Detecting Fraud in Auto Insurance Claims

Analysis by: Filip Gvardijan

Use Case

Insurance companies handle lots of claims, some of which are fraudulent. Reducin Detecting fraudulent claims before payment would enable company to act and by doing so, company could improve efficiency and reduce cost.

Scientific approach can help the insurance company to:

1. **Explore influential factors** that correlate with fraudulent claims.
2. **Predict fraudulent claims** in automated way using Machine Learning algorithm.

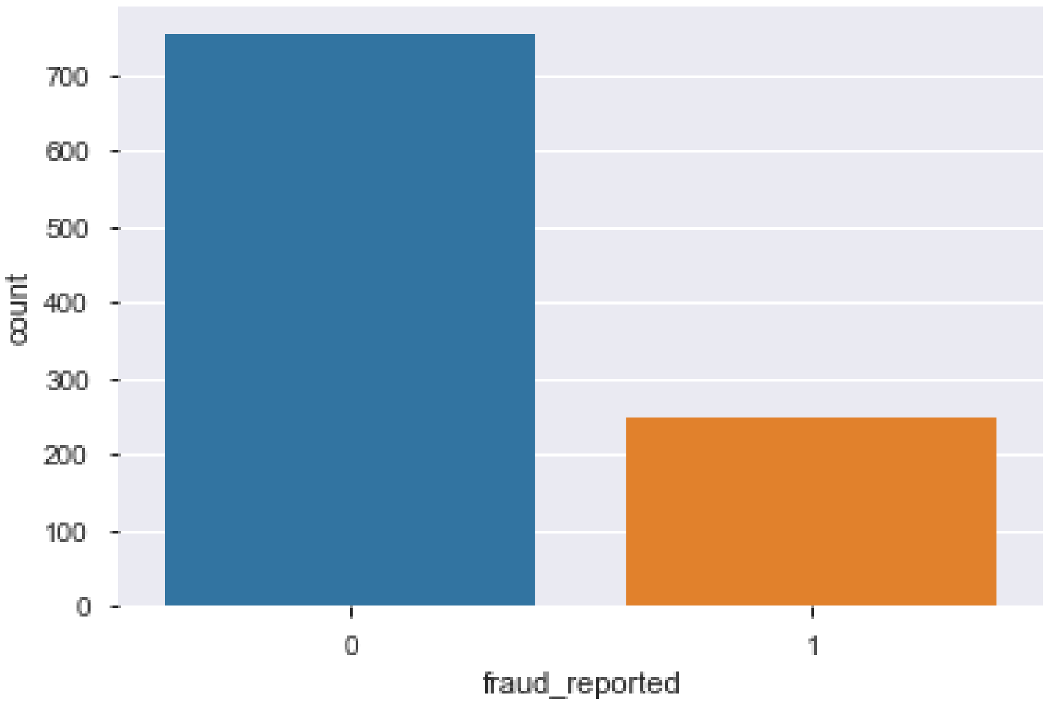
Dataset

Dataset is on car insurance claims. Dataset provides details about customer, insurance policy, incident and cost.

- Q1 2015
- 1000 incidents
- 1 record per incident
- 40 features

Fraud percentage is 24.7%

Historical data shows that 24.7% of incidents are frauds.



