Vehicle Insurance Fraud Analysis

Christine Gendron
Mitchel Diaz
Steven Dookhantie

Our Purpose:

Vehicle Insurance Fraud is an expensive, increasingly common, and avoidable problem for insurers.

*** find a stat **

This is not exclusive to auto insurers. Sophisticated fraud detection processes will be essential to the entire Financial Services Industry moving forward.

That’s why this project seeks to use machine learning techniques to understand and predict fraud.

The Questions:

Can we accurately predict whether an insurance claim is fraudulent?

Which features are correlated with higher likelihood of fraud?

The Data:

Our dataset was generated by **Angoss KnowledgeSEEKER**, a provider of systems for predictive analytics, and found on **Kaggle**.

This set features over 15,000 claim samples with 33 columns, including information on:

- *Demographics of claimant*
- *Details on the vehicle in question*
- *Information about the claimants' policies.*
- *Whether fraud was detected*

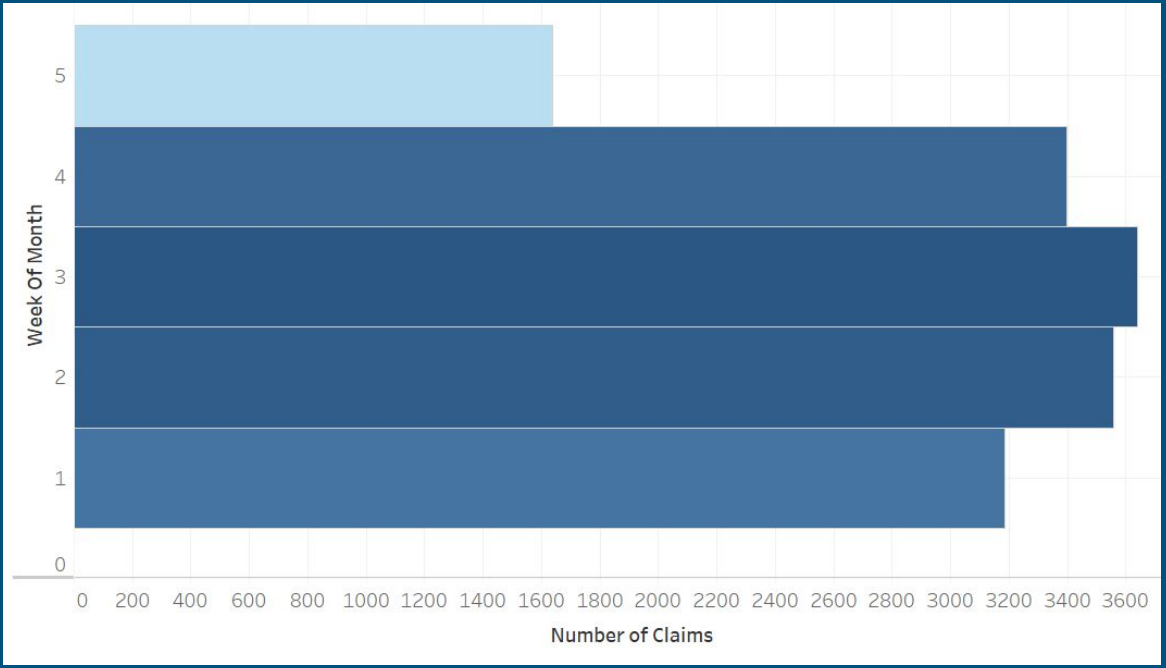
Phase 1: Exploration

Summarizing the data using Python

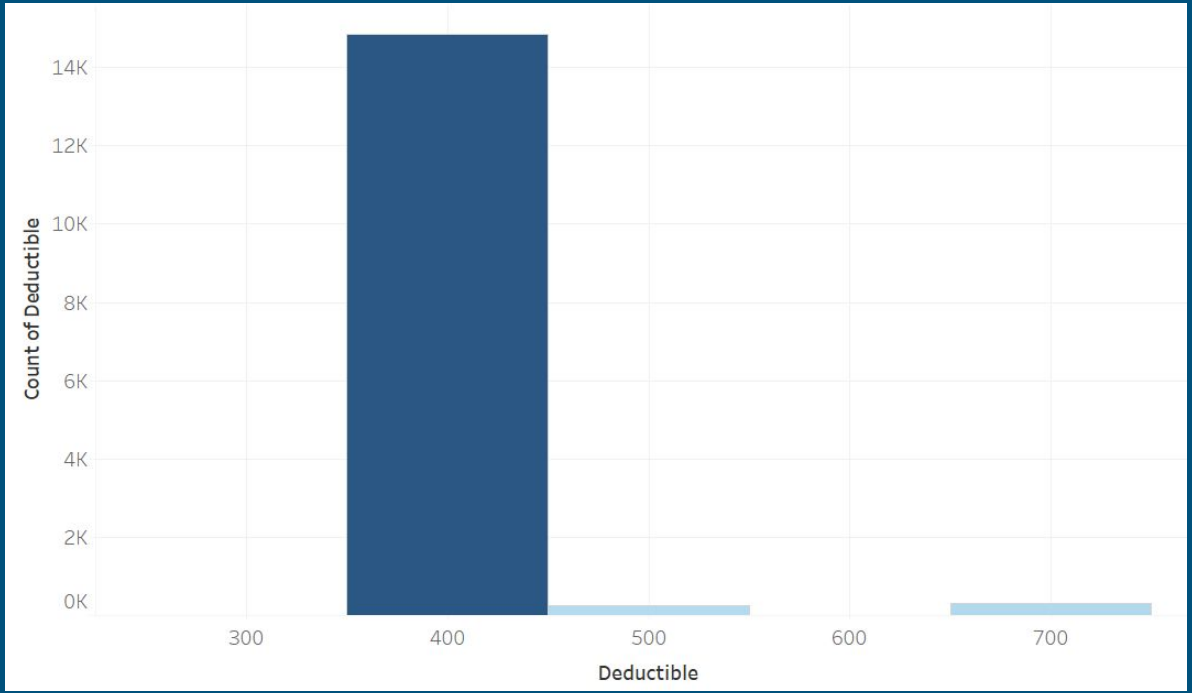
- .describe()
- .dtypes

Dropping Unnecessary Columns

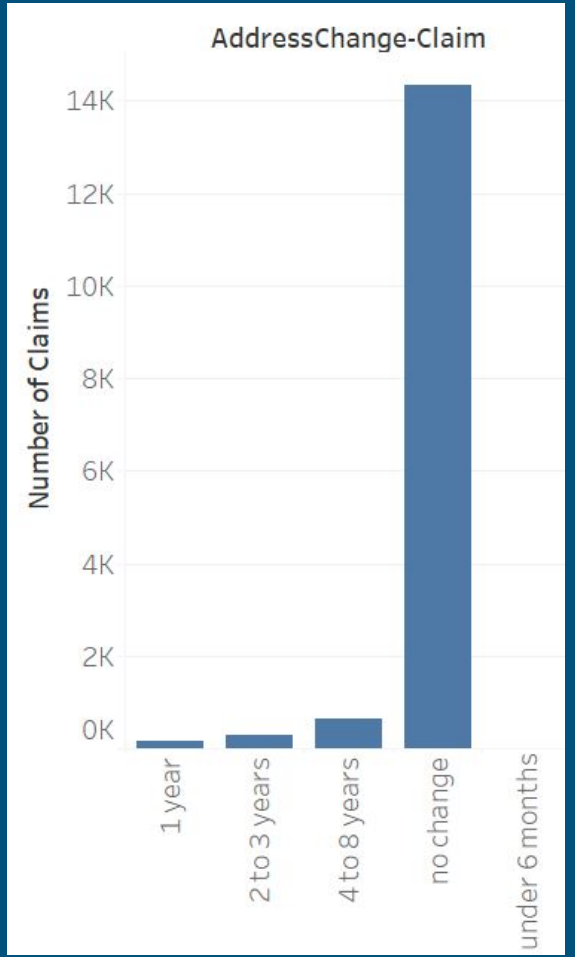
Week Of Month was dropped, since 1-4 were similar, and week 5 of a month always has fewer days than the others.



