# Car Insurance Claim Prediction:

**Context:** The company has shared its annual car insurance data. The task is to analyse this data to uncover real customer behaviours and patterns.

**Content:** The dataset consists of columns representing real-world features. The outcome column indicates whether a customer has filed a claim (1) or not (0). The data includes 19 features, with 18 of them representing various logs recorded by the company.

# Problem Statement:

Given the car insurance dataset, develop a model that can predict whether a customer will file a claim (1) or not (0) based on the available features. The goal is to analyse the data, extract meaningful insights, and build a predictive model that can assist the company in identifying high-risk customers and optimizing their insurance policies.

## Data Discovery:

1. Load the car insurance dataset from a CSV file into a Pandas Data Frame.
2. Display the first 5 rows of the dataset.
3. Check the data types of each column in the dataset.

## Data Cleaning and Processing:

1. Identify and handle missing values in the dataset. How to decide whether to impute or drop them? Perform the necessary steps for the respective columns.
2. Identify if any columns have incorrect data types (e.g., age stored as a string). convert them to the correct data type?

## Exploratory Data Analysis (EDA):

1. Analyse the distribution of the 'age' column in the dataset.
2. Find the correlation between numerical features in the dataset.
3. Visualize some EDA charts and highlight key insights from the data.

## Feature Engineering:

1. Bin columns such as 'Driving Experience'(e.g.: 0-5 as beginner, 5-10 as intermediate) and 'Age'(eg:10 -20 as teen, 20-30 young etc ).
2. How do you handle categorical variables in the dataset, especially if they need to be used in a machine learning model? If there are any categorical variables, handle them appropriately.
3. How do you perform feature scaling on the data?
4. If the dataset is imbalanced, how do you handle this scenario? Implement the necessary steps to address the imbalance.

## **Train:**

1. Which types of models would be best considered for predicting binary classification?
2. Train three of the best considered models above. Use some hyperparameters while training.
3. What evaluation metrics would be used to assess the performance of your claim prediction models?
4. Perform visualizations for the predicted models and identify which model is performing the best.
5. How would you explain the bias-variance trade-off in the context of the predicted models?

## Deployment and Monitoring:

1. What factors should be considered when deploying a trained model into production?
2. What are the different ways of deploying a model?
3. How would you monitor its performance over time?