Vehicle Insurance Prediction

Problem statement

The Dataset contains information on policyholders having the attributes like policy tenure, age of the car, age of the car owner, the population density of the city, make and model of the car, power, engine type, etc. You need to develop a machine learning model for identifying the potential policyholders. Potential policyholders are those who may opt for insurance services in future.

Variable	Description
policy_id	Unique identifier of the policyholder
policy_tenure	Time period of the policy
age_of_car	Normalized age of the car in years
age_of_policyholder	Normalized age of policyholder in years
area_cluster	Area cluster of the policyholder
population density	Population density of the city (Policyholder City)
make	Encoded Manufacturer/company of the car
segment	Segment of the car (A/ B1/ B2/ C1/ C2)
model	Encoded name of the car
fuel_type	Type of fuel used by the car
max_torque	Maximum Torque generated by the car (Nm@rpm)
max_power	Maximum Power generated by the car (bhp@rpm)
engine_type	Type of engine used in the car
airbags	Number of airbags installed in the car
is_esc	Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not.
is_adjustable_steering	Boolean flag indicating whether the steering wheel of the car is adjustable or not.
is_tpms	Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not.
is_parking_sensors	Boolean flag indicating whether parking sensors are present in the car or not.
is_parking_camera	Boolean flag indicating whether the parking camera is present in the car or not.
rear_brakes_type	Type of brakes used in the rear of the car
displacement	Engine displacement of the car (cc)
cylinder	Number of cylinders present in the engine of the car
transmission_type	Transmission type of the car
gear_box	Number of gears in the car

steering_type	Type of the power steering present in the car
turning_radius	The space a vehicle needs to make a certain turn (Meters)
length	Length of the car (Millimetre)
width	Width of the car (Millimetre)
height	Height of the car (Millimetre)
gross_weight	The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg)
is_front_fog_lights	Boolean flag indicating whether front fog lights are available in the car or not.
is_rear_window_wiper	Boolean flag indicating whether the rear window wiper is available in the car or not.
is_rear_window_washer	Boolean flag indicating whether the rear window washer is available in the car or not.
is_rear_window_defogger	Boolean flag indicating whether rear window defogger is available in the car or not.
is_brake_assist	Boolean flag indicating whether the brake assistance feature is available in the car or not.
is_power_door_lock	Boolean flag indicating whether a power door lock is available in the car or not.
is_central_locking	Boolean flag indicating whether the central locking feature is available in the car or not.
is_power_steering	Boolean flag indicating whether power steering is available in the car or not.
is_driver_seat_height_adjustable	Boolean flag indicating whether the height of the driver seat is adjustable or not.
is_day_night_rear_view_mirror	Boolean flag indicating whether day & night rearview mirror is present in the car or not.
is_ecw	Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not.
is_speed_alert	Boolean flag indicating whether the speed alert system is available in the car or not.
ncap_rating	Safety rating given by NCAP (out of 5)
is_claim	Outcome: Boolean flag indicating whether the policyholder file a claim in the next 6 months or not.

Part 1 - Feature Engineering and Model building

1. Data cleaning:
    - The dataset is large and it may contain some data error.
    - In order to get clean, error free data some pre-processing requires.
    - Perform necessary data cleansing, convert data types in appropriate data types, remove duplicates, remove inapplicable columns, change date formats.
    - Visualize missing data using bar plot and heat map. Handle missing values (if required).
    - Save the cleaned data.

2. EDA:
    - Observe the variables and their distribution.
    - Create new columns if required.
    - Save the output data.
3. Data Modeling:
    - Handle the categorical column data.
    - Prepare categorical features for correlation matrix.
    - Employ label encoding technique i.e. use label encoder to convert categorical values to the numerical values.
4. Model Building:
    - Split the data in training and testing data set.
    - Build following machine Learning Models:
        - Logistic Regression Model
        - KNN Model: use n_neighbours = 100
        - SVM Model
        - Decision Tree
    - Evaluate performance of all these modules.

- Columns to create: 1. torque to rpm ratio = torque / rpm

  2. power to rpm ratio = power / rpm
- Identify the best performing model for finding potential customers.
- Save that model into .pkl file.
- Total Expected columns for Model building must be above 85
- Model Accuracy Must Be below 90%

Part 2 - Prediction using Streamlit Application

- Create interactive UI by using Streamlit.

Part 3 - Visualization using Tableau Application

- To design an interactive Tableau dashboard that visually explores and analyzes key attributes influencing insurance adoption. This dashboard will support the machine learning model development by providing intuitive insights, helping business users and data scientists understand the data distribution, trends, and potential segmentation of customers.