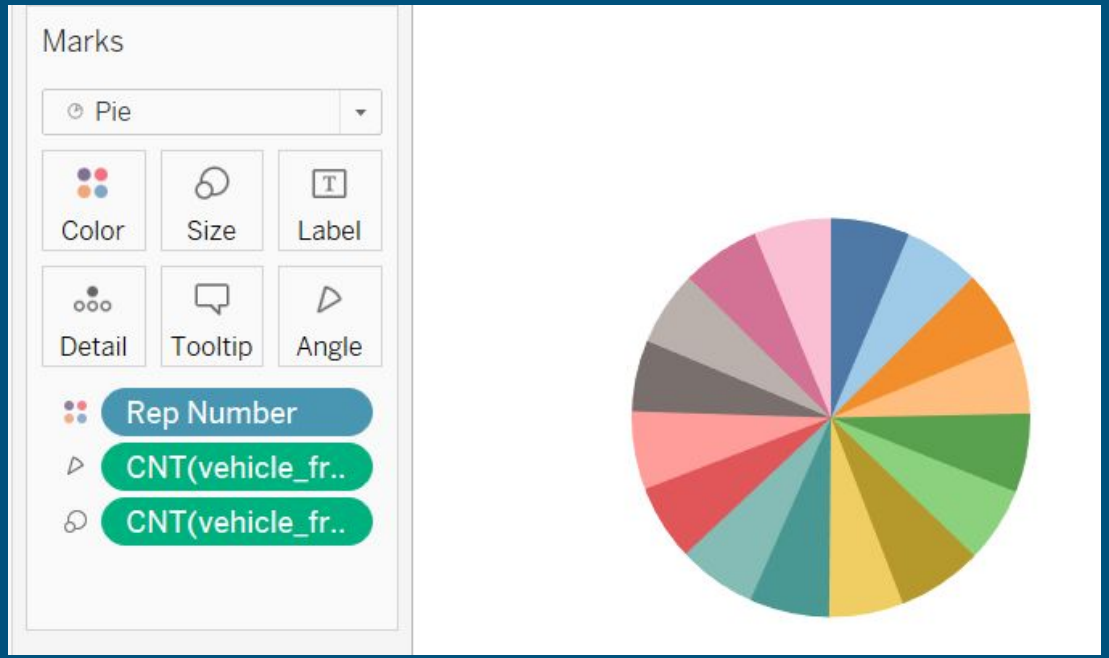
Deductible was dropped, since nearly all claims had a deductible of $400.



AddressChange-Claim was dropped, since most claims fell under “no change”, and the meaning of this column is ambiguous.



RepNumber was dropped, since there was a relatively even spread across rep numbers, and the ID number of the insurance rep handling the claim is not relevant to this analysis.



Phase 2: Analysis

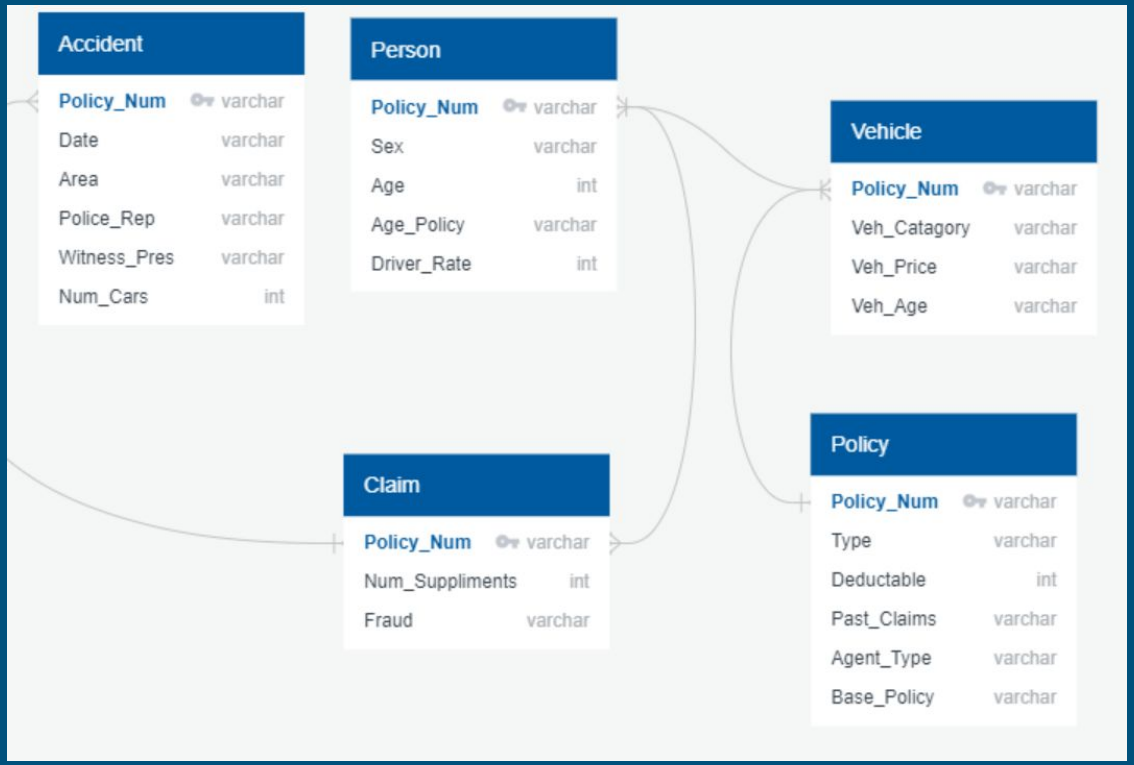
Slides will include:

- Rundown of tech, languages, tools, and algorithms used
- High-level overview of analysis/ml process (real details will be in readme/repo)
- Database stuff? (or can this just be reference in the readme/repo?)
- Visualizations and demo of interactive dashboard illustrating results
- Recommendations for future analysis
- What we would have done differently in retrospect

The Database:

- The data is originally in a csv format
- This analysis uses SQL to to store this static data
- This database is connected to our Machine Learning model via a connection string

Our ERD:


Accident
Policy_Num varchar
Date varchar
Area varchar
Police_Rep varchar
Witness_Pres varchar
Num_Cars int
Person
Policy_Num varchar
Sex varchar
Age int
Age_Policy varchar
Driver_Rate int
Vehicle
Policy_Num varchar
Veh_Catagory varchar
Veh_Price varchar
Veh_Age varchar
Claim
Policy_Num varchar
Num_Suppliments int
Fraud varchar
Policy
Policy_Num varchar
Type varchar
Deductable int
Past_Claims varchar
Agent_Type varchar
Base_Policy varchar

The Dashboard:

We will use Tableau to visualize correlations between various features and whether fraud was detected. Examples:

- Are more fraudulent claims filed by men or women?
- Is fraud associated with particular types of vehicles?
- Are people with certain policies more likely to commit fraud?

ADD LINK TO DASHBOARD WHEN COMPLETE

The Model: Logistic Regression

Our dependent variable (FraudFound) is binary- either fraud was found, or it was not. Logistic Regression is the most efficient way of predicting this type of outcome.

INSERT CONFUSION MATRIX when complete

INSERT ACCURACY SCORE when complete