

# **Car Insurance Claim Prediction**

Jingjing Yang

5G00FT12-3003 AI and Machine Learning Practical work 1

13th October 2023

# **Description of the work**

The dataset used in this project is the <u>Car Insurance Claim Prediction dataset</u> downloaded from Kaggle. The dataset contains 58,592 rows and does not have any missing values or duplicate records. It provides various columns related to insurance policies and car details.

This work aims to analyze the car insurance claim prediction dataset and develop a predictive model to determine whether a policyholder will file a claim in the next 6 months or not based on various factors.

# **Data preparation for the training**

## **Summary of columns (partially)**

#### 1. Policy Information:

- policy\_id: Unique identifier of the policyholder.
- policy\_tenure: The duration of the policy, measured in hours, representing the elapsed time of the policy holder over a full year.

#### 2. Car Information:

- age\_of\_car: Normalized age of the car in years.
- age\_of\_policyholder: Normalized age of the policyholder in years.
- area\_cluster: Cluster or category of the area (C1-C22).
- population\_density: Population density of the area.
- make, segment (A/B1/B2/C1/C2/Utility), model (M1-M11)
- fuel\_type: Type of fuel used by the car (Petrol/Diesel/CNG).

### 3. Technical Specifications:

- max\_torque: Maximum torque produced by the engine(Nm@4400rpm),9 unique values.
- max\_power: Maximum power produced by the engine(bhp@6000rpm),9 unique values.

Various technical specifications such as engine\_type,
rear\_brakes\_type (Drum/Disc), engine displacement, number of cylinder (3/4),
transmission\_type, number of gear\_box (5/6), steering\_type, space of turning
\_radius in meters, etc.

#### 4. Car Dimensions:

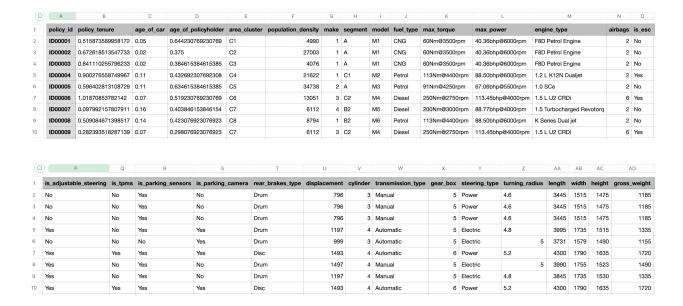
- length, width, height: Dimensions of the car in millimetre.
- gross\_weight: Maximum allowable weight of the fully-loaded car.

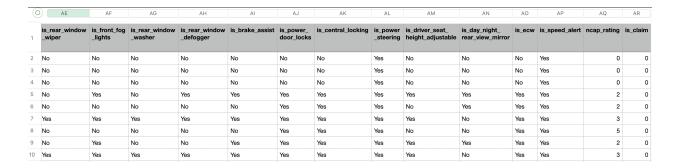
#### 5. Safety Features:

Various columns indicating features like number of airbags (1/2/6), the presence of esc (Electronic Stability Control), adjustable steering wheel, tpms (Tyre Pressure Monitoring System), parking sensors, parking camera, fog lights, brake assist, power door look, central locking, power steering, ecw (Engine Check Warning), speed alert, etc.

#### 6. Insurance and Claims:

- ncap\_rating: Safety rating by NCAP (New Car Assessment Program), out of 5.
- is claim: Indicates whether a claim has been made in the next 6 months.





## data preprocessing

• Encoding all the boolean data into numerical values to fit machine learning models

```
['is_esc', 'is_adjustable_steering', 'is_tpms', 'is_parking_sensors', 'is_parking_camera', 'is_front_fog_lights', 'is_rear_window_wiper', 'is_rear_window_washer', 'is_rear_window_de fogger', 'is_brake_assist', 'is_power_door_locks', 'is_central_locking', 'is_power_steering', 'is_driver_seat_height_adjustable', 'is_day_night_rear_view_mirror', 'is_ecw', 'is_speed_alert']
```

· Convert categorical variables into numerical representations using dummy encoding

```
['area_cluster', 'segment', 'model', 'fuel_type', 'max_torque', 'max_power', 'engine_type', 'rear_brakes_type', 'transmission_type', 'steering_type']
```

Oversampling using SMOTE (Synthetic Minority Over-sampling Technique)

The majority class (54844 No claim) significantly outnumbers the minority class (3748 Claim), so oversampling is performed using SMOTE with a specific ratio to avoid either oversampling excessively or a classifier predicts only the majority class.

Standardization

To scale numerical features and bring them to a similar scale, method **standardscaler** is used.

# Relevant metrics for the cases

Metric	Description
Confusion matrix	A table that summarizes the performance of a classification model by showing the counts of true positive, true negative, false positive, and false negative predictions.
Accuracy	The ratio of correct predictions to the total number of predictions made by the model.
Precision	The ratio of true positive predictions to the total number of positive predictions made by the model.
Recall	The ratio of true positive predictions to the total number of actual positive instances.

Model: LogisticRegression Confusion Matrix: [[8580 2411] [5214 2991]] Accuracy: 0.6027818295478224 Precision: 0.5536838208071084 Recall: 0.36453382084095065 Model: DecisionTreeClassifier Confusion Matrix: [[5446 5545] [1502 6703]] Accuracy: 0.6328922692227548 Precision: 0.547273024167211 Recall: 0.8169408897014016 Model: RandomForestClassifier Confusion Matrix: [[10100 8911 [ 1072 7133]] Accuracy: 0.8977391123150656 Precision: 0.8889581256231306 Recall: 0.8693479585618525

### Conclusions of the results

The models were validated using a test dataset that was not seen during training, providing an unbiased evaluation of their performance on new, unseen data.

**Logistic Regression** model shows moderate overall performance but struggles to correctly identify positive cases (low recall).

**Decision Tree Classifier** model has a better recall, but precision is compromised.

Random Forest Classifier model has high accuracy and balanced precision and recall.

The results indicate that the Random Forest Classifier model outperforms others, and is potentially usable in real-world applications for predicting insurance claims.. This effectiveness may be attributed to the dataset's high-dimensional nature and complex relationships.

#### **Room for Improvement:**

- Exploring additional relevant features or creating new features (such as torque/rpm ratio, power/rpm ratio) could improve model performance.
- Fine-tuning the hyperparameters of the models through grid search or randomized search might yield even better results.
- Explore other advanced models or ensemble techniques.