## Introduction

The objectives of this project are to:

* Conduct an exploratory data analysis (EDA) to familiarize yourself with the data, to see what the data seems to say, and to uncover and manage any issues with the data.
* This would include, but is not limited to, things such as:
  + variables that need to be transformed
  + outliers
  + data errors
  + missing data
  + duplicate records
* Given the business context of the problem (described below), identify appropriate target and predictor variables.
* Create a generalized linear model (GLM) to address the business problem, evaluate model performance, and use the evaluation results to improve model performance.
* Interpret the findings of the predictive analytics model from a technical perspective.

The requirements of this project are to:

* Complete the project by the deadline,
* Submit a single code file, and assessment data set (Scored.csv), and,
* Respond to any questions about your project, if requested, to complete the project requirements.

You have been provided with the necessary datasets (explained in more detail below) and these instructions. No other data sets are needed to complete this project.

## Problem Description

For this project, you are contracted as a consultant by senior level management who have no actuarial background to refresh auto rates. The company does not want the in-force rates to be considered under this contract (i.e., do not consider the in-force rates in the model that you build). You are tasked with estimating the expected pure premium for a private passenger automobile book of business where the expected pure premium can be calculated as expected frequency multiplied by expected severity. The expected frequency is the expected number of claims for an exposure (number of claims divided by exposures), and the expected severity is the average expected cost of a claim (incurred loss divided by the number of claims). Expenses such as commissions, taxes, or loss adjustment expenses and profit should not be considered in this project.

### Problem to be Solved

You are to analyze the data that has been provided to you and perform exploratory data analysis (EDA) to ensure that the data comports with the data definitions in the data dictionary provided below. During this phase, you are also to review the data to determine if any variables require transformation. You are also asked to determine whether there are any outliers, obvious errors, or duplicates of unique keys in the data and treat those appropriately. You are to also identify and treat valid missing data elements by removing them from the data. **Data imputation is outside the scope of this project** and therefore all records with missing data should be removed. The EDA phase will also include joining the various datasets into one dataset to begin modeling.

Note that the claims dataset will only contain claims for those policies that have incurred a claim during the experience period. If a policy did not incur a claim during the experience period, it will not have a matching record in the claims dataset (auto-claims file). In both the auto-claims file and prior.claims file, there’s not a record for every policyholder. If a policyholder does not have any prior claims, then there will be no matching record in the prior-claims data set. For this policyholder, a value of zero in the column prior.claims is appropriate. This should not be confused with a customer that has a missing value in any other attribute.

Using the cleansed data, determine a pure premium model. Using an iterative approach and best practices, you are to select the models that produce the best frequency, severity, and ultimately a pure premium model to estimate the pure premium on the data available in the Assessment file (Scored.csv). The policies in the Assessment file should not be used in the model building process.

Once the predicted pure premium has been determined for the policies in the Assessment dataset, you are to write the “id” and predicted pure premium to the Scored.csv file for grading.

## Datasets

The data used in this project was simulated and should not be relied upon in any way for any use outside of this project. The sole purpose of this simulated data is for this CAS project.

Every dataset includes the variable ‘id’, which is simply the identification number of the policy and the connection point (primary key) of the following five modeling datasets.

**Auto-Customer Dataset**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | **Variable Type and Valid Entries** |
| age | Age of the insured | Numeric (16 to 97) |
| gender | Gender of the policyholder | Categorical (2 bins): M and F |
| marital.status | Marital status of the policyholder | Categorical (5 bins): Single, Married, Separated, Divorced, and Widowed |
| yrs.licensed | Number of years the policyholder has had their driver’s license | Numeric (0 to 76) |
| region | Territory in which the policyholder garages their vehicle | Categorical (14 bins): Letters A through N |

**Auto-Vehicle Dataset**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | **Variable Type and Valid Entries** |
| vehicle.age | Age of the vehicle on the policy | Numeric (0 to 33) |
| vehicle.value | Value of the vehicle on the policy | Numeric (80 to 90,000) |
| body | Describes the body style of the vehicle | Categorical (3 bins): convertible, sedan, and suv |
| seats | Number of occupants the vehicle can hold | Categorical (3 bins): 4, 5, and 8 |
| drive | Describes the drivetrain of the vehicle. AWD = All Wheel Drive, FWD = Front Wheel Drive, and RWD = Rear Wheel Drive | Categorical (3 bins): AWD, FWD, and RWD |
| mpg | Miles per gallon that the vehicle can obtain | Numeric (19 to 53) |
| length | Length of the vehicle in centimeters | Numeric (161 to 210) |
| height | Height of the vehicle in centimeters | Numeric (54 to 77) |
| hp | Amount of horsepower the vehicle outputs | Numeric (95 to 440) |
| engine.size | Size of the vehicle's engine in liters (Rounded to the nearest decimal) | Numeric (1.2 to 6.0) |
| cylinders | Number of cylinders in the vehicle's engine | Categorical (4 bins): 3, 4, 6, and 8 |
| fuel | Type of fuel the vehicle consumes | Categorical (4 bins): diesel, flex, gasoline, and gasoline/hybrid |

**Auto-Prior Dataset**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | **Variable Type and Valid Entries** |
| prior.claims | Total number of claims the insured has had prior to joining the Company | Numeric (1 to 10) |

**Auto-Exposure Dataset**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | **Variable Type and Valid Entries** |
| nb.rb | Indicator to determine if the policy is a new business or a renewal where NB = New Business and RB = Renewal Business | Categorical (2 bins): NB and RB |
| exposure | Total number of exposures for a policyholder in the experience period | Numeric (0 to 1) |
| year | Policy year | Numeric (2018 to 2022) |

**Auto-Claim Dataset**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Variable Description** | **Variable Type and Valid Entries** |
| clm.count | Number of claims insured has incurred during the policy period | Numeric (1 to 5) |
| clm.total | Total amount incurred from all claims incurred during the experience period for the insured | Numeric (800 to 80,000) |

**Assessment Dataset**

This assessment dataset is not to be used in model development or model validation – view this data as the type of data your model will receive when it is in production. It contains the same information as the previously mentioned datasets aside from Auto-Claims but on a different group of insureds. Upon completion of the model build and validation process, you are required to use your model and provide pure premium prediction (i.e., scores) for each insured in the Assessment data. The Assessment results should be saved as a .csv file (named ‘Scored’) and contain two columns – ‘id’ and ‘Prediction’. The Reviewers will use the ‘Scored’ data to evaluate your model.