**Summary**

The project focuses on three datasets to tackle auto insurance fraud detection, combining traditional claim data, vehicle damage images, and detailed policyholder information. The "Car Parts and Car Damages" dataset contains annotated images of vehicle damages, with polygons marking specific parts and damage types, enabling supervised learning to identify patterns of fraud through visual damage analysis. The "Vehicle Insurance Fraud Detection" dataset includes over 15,000 records with accident details, policyholder demographics, and claim-related information, which can be used for both supervised and unsupervised learning to predict fraud by identifying high-risk features like vehicle type and prior claims history. Lastly, the "Insurance Claims" dataset offers 1,000 records of claims, including information on incident types, vehicle details, and whether fraud was reported, making it well-suited for logistic regression and anomaly detection. Together, these datasets allow for cross-analysis, enabling a more robust fraud detection system that incorporates image classification, demographic profiling, and claim pattern recognition to reduce fraudulent claims, improve processing efficiency, and enhance customer satisfaction.

**Industry of Focus: Auto Insurance**

As the team’s expertise in Data Science matured throughout the course we became keen on selecting Auto Insurance as our industry of focus. Every year, property damage fraud is estimated to cost the shareholders $40 billion in fraudulent claims. This accounts to $400-700 of each policyholder’s premium going into paying out these frauds. (FBI, 2010) This shortfall, although unfortunate, makes the subtopic an appealing choice for both short-term analysis and long-term projects like capstone and thesis.

One of the subtopics chosen within the auto insurance industry is *“fraud detection”* we have found that many of the datasets we had previously chosen for our capstone one project could be used as the primary and supplementary datasets to formulate hypothesis about using these data to detect likeliness of fraudulent claim being filed.

In addition to more traditional fraud detection datasets, we have also decided to add another which offers a more unique and creative approach to the problem, image detection and classification.  Oftentimes, fraudulent claims can be prevented as well as assisted through using image detection to detect where the damage is as well as how severe the damage is; however, image detection on it’s own relating to where the damage on the vehicles is is not enough to determine or make an accurate prediction on fraud, when combining it with the other two datasets, it becomes possible.

Merit of choosing these two subtopics is that they can both add value to each other’s algorithms. For example, one of the possible tools we can build with fraud detection dataset is an AI flagging algorithm that flags the insurance company if the information revolving around the claim suggests that the claim is likely to involve fraud. This can then be supplemented with photo analysis software to detect the severity and type of fraud such as enhanced damage, staged accident, and falsified age of damage to make recommendations to the agent about their decision making process. For this reason we found that all three datasets chosen offered equal significance to the overall goal of our group and did not choose one to be a specific top dataset.

**Analysis of the Top Datasets**

**“Car Parts and Car Damages” (Humans In The Loop, 2023)**

Summary of Data (pandas.describe)               Example PolygonA car with a pink and green paint on it

Description automatically generatedA computer screen with white text

Description automatically generated

As insurance companies move towards a more virtual inspection system (Dell, 2024) This dataset provides a crucial “next step” in not only processing claims faster but more accurately identifying fraud. Supplemented with the other datasets we can draft a comprehensive model that detects fraud in very early stages (by profiling the customer’s demographic and policy information) and at the payout stage when the body shop’s provide photos for insurances to review. More importantly, all three datasets can be used to train each other’s model and dataset - working in unison in formulating a continually improving neural network or decision trees . For example, if  supervised learning on the dataset has found fraudulent damages like rust or self-vandalism, the policy information revolving that claim can be used to train the other datasets by adding another data point that has or has not triggered the “Fraud Found” target data.

Analyzing the “Car Parts and Car Damages” dataset, details of the supervised learning on images are provided. Trainers used classification system consisting of 40 classes and 24851 different polygons each detailed with type of parts/damage , shape of polygon, and color of the polygon (HEX) This data allows for *supervised* learning by providing target data like “Type of Damage” or “Part Damaged” which marks what kind of damage is present on the polygon identified. Allowing us to compare if certain markers have any correlation to this target value, such as, if a black perpendicular polygon suggests a dent or missing part.The data is better suited for supervised learning but opportunities for unsupervised learning exists after the model is fully built out and identifying patterns. We may find patterns like more parts(polygons) being affected on a left side collision than a right side collision. *Clustering* may be performed with one or more of the vehicle information. We may group data by the vehicle type, manufacturer, or age. Dataset is best suited for *anomaly* *detection* as data provides many points of *profiling*. For example, we may be interested in investigating certain damages if a polygon for a certain dent is abnormal for a polygon with a similar vehicle profile. Dataset allows for *regression* analysis with numerical values like “color”. One key data that can be helpful in addition to the current set would be damages caused from non- collision damages like flood or fire as these are common types of fraudulent accidents.

This dataset can be coupled with “Insurance Claims” to add more opportunities for profiling. For example, we can group similar vehicle types and hypothesize questions based on the pivot data for “Insurance Claims”, “Fraud Reported” by seeing which types of damages and parts are most involved in fraudulent claims. We can then make a model based on our previous findings to potentially create an enterprise ready software that can detect fraud in vehicle images.

**“Vehicle Insurance Fraud Detection” (Kapoor, 2023) Dataset**

Summary of Data (pandas.describe)               Names of DataA screenshot of a black and white screen

Description automatically generatedA computer screen shot of white text

Description automatically generated

Sample Data Format

A white background with many small text boxes

Description automatically generated with medium confidence

“Vehicle Insurance Fraud Detection” dataset provides an in-depth and robust data of the automobile insurance policies and if or if not fraud was identified. The dataset totaling at 15,420 records accident details such as the month and day of the week (numeric), vehicle information including make and category(text), and demographics of the policyholders like age, gender, and marital status(boolean/categoric). The dataset also tracks claim-related details such as police report filings, witness presence, and prior claims history, along with derived metrics like the number of days between the policy start and the accident. Combined with the set’s key pivot data, “Fraud Found”, the dataset provides valuable information for trend analysis, predictive modeling, and fraud detection studies.

The car claims dataset, which includes both categorical and numerical data, works well for both *supervised* and *unsupervised* learning. For *supervised* learning, we could use it to predict fraud in claims by treating the Fraud column as the target.   A *decision tree* could help by breaking down the data into clear, interpretable rules to classify claims, while *k-nearest neighbors* could compare new claims to similar past ones to determine their category. For more complex patterns, neural networks can uncover subtle relationships between features, offering deeper insights into what might indicate fraud. Dataset is best suited for *anomaly* *detection* as data provides many points of *profiling*. For example, we may be interested in investigating certain demographics that have abnormally high frequency of fraud. Dataset allows for *regression* analysis with numerical values like “Age of Vehicle”.

On the unsupervised side, *clustering* could uncover natural *groupings*, like claims by car make, vehicle category, or accident area, giving insights without needing a specific target. These clusters could reveal trends, such as which types of claims are most frequent in certain areas or for certain vehicle categories.

This dataset works well in unison with “Insurance Claims” (AQQAD,2023) both datasets have detailed demographic information like policyholders age and marital status which can be used to cluster and profile the same data with similar clusters and profiles. Both dataset then can supplement each other for data that the other lacks. For example, after combined clustering and profiling is performed, “Insurance Claims” can provide how severe the injuries were for people in that profile while “Vehicle Insurance Fraud Detection” can provide data on how the fault was determined for the data in that cluster. Together we can hypothesize questions like “Are customers in a certain age group that are unmarried more or less likely to be involved in fraud, if so are they more or less involved in fraud with injuries claimed or is it just with property damage?” or “Is fraud more likely to happen with cars in year 3(vehicle lease end date), or with very old vehicles (self-caused accidents)?”

The data set lacks some information about individuals, such as education level, salary, etc., but “Insurance Claims” can complement this data set. We can use this dataset to verify whether “Insurance Claims” and the model are accurate or if there is an overfitting situation. Another set of information that will aid in formulating our model would be the claim severity and type of loss (Ex. fire, flood, hit and run). But the set itself did not have any records with missing values. After the model is established and trained, we can use this model to help “Car Parts and Car Damages” and model predict the potential insurance fraud behavior, and then confirm the fraud behavior of users in combination with its model. This dataset can help accurately predict the likelihood of a claim being fraudulent by analyzing factors such as the type of vehicle, accident area, policy details, and claim history. It identifies the features and policy characteristics most strongly associated with frequent claims. Additionally, it highlights the most common accident types and locations where fraudulent claims are filed, providing valuable insights to detect and prevent auto insurance fraud.