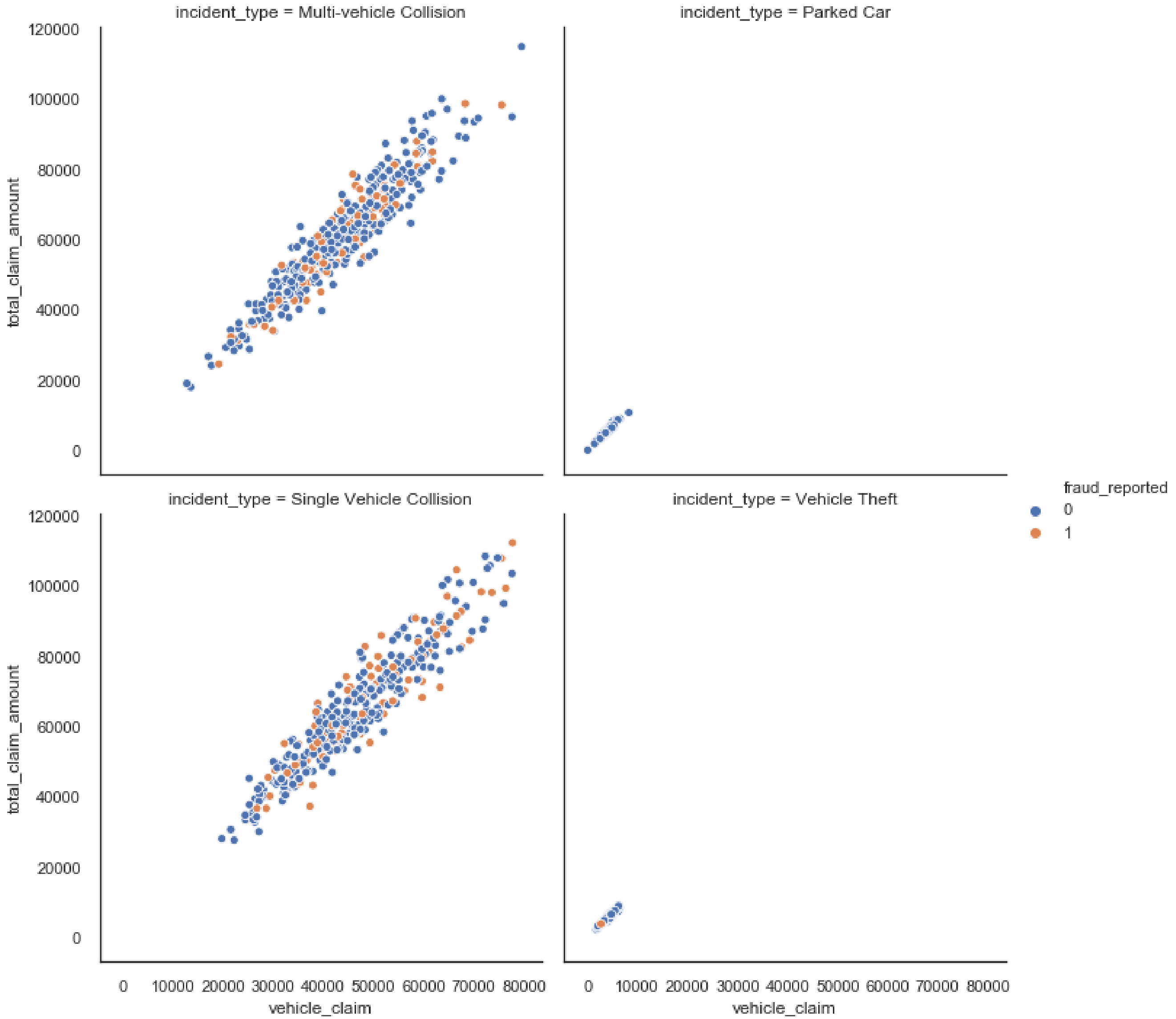
753 247

Collisions fraud is more frequent and costs are high

- Incidents in traffic make majority of cases

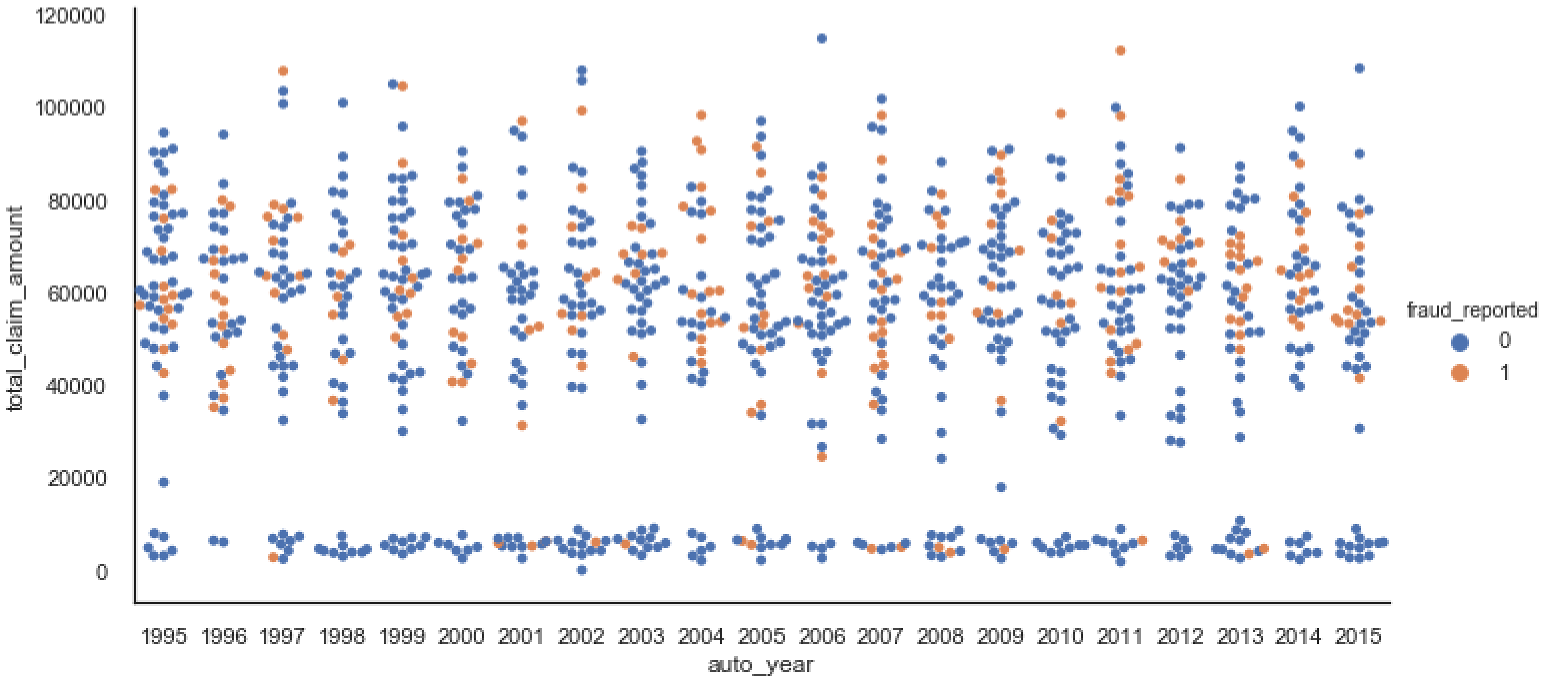
- Vehicle claim amount is major contributor to total claim amount

- Fraud is evenly represented in single and multi-vehicle incidents



Car age does not matter

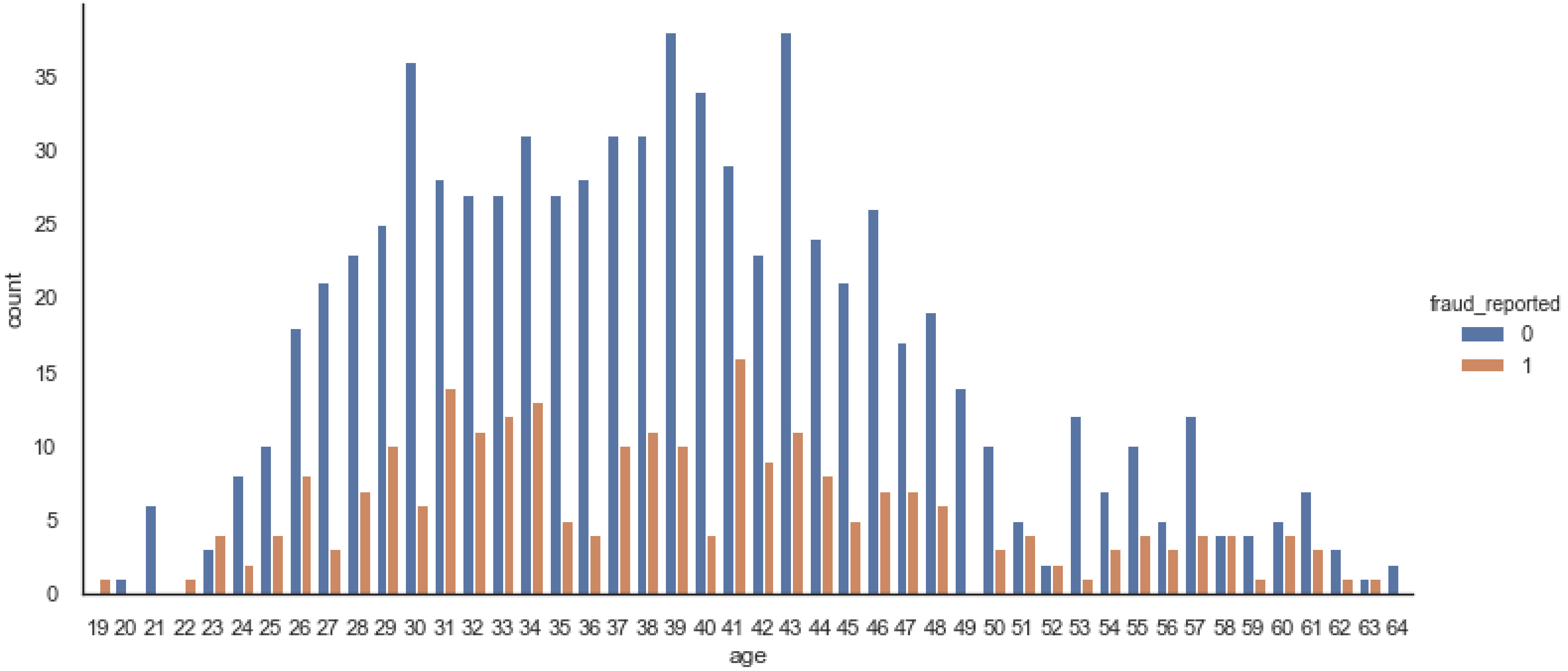
- Fraud is committed across the range of auto years



Customer age does matter

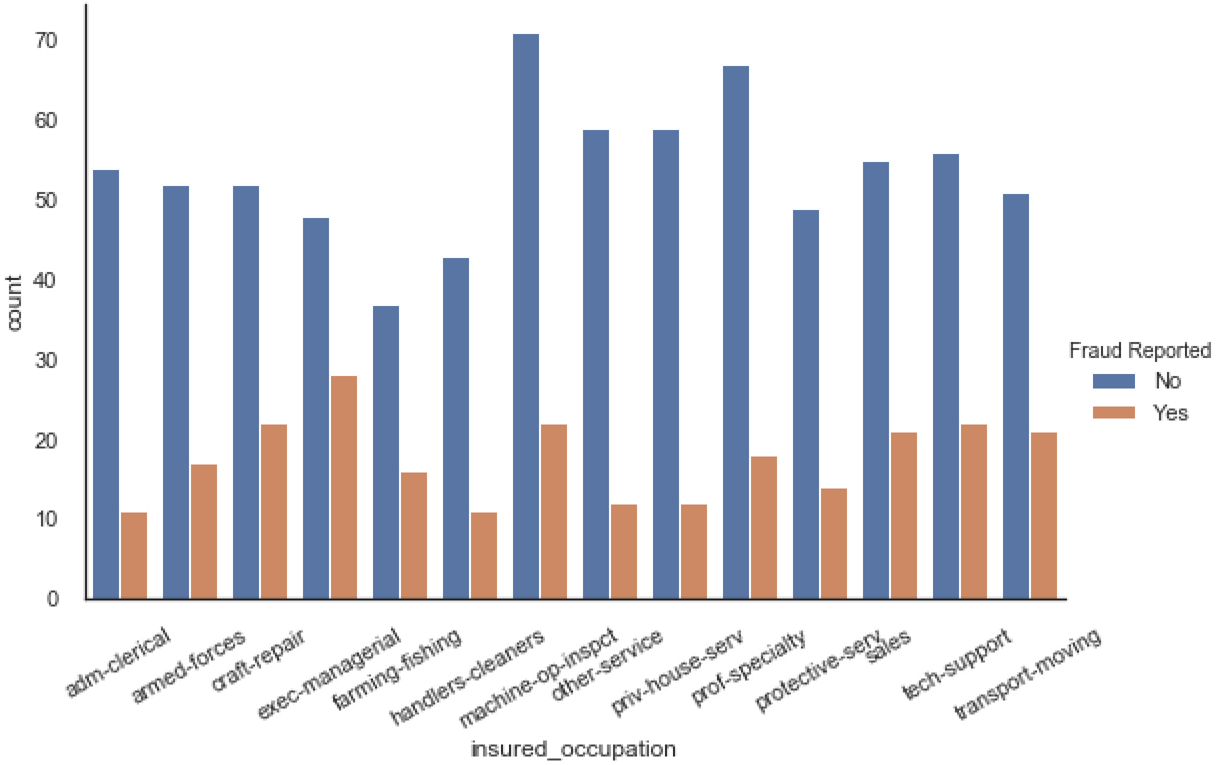
- Some age groups committed relatively more fraud
  - Under 24
  - Above 55

- Others are more or less equally likely to commit fraud



Managers seem more inclined to fraud ?

