Prediction of Loss Ratio for Auto Insurance Policies

Business and Data Understanding

The auto insurance industry provides insurance for vehicles and drivers from various facets of life. To predict the loss ratio the company has to avoid inappropriate pricing of policies and vary their rates accordingly. The data that has been provided by the auto insurance company is used to predict the ratio of losses to premiums earned by the company. The loss ratio is calculated by using claims paid by the insurance company plus adjustment expenses against the total amount earned by the premiums. The insurance company has rich historical data of the customers and insurances they have been using from previous years, hence we can use data science techniques to solve this business problem. To show an instance of the loss ratio, assume that the company pays 80$ in a claim for every 160$ premium amount collected, the loss ratio would be 50 per cent. The labelled training data can be used to predict the loss ratio of new data, thus this is a supervised problem. The target variable is the loss ratio, which is a numerical value. Regression techniques can be used to solve this business problem since the target variable is a numeric value. The testing and training data set provided are almost similar, but for loss amount, claim amount and loss ration that is extra in the training data set. Most of the attributes are defined precisely, but still, we have to take extra care with the data during the data preparation step. The next step is to identify the crucial attributes that will help us to predict the loss ratio accurately. Then we use these selected attributes in our data frames to build our models. There are multiple avenues we can explore after this step, for example, we can train a model using both train and test dataset to predict the loss ratio. Another approach is to train a model using training dataset and test it on test data to predict the loss ratio.

Data preparation

The most tedious process in data science approach is data preparation, it takes around 70% of the resources and time for this particular step. To train our model, first, we need to clean the data which will help us to build better models. Data preparation includes various steps like null value replacement, missing values, median or average calculation, removing outliers, reformatting the data to our needs, making corrections to the data, combining data sets to suit our problem and many more. This step may include deleting the whole record, replacing null values or dropping an attribute which does not help in our prediction. Every data scientist may have a different approach for data preparation, but the main motive is to prepare our data that is more suitable to solve the business problem on hand. Techniques like model-based imputations can be used to predict the model for the target variable in the data set that contains missing values. Mean imputation, ratio imputation and regression imputation are few approaches we can use. The trends in data can be used for approaches involving Interpolation and Extrapolation. If the data has a categorical value it would be useful to convert them to numerical values by using dummy variables, one hot encoder and label encoder. In the auto insurance project, we have converted NA values to 0 in vehicle bodily injury column. Removed unknown values from various columns like vehicle anti-theft device, vehicle passive restrained columns and many more. We have dropped columns like vehicle territory, vehicle annual miles, vehicle collision coverage indicator and many more that are not important to the analysis.

Modelling

The modelling technique we used here is regression modelling technique. This technique uses a statistical approach for estimating the values between the target variable and the other attributes present in the dataset. In the previous assignment, we have used Lasso and Ridge Regression to predict the target variables value. It completely depends on the problem at hand, we can use various regression techniques to solve this problem. We want to solve this business problem using different techniques and compare them to select the best fit for us. We have to be vary of outliers while developing these models. Visualization techniques like histogram, boxplot, scatter plot and other techniques can be used the behaviour of our data. The most correlated features like annual premium, vehicle make year, policy company. Calculation of least square distance between the different data points will help to analyse the problem set. We have to see how different types of data behave with the selected models.

Evaluation and Deployment

The evaluation and deployment process should involve the business stakeholders because at the end of the day they are the people who decide if the model is a good fit for them or not. The data scientists job is to explain how the model will be profitable to the company using visualization tools, so people should be able to understand the data science jargon. We have to explain how the classification threshold is selected and utilized in our model. The next step in the process is to experiment with our model with test data and see how it works. If the stakeholders believe that the model increases their profitability and gives them an added advantage against their competitors they will implement the model to solve their business problem. For evaluation, we can use cross-validation techniques namely 5-fold, 10-fold evaluation to check for model overfitting. In terms of the auto insurance data set we are using the project, we have to show to the stakeholders that the predictive power of our model will indeed reduce the loss ratio for the company. The expense of this approach should be justified for the company to implement this approach to solve their business problem.