Tampere University of Applied Sciences

Car Insurance Claim Prediction

Jingjing Yang

5G00FT12-3003 AI and Machine Learning
Practical work 1

13th October 2023

Used algorithms: Logistic Regression, Decision Tree, Random Forest

Description of the work

The dataset used in this project is the Car Insurance Claim Prediction dataset 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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	policy_id	policy_tenure	age_of_car	age_of_policyholder	area_cluster	population_density	make	segment	model	fuel_type	max_torque	max_power	engine_type	airbags	is_esc
2	ID00001	0.515873589958172	0.05	0.644230769230769	C1	4990	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	No
3	ID00002	0.672618513547733	0.02	0.375	C2	27003	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	No
4	ID00003	0.841110255796233	0.02	0.384615384615385	C3	4076	1	A	M1	CNG	60Nm@3500rpm	40.36bhp@6000rpm	F8D Petrol Engine	2	No
5	ID00004	0.900276558749967	0.11	0.432692307692308	C4	21622	1	C1	M2	Petrol	113Nm@4400rpm	88.50bhp@6000rpm	1.2 L K12N Dualjet	2	Yes
6	ID00005	0.596402813108729	0.11	0.634615384615385	C5	34738	2	A	M3	Petrol	91Nm@4250rpm	67.06bhp@5500rpm	1.0 SCe	2	No
7	ID00006	1.01870853782142	0.07	0.519230769230769	C6	13051	3	C2	M4	Diesel	250Nm@2750rpm	113.45bhp@4000rpm	1.5 L U2 CRDi	6	Yes
8	ID00007	0.097992157807911	0.16	0.403846153846154	C7	6112	4	B2	M5	Diesel	200Nm@3000rpm	88.77bhp@4000rpm	1.5 Turbocharged Revotorq	2	No
9	ID00008	0.509084671398517	0.14	0.423076923076923	C8	8794	1	B2	M6	Petrol	113Nm@4400rpm	88.50bhp@6000rpm	K Series Dual jet	2	No
10	ID00009	0.282393518287139	0.07	0.298076923076923	C7	6112	3	C2	M4	Diesel	250Nm@2750rpm	113.45bhp@4000rpm	1.5 L U2 CRDi	6	Yes

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	is_adjustable_steering	is_tpms	is_parking_sensors	is_parking_camera	rear_brakes_type	displacement	cylinder	transmission_type	gear_box	steering_type	turning_radius	length	width	height	gross_weight
2	No	No	Yes	No	Drum	796	3	Manual	5	Power	4.6	3445	1515	1475	1185
3	No	No	Yes	No	Drum	796	3	Manual	5	Power	4.6	3445	1515	1475	1185
4	No	No	Yes	No	Drum	796	3	Manual	5	Power	4.6	3445	1515	1475	1185
5	Yes	No	Yes	Yes	Drum	1197	4	Automatic	5	Electric	4.8	3995	1735	1515	1335
6	No	No	No	Yes	Drum	999	3	Automatic	5	Electric	5	3731	1579	1490	1155
7	Yes	Yes	Yes	Yes	Disc	1493	4	Automatic	6	Power	5.2	4300	1790	1635	1720
8	Yes	No	Yes	No	Drum	1497	4	Manual	5	Electric	5	3990	1755	1523	1490
9	Yes	No	Yes	No	Drum	1197	4	Manual	5	Electric	4.8	3845	1735	1530	1335
10	Yes	Yes	Yes	Yes	Disc	1493	4	Automatic	6	Power	5.2	4300	1790	1635	1720

	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR
1	is_rear_window_wiper	is_front_fog_lights	is_rear_window_washer	is_rear_window_defogger	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	ncap_rating	is_claim
2	No	No	No	No	No	No	No	Yes	No	No	No	Yes	0	0
3	No	No	No	No	No	No	No	Yes	No	No	No	Yes	0	0
4	No	No	No	No	No	No	No	Yes	No	No	No	Yes	0	0
5	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2	0
6	No	No	No	No	No	Yes	Yes	Yes	No	Yes	Yes	Yes	2	0
7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	3	0
8	No	No	No	No	No	Yes	Yes	Yes	No	No	Yes	Yes	5	0
9	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	2	0
10	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	3	0

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_defogger', '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   891]
 [ 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.