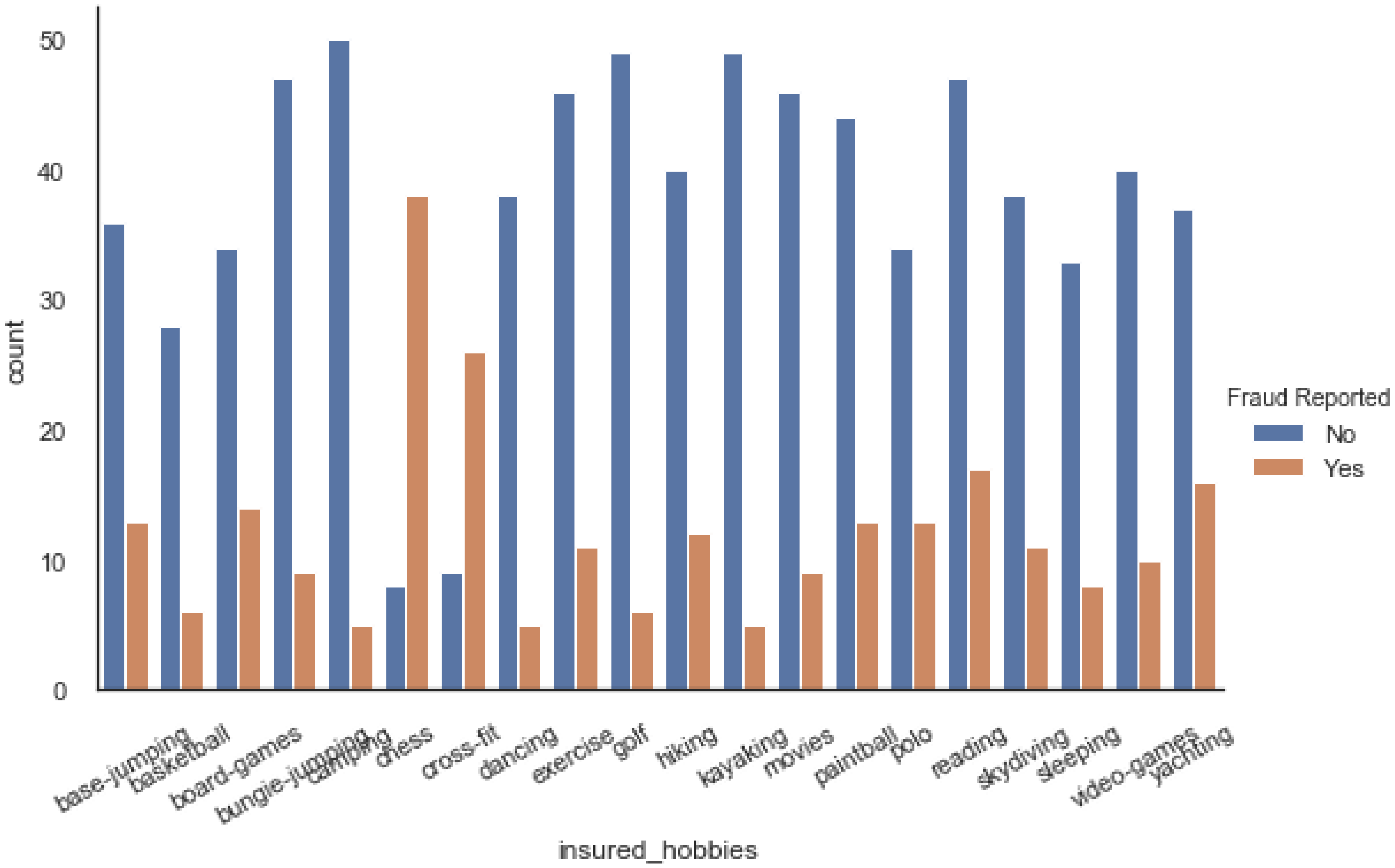
- Incidents involving people who are **exec-managerial** are **37% fraud**



Some hobbies stand out…

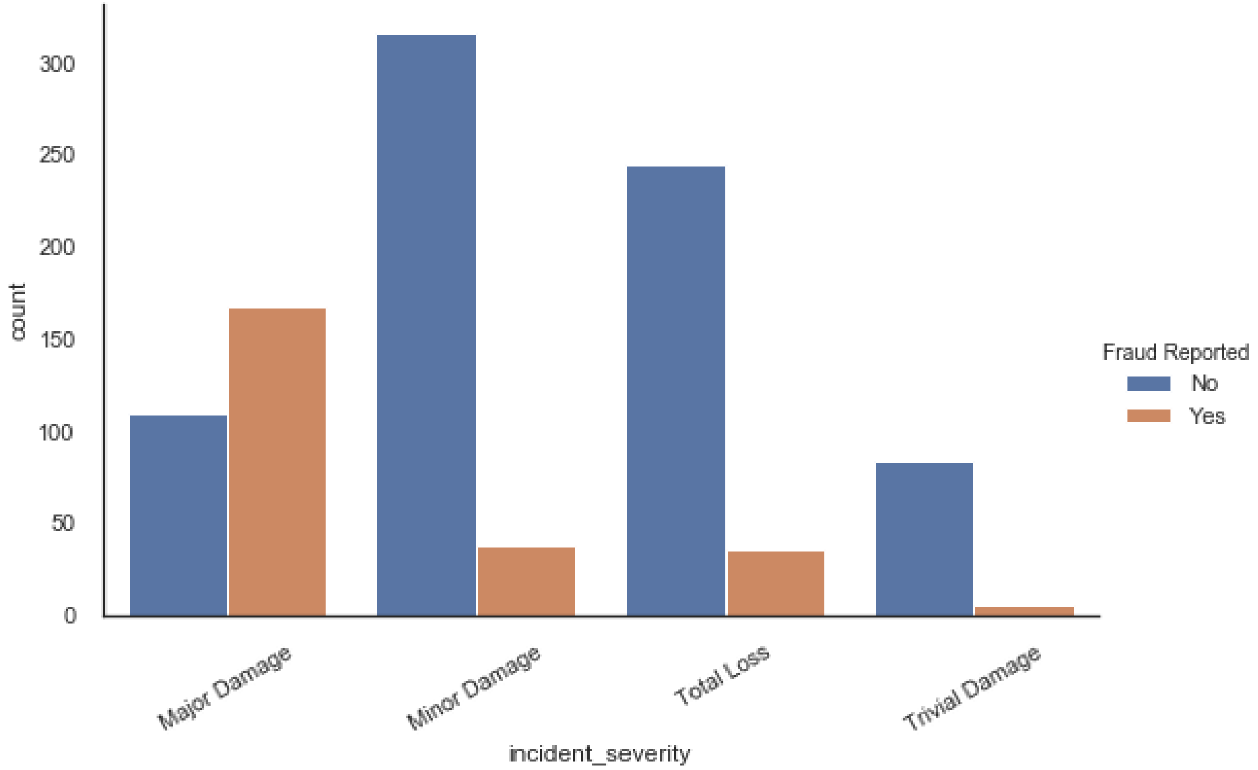
- Incidents involving people who play **chess** are **83% fraud**

