Proposal for the Auto Insurance data science project

Introduction:

The Auto Insurance dataset given consist of training dataset and 330 testing portfolios. Training dataset consists of more than 400K entries and 69 columns. Testing portfolios have 5 columns less than training dataset hence we need to check for those columns while predicting the target. The Training dataset has numerical as well as categorical data. The goal of the project is to predict the loss ratio.

Business and Data Understanding

- The exact is the business problem to be solved is :
  The training data contains a set of auto policies including a number of policy level attributes as well as Annual Premium and Loss Amount. The problem to be solved is : to predict the natural logarithm of the loss ratio of a portfolio of auto insurance policies in testing dataset.
- Business entity does an instance/example correspond to:
  Each Record in dataset represents the details of each auto insurance policy like policy number, policy installment , vehicle make year, annual premium , losses , loss ratio.
  The business entity here is the loss ratio that is to be predicted for testing portfolios.In the dataset, 50 % of the policies are mispriced by more than 10% plus minus upto 50. Loss ratios can be used to reduce or increase rates.
  Eg. Permissible loss ratio = 1-expense ratio = 0.7
  Portfolio loss ratio = 0.666
  Rate Change Factor = 0.666/0.7-1 = -0.04762
  Results in a rate reduction of 4.762%
  Loss reserving determines the present liability associated with the future claim payments. Expected Loss Ratio can be used to determine the estimated losses.
  Loss ratio = Total losses/total annual premium
- The problem is supervised problem. The Target attribute is the loss amount which can be then used to calculate the loss ratio. The loss ratio has a range between 0.0 to 24787.14. Loss amount can take up values between 0.0-1.072292e+06.
- For supervised problems: will modeling this target variable improve the stated business problem, