## Problem Statement

In the USA, drivers are required to have at least third party insurance. Comprehensive insurance is also available and covers damage to the insured car & injury in addition to third party insurance.

Insurance Companies are concerned with paying out fraudulent claims. The purpose of this project is to quantify the impact of fraudulent claims and devise a decision making model to detect validity of claims and whether more investigations is necessary to prevent fraud.

1. Clean data
2. Quantify or analyse impact and frequency of fraudulent claims.
3. Design and deploy a model that can detect claims that are likely to be fraudulent and therefore require further validations before paying out.

## Clean data:

Missing Data & summaries in excel and Power BI

*I used Pivot tables and PowerBI dashboards to impute missing data. See comments in excel for the way data was cleaned. Another way to clean up data could be implemented in python code.*

There are 22 missing values in the overall data.

21 rows, each missing only 1 value and 1 row missing 2 inputs (policy\_annual\_premium & incident\_state)

10 columns are missing data in the range of 1 to 4 values.

Conclusion: Given on 1000 observations / rows and 39 columns, having only 22 missing values spread across variables and row labels, we can impute this data rather than removing it altogether.

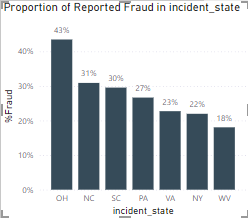
178 collision types were fill in as “?”, these were replaced by “Unknown”.

360 property claim “?”. Replace with Yes if there is property claim value, with No if property claim value is zero.

343 records have “?” on policy report. Made these “Unknown”

## Data Analysis via Power BI and Python

OH as an incident state have higher proportion of fraudulent claims, at 43% compared to average of 25%. Females make more fraudulent cases (67%) than males here although data is very limited to make such inference.



A table with numbers and percentages

Description automatically generated

Fraudulent claims tend to be reported to other authorities than policy. There is surprisingly low proportion of fraudulent claims in the category of claim that are not reported to any authorities.

A graph of a graph

Description automatically generated with medium confidence

Overall, gender and education level do not have material impact on likelihood of claims being fraudulent or not.

## Fraud Detection Model

A fraud detection model was developed in python. The aim of the model is to be able to classify cases that should be investigated for fraud prior to paying out in order to minimise claim payout on invalid claims. This has such long-term implications in improving profitability of insurers and minimising fraud overall.

Logistic regression, Decision tree and Random forest were fitted. Random forest performed better than the other two and would therefore be selected model.

The data of only 1000 observations is too low for this class of insurance, the results would need to be validated in a larger dataset where accuracy may be much higher than the 0.77% achieved in this model.