Auto Insurance data science project

Sharmistha-801061221

1. Business and data understanding

a) What exactly is the business problem to be solved?

The goal is to reduce the loss of an auto insurance company via an analysis of natural logarithmic of loss ratio using a predictive model, we will predict the natural logarithmic of the loss ratio of a portfolio of auto insurance policies.

We have a testing data set of 330 policy portfolios each having around 1000 auto policies.

The training data contains of a set of auto policies including several policy levels attributes as well as Annual Premium and Loss Amount.

b) Is the data science solution formulated appropriately to solve this business problem?
Yes.

- The loss ratio of a policy is the Loss Amount divided by the Premium
- The loss ratio of a portfolio of policies is the sum of all the Loss Amounts of all the policies in the portfolio divided by the sum of all the Premiums in the portfolio
- The target is the natural logarithm of the loss ratio of a portfolio

c) What business entity does an instance/example correspond to?

- The target entity is Loss Amount
- The main features are: Annual_Premium, Vehicle_Make_Year, Vehicle usage, Vehicle Symbol, Number of drivers, Vehicle performance, Miles to work, Vehicle Age, Total drivers, Driver's Age, Vehicle_Anti_Theft_Device etc.

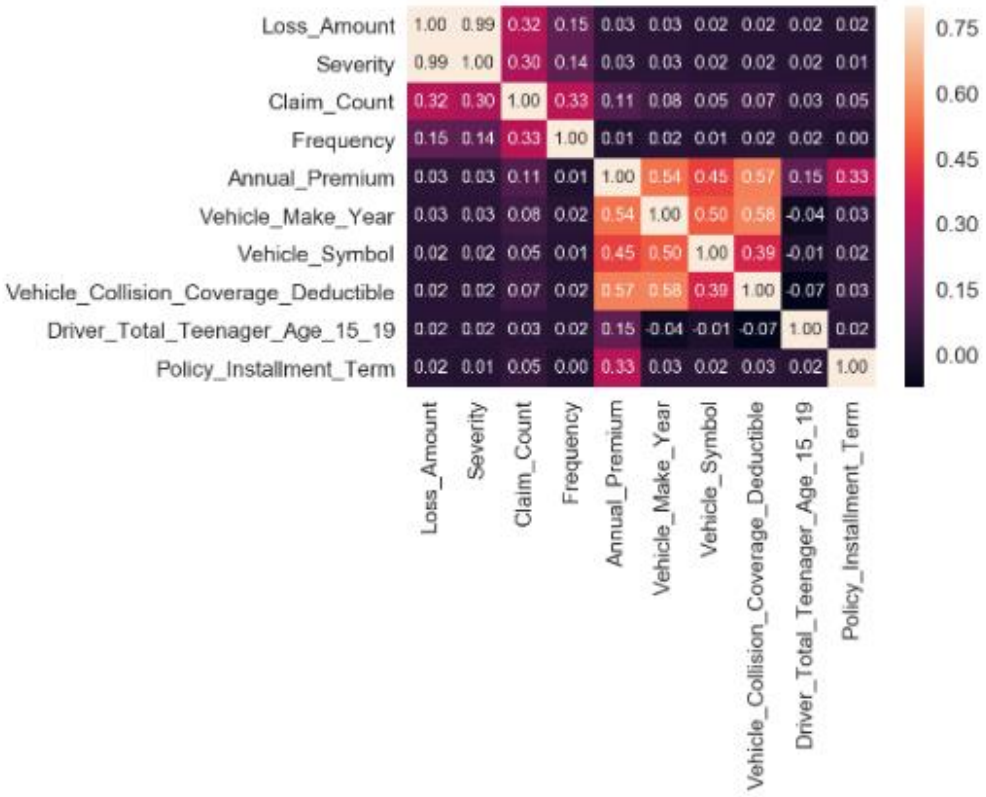
d) It is a supervised regression problem where the target and other attributes are well defined

2. Data preparation

- The data is represented in the form of table with 424431 entries and 67 attributes, the data contains auto insurance policies of one year. We must do feature engineering on the attributes and select the most relevant attributes for our model.
- Out of 68 features we have 27 categorical and 40 numerical
- Categorical attributes will be converted to numerical attributes using one hot encoding
- We will use "Correlation matrix Heat Map" to determine the most relevant features

Auto Insurance data science project

Sharmistha-801061221



Top 10 numerical features strongly correlated to Loss Amount

- Some of the features are interpreted as numerical but it is categorical, for e.g. Vehicle_Make_Year. We need to transform such attributes to string data type to make it categorical
- We will check for outliers in the data (e.g. claims of $1 million) in our dataset we have attribute “Vehicle_Days_Per_Week_Driven” which has values as 8, -1 and 9.
- “Vehicle_Annual_Miles” attribute is not useful as it has values for only 16 entries
- We will adjust the missing/null values in the data, by eliminating/ considering mean or default values as per the business knowledge and experts advise
- We will standardize the training data before solving the problem, as it results in faster convergence of machine learning algorithm. Will achieve this by calculating the mean and variance of an attribute
- As the data is Imbalanced, we must fix the skewed feature so that our model will be more accurate while making prediction

Auto Insurance data science project

Sharmistha-801061221

3. **Modelling**
   - The target variable is real and continuous value, we can proceed with supervised machine learning regression algorithm
   - The simplest is linear regression
   - We will use cross validation technique for training the model, as it is more sophisticated training and testing procedure and it also helps in choosing optimal regularization parameter
   - We will be using both lasso and ridge regularization co-efficient
   - We will use ensemble method to train the models i.e. training different models (linear regression, support vector machine, Random forest, Adaptive boosting etc.) and then make predictions using the average of individual predictions
   - In ensemble modeling we would resample the data for each model using with replacement and without replacement of training data, basically using different training dataset to train different model
   - Finally combining the output of all the models will result in improved performance

4. **Evaluation and Deployment**
   - We would like to have a domain knowledge expert validation
   - The evaluation uses cross validation hold out technique
   - The result will contain two columns “ID” and “ln_LR”. ID will identify the portfolio in the testing set and ln_LR is the target prediction i.e. the natural logarithmic of loss ratio of the portfolio
   - Based on the losses the auto insurance company can adjust their premium amount
   - They can create new rules specific to attributes to maximize profit
   - This will help in managing the risk and incur profit
   - The result will be evaluated based on the Mean Absolute Error (MAE); we can plot a regression line graph to show our results to the stakeholders