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.

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

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 Tools:

- Python, including multiple libraries (scikit-learn, pandas)
- The Logistic Regression Algorithm
- SQL
- Tableau
- HTML
- CSS

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*

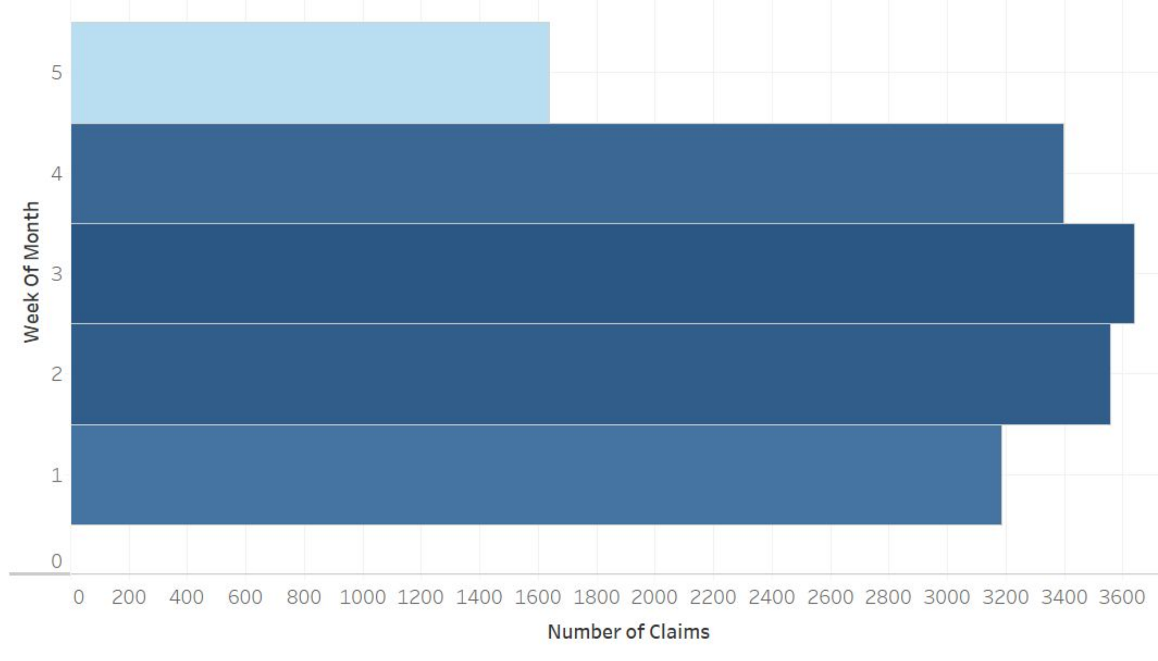
Data Exploration

