 10. Can create a predictive model to predict if a claim is fraudulent based on the other variables (such as insured_education_level, insured_occupation, insurced_hobbies, incident_type, incident_severity, incident_hour_of_the_day, bodily_injuries, witnesses, total_claim_amount, injury_claim, property_claim, and vehicle_claim)?