LIBERTY INSURANCE CASE STUDY

Problem Statement

You are given a dataset containing policy information of motor insurance customers and the total claims they have filed with an insurance company. The goal of this case study is to work towards building a model to predict the total number of claims a customer is going to file with the company.

Objectives and Expected Output:

1. Build a model, using one or more techniques other than a GLM, to predict the frequency of claims (claim_count per time of policy) a customer's going to make with the company.

2. In this case study, we are interested in all the steps involved in the model building process. Particularly data clean up, engineering new features, modelling and evaluation. Log and report all the key decisions made along the process.
3. Present your findings in a format you find most appropriate—Jupytr notebook, RMarkdown, PowerPoint, etc. The expectation is this can be used to communicate to a non-technical business audience. In the interview, you will be asked about your modelling decisions and key findings
4. Email the Presentation & all your source code.

The expectation is to spend ~6 hours completing this case study, but you won't be held or evaluated based on time spent (either less or more).

Data file (casestudy_data.csv) is provided as a separate file. Description of data fields is given below.

DATA DESCRIPTION:

- **policy_desc:** Policy Identifier; Primary Key which is unique for every policy
- **claim_count**: Total Claims (This is the response you should predicting); *Numeric Variable*
- **cat_areacode**: Area Code; *Categorical Variable*
- **num_vehicleAge**: Age of the vehicle; *Numeric Variable*
- **num_noClaimDiscountPercent**: Percentage of discount applied to policy premium based on claim history. If value is greater than 100 then policy premium was increased, if it's less than 100 a discount was applied. A value of 100 means the premium remain unchanged; *Numerical Variable*
- **cat_carBrand**: Insured Vehicle Brand; *Categorical Variable*
- **num_populationDensitykmsq**: Population density of the city the policy holder lives in; *Numerical Variable*
- **cat_Region**; Region of the country the policy holder lives in; *Categorical Variable*
- **ord_vehicleHP**: Vehicle Horsepower; This feature is anonymised but maintains the same ordinality; *Ordinal Variable*
- **num_exposure**: Exposure time of policy. Time period within which the claims were made; *Numerical Variable*
- **cat_fuelType**: Insured Vehicle Fuel Type; *Categorical Variable*
- **num_driverAge**: Age of the Policy Holder; *Numerical Variable*