- Incidents involving people who do **cross-fit** are **74% fraud**



Major incidents need special attention

- Incidents with Major Damage are fraud in **61% of the cases**

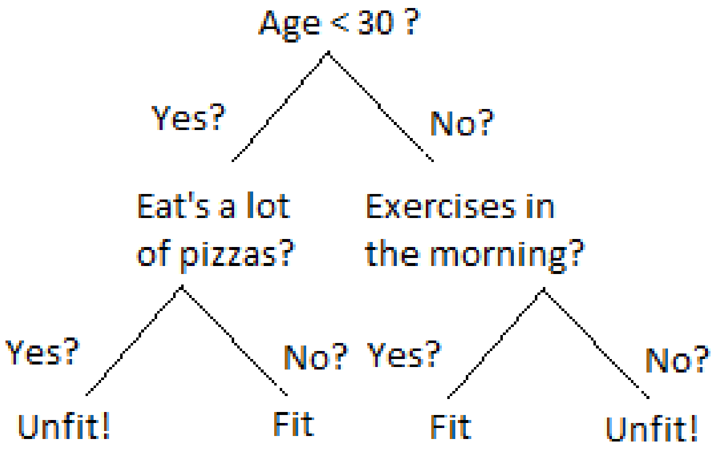


Using Machine Learning to predict Fraud

- Fraud prediction is binary classification type of problem.
- Algorithms based on decision trees perform very well on this type of problem.

✓ ✗

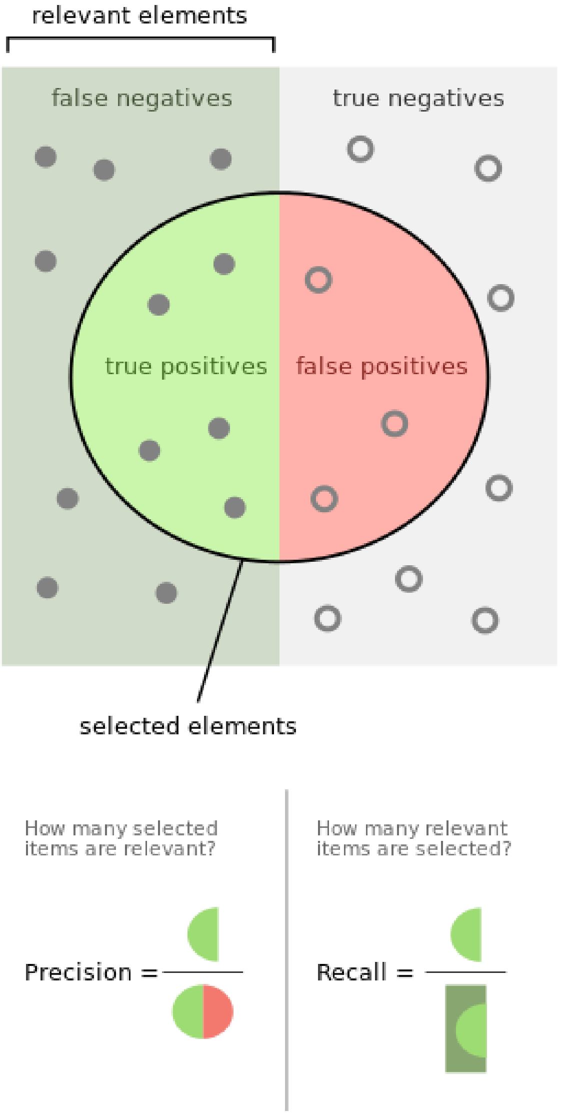
Is a Person Fit?