Summarizing the data using Python

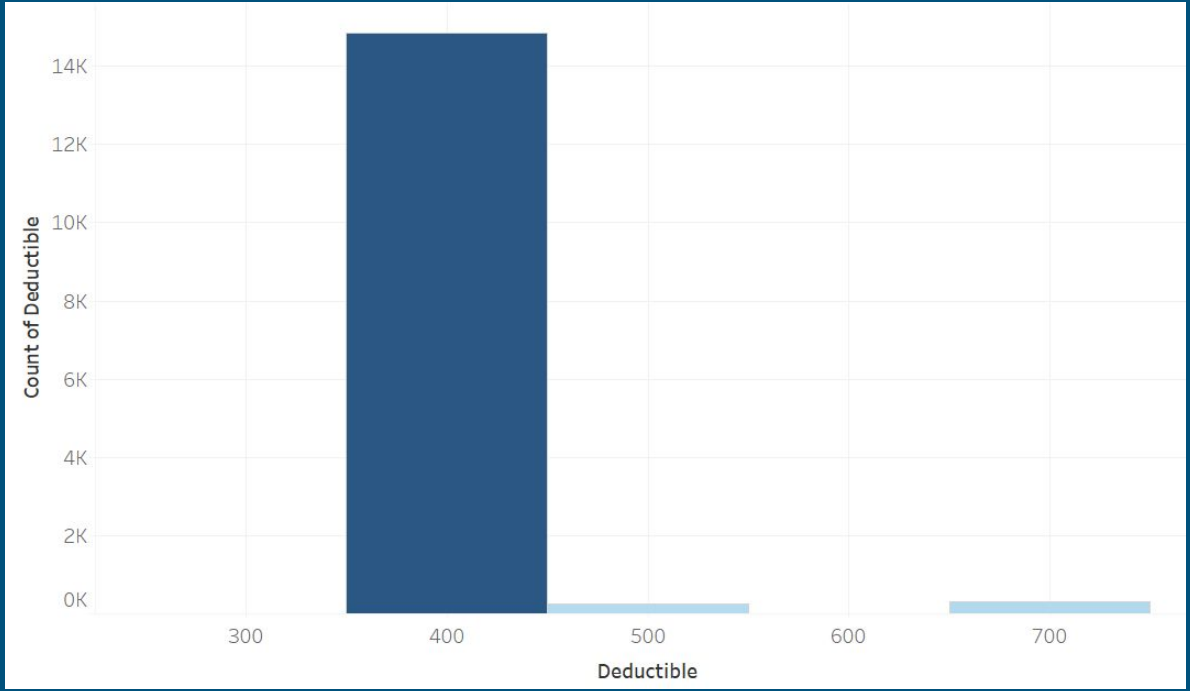
- .describe()
- .dtypes

Dropping Unnecessary Columns

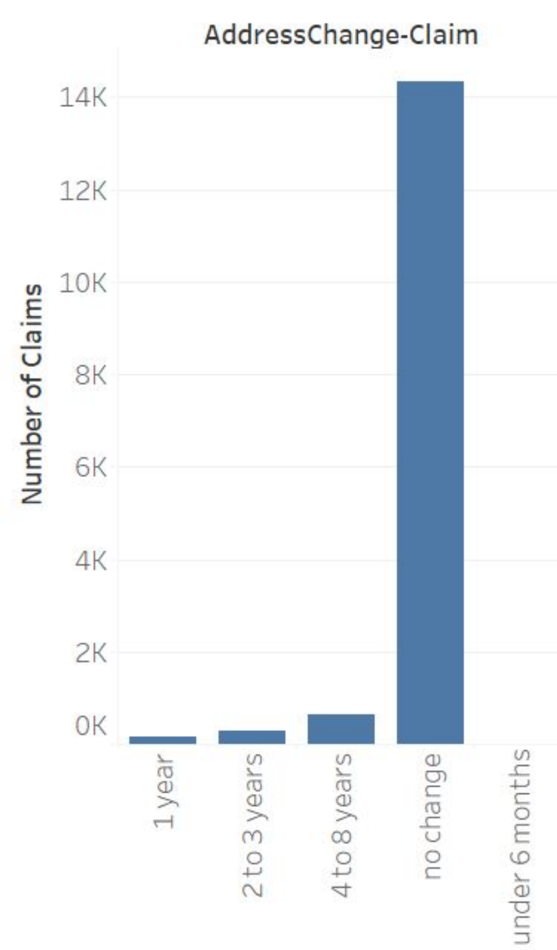
Week Of Month was dropped, since 1-4 were similar, and week 5 of a month always has fewer days than the others.



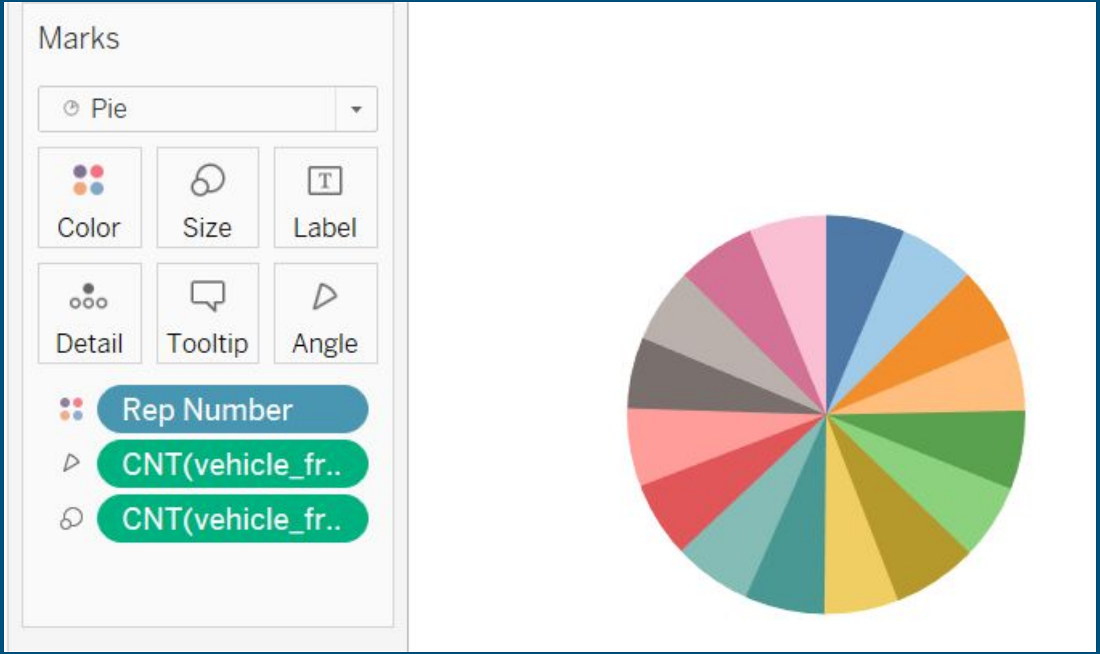
Deductible was dropped, since nearly all claims had a deductible of $400.