**“Insurance Claims” (AQQAD 2023) Dataset**

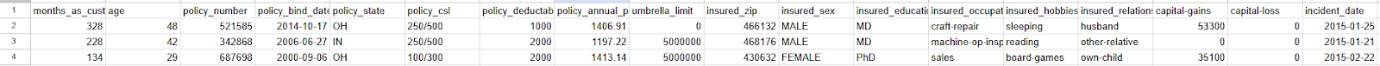
Summary of Data (pandas.describe)               Names of Data

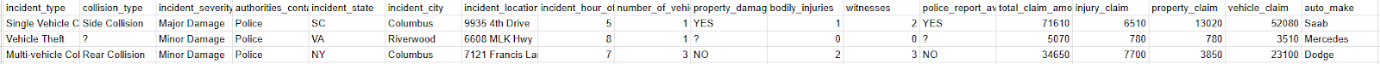
A screenshot of a computer screen

Description automatically generatedA screen shot of a computer screen

Description automatically generated

Sample:





A screenshot of a data sheet

Description automatically generated

“Insurance Claims” dataset provides 1,000 insurance claims, in the format of a csv file, with each row representing a new claim.  The data contains key information regarding the claimant ranging from numeric values, such as the fields of “months\_as\_customer”, “age”, “policy\_number”, as well as “policy\_deductible”, as well as values that are Strings, such as ”insured\_occupation”,”insured\_education\_level”, “insured\_hobbies”, “auto\_make”, and “auto\_model”. In addition to this the dataset contains key information regarding the accident itself, with fields such as “incident\_type”,”collision\_type”,”incident\_severity”, and “vehicle\_claim”. Lastly, the dataset also contains a single character field indicating whether or not fraud was reported, represented by a Y or N value.  Several hypotheses can be made from the wide range of features that are provided through this dataset, such as “Do single car collisions indicate a higher likelihood of fraud?”, “What types of incidents indicate a higher likelihood of fraud?”, as well as “Do certain vehicle makes have a higher likelihood of a fraudulent claim?”.

This data allows for supervised learning due to it having information as to if fraud was reported in the “fraud\_reported” field.  While there is a mix of both numeric and string data, the string data can be converted into numbers through feature engineering, for example, rather than using the “incident\_location” field, the zip codes could be used, and if there was a desire to be more specific, a field could be added indicating weather or not the incident occurred in a rural, suburban, or urban environment.  The types of supervised data that this data is well suited for ranges, but it is best suited for a form of logistic regression, due to the variety of extensive features this dataset has, and predict the probability that a new claim should be classified as fraud or not, this dataset has a large number of entries so a form of holdout or cross validation testing would also work well to verify the models performance and determine which fields have the most impact on the prediction, to save compute when using the model in a business case, which can be over a much larger amount of data. This data is also well suited for unsupervised learning, which can be used to determine if there are any relationships between some of the features such as “authorities\_contacted” and whether or not they could be used to find a relationship with fraud being found.  Lastly, this data would be well suited for a neural network due to the amount of entries of data, as well as how many features there are, a single model might not be able to capture all of the fields correctly, so a neural network created from this dataset may have the best performance, as well as have the least amount of overfitting.

While this dataset stands out due to the large number of fields and good insights would be able to be gained through the data, it fits in well with the other two datasets provided, when paired with “Vehicle Insurance Fraud Detection” (Kapoor, 2023), the data is able to fill in some of the gaps that this dataset has, for example,  the “DriverRating” as well as “Fault” features could be used to provide more valuable insights into the types of drivers and if they are more likely to commit fraud.  In addition, when paired with the “Car Parts and Car Damages” dataset of images (Humans In The Loop, 2023), the fields of “collision\_type” as well as “incident\_secverity” could be paired with the images to predict the probability of fraud based on data from the claim, as well as images of the vehicle after the accident.  A neural network composed of the three datasets would perform particularly well and offer a wide range of ways to predict fraud, ranging from images, locations, as well as information that is available on the claimant.

**Summary**

In this milestone project, our group explored a range of data sets on auto insurance fraud, and ultimately we selected three of the most important data sets to help us build and test our model for our future work. From category and numerical records, the insurance industry can use this data to locate potential fraud. Through the record of image, the insurance industry can train AI to quickly and accurately identify and predict the compensation needed to repair cars, as well as possible insurance fraud. By training models, it can effectively help the insurance industry to reduce labor costs and the possibility of fraud, speed up the processing time of cases, and improve customer satisfaction.

**References**

Aqqad, A. (2023, August 22). *Insurance\_claims*. Mendeley Data. <https://data.mendeley.com/datasets/992mh7dk9y/2>

Dell, E. (2024, June 21). How virtual inspections can benefit insurance customers. Digital Insurance. https://www.dig-in.com/news/how-virtual-inspections-can-benefit-insurance-customers

FBI. (2010, March 17). *Insurance fraud*. FBI. <https://www.fbi.gov/stats-services/publications/insurance-fraud>

Humans In The Loop. (2023, June 9). *Car parts and car damages*. Kaggle. https://www.kaggle.com/datasets/humansintheloop/car-parts-and-car-damages

Kapoor, K. (2023). Vehicle Insurance Fraud Detection. Retrieved from <https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection>

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O’Reilly.