Evaluating Performance

We are dealing with binary classification and an unbalanced dataset, so the chosen evaluation metric is **f1-score**.

- f1-score considers both the *precision p* and the *recall r* of the test to compute the score.
- This makes it a very robust measure in which false positives and false negatives are penalized.



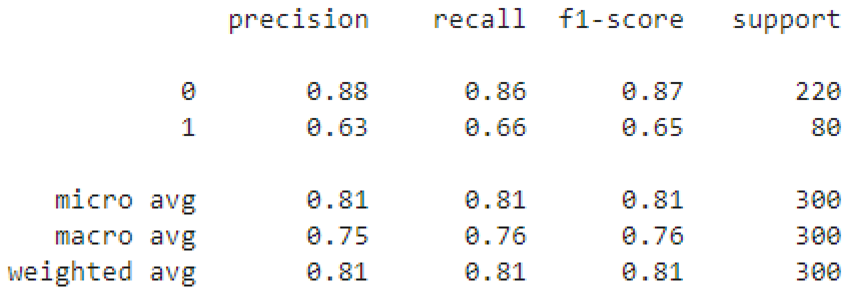
Gradient Boosted Tree Model

Best performing model is **LightGBM** with **f1-score 0.65** for detecting fraud.

- **53 / 80** fraud cases predicted correctly
- 27 / 300 false negatives (not fraud, predicted fraud)
- 31 / 300 false positives (fraud, predicted not fraud)

Model is trained with clean dataset with most influential features extracted.

Classification report

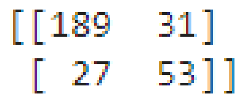


```
              precision    recall  f1-score   support

           0       0.88      0.86      0.87       220
           1       0.63      0.66      0.65        80

   micro avg       0.81      0.81      0.81       300
   macro avg       0.75      0.76      0.76       300
weighted avg       0.81      0.81      0.81       300
```

Confusion matrix:

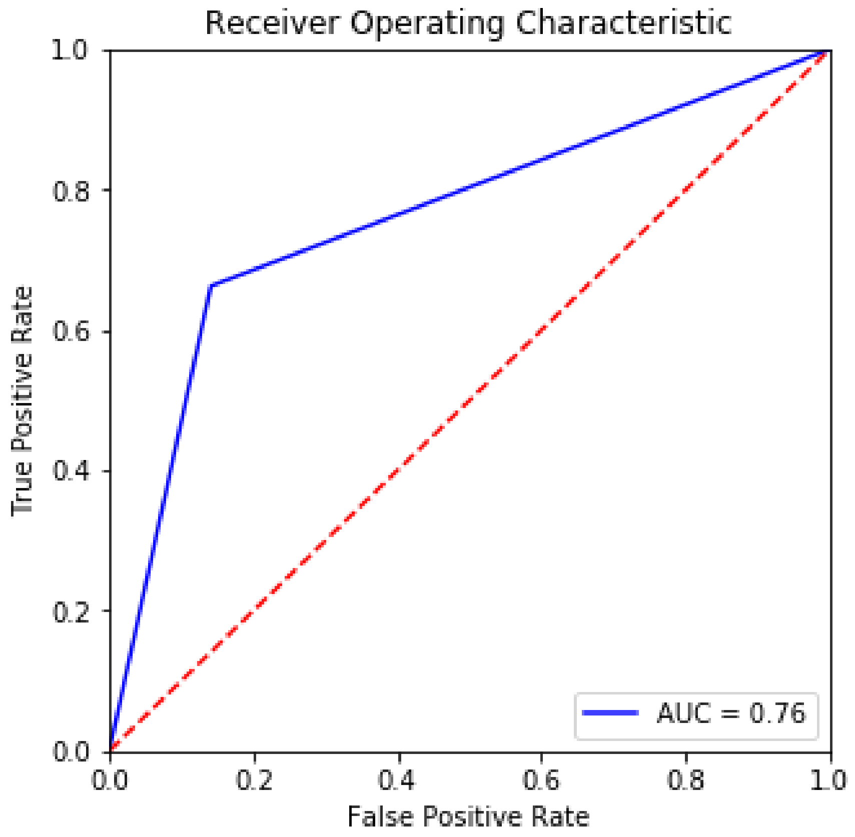


```
[[189  31]
 [ 27  53]]
```

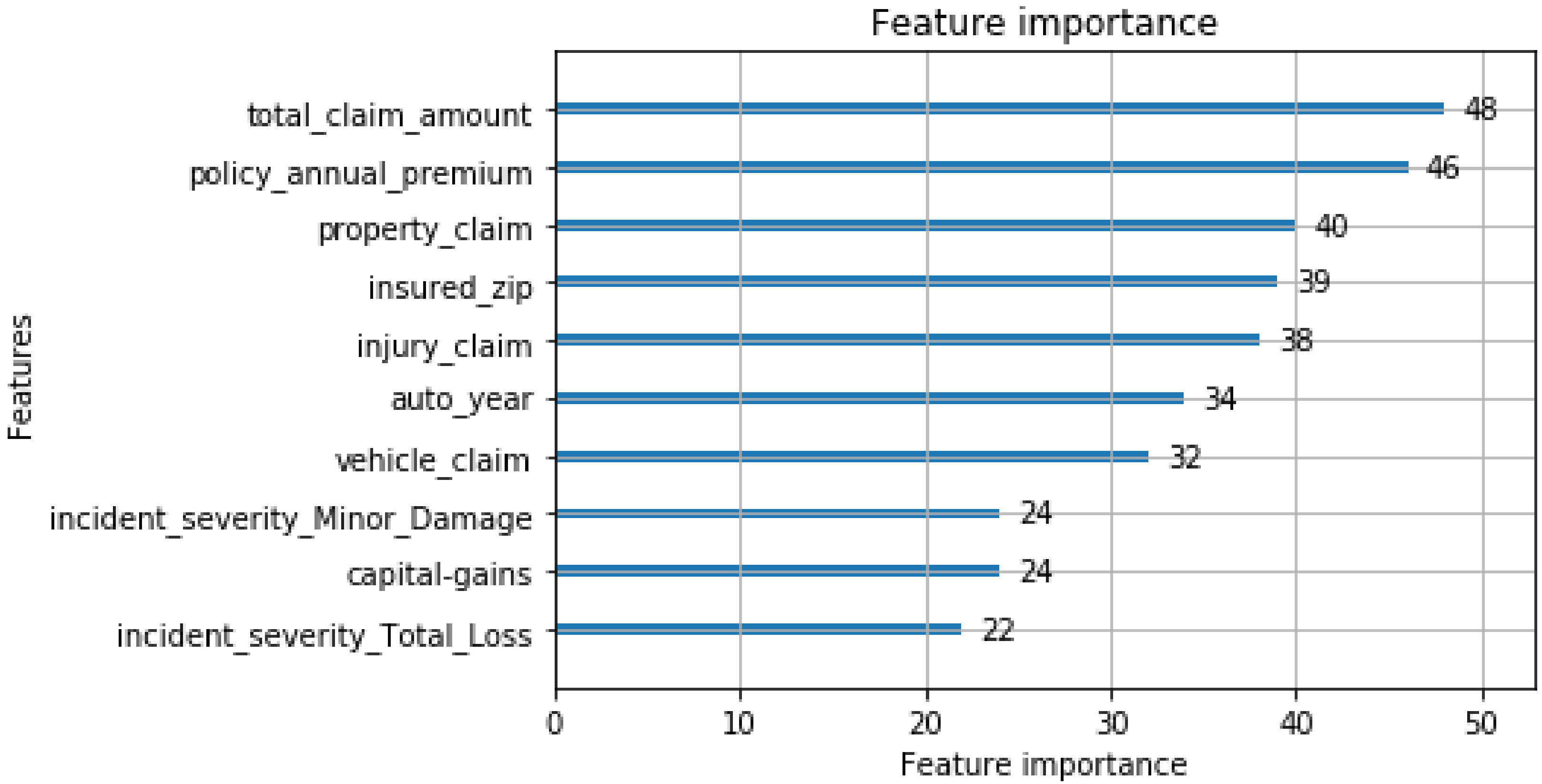
Performance curve

Area under the curve for our model is **0.76**.

Meaning that our model gives *76% probability* that a randomly chosen *positive instance (fraud) is ranked higher* than a randomly chosen *negative instance* (not fraud).



Most important features in the model



Using the model

Trained model can be used when new insurance claim is received.

1. Raw data is collected and stored.
2. Script for data cleaning and transformation is run on raw data.
3. Prepared data is fed into the pre-trained model.
4. Model returns a probability that claim is fraudulent.
5. This information can be used to adjust further steps in processing the claim.

Thank you