Predicting annual premium for Motor Insurance

Problem Statement and Context

One of the most difficult tasks insurance companies are dealing with, is setting up a correct annual premium for an insurance policy. This is very critical because it can lead to huge company losses if not set properly and if risk is not well evaluated. Annual premium is the price the insurer is paying on a yearly basis to have his vehicle insured against different risks (e.g. accident, natural catastrophes, theft). This means that in case of any of these insured risks, insurance company will cover the damage for the vehicle. If the vehicle or the driver is risky, insurance company needs to set higher premium to compensate for it, because there is a higher chance some damage will occur. Riskier customers are those who had previous claims or accidents with the vehicle, drivers who recently got a driving licence or also older vehicle models.

Scope of solution space

The insurance company has collected information about both the driver and the vehicle, plus policy data (annual premium and sales channel) which can be used to build the prediction model. They are now interested in building a model that can suggest what an annual premium should be based on the available attributes.

Based on the available data, I will perform an analysis to understand what the most important features are for deciding on the premium and use those in building a model. It would be useful to check both linear regression and random forest regression in order to see which one predicts the dependant variable (annual premium) more accurately.

Constraints

Data might not be sufficient to properly predict prices as more attributes might be needed. Also, in some attributes there might be less heterogeneity in responses which can lead to having to exclude it from the model regardless of its importance for pricing.

Data sources

The dataset is available in csv format. The data set consists of 10 attributes and 127.000 observations. Those 10 attributes show the age and gender of the driver, as well as the region where he lives, while also depict the characteristics of the car such as age and whether it was previously damaged or not. We also know if the vehicle was previously insured and what is the annual premium now. Last, but not least we can also track the sales channel used.