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.



Missing Values

We handled missing values using two functions:

1) **missing_value(*dataframe, value*)** identified any missing values

2) **impute_missing(*cat_df, num_df, list_of_cols, dict_of_encoder*)** used KNNImputer to impute those missing values.

Dropping Highly Correlated Features

A For Loop drops numerical features with a correlation of above 0.8, since those features would make the algorithm more complex without adding significant information.

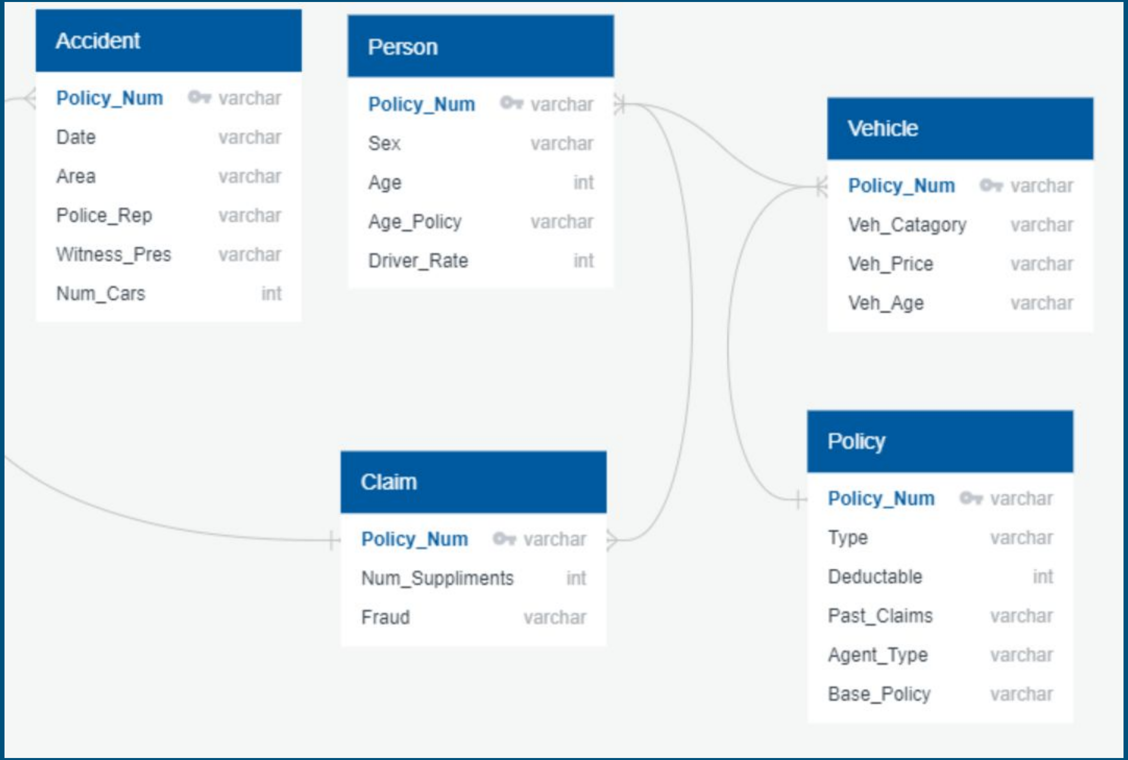
THE ITEM YOU ARE ABOUT TO DROP HAS A HIGH VALUE.

Phase 2: Analysis

The Database:

- The data was originally in a csv format
- This static data is stored in a PostgreSQL database
- 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 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.

Machine Learning Outcomes:

INSERT CONFUSION MATRIX when complete

INSERT ACCURACY SCORE when complete

The Dashboard:

Our dashboard, built using HTML and CSS, features:

- Tableau visualizations
- Images and summary of our Machine Learning Analysis

ADD LINK TO DASHBOARD WHEN COMPLETE

Takeaways:

Future analysis ideas/what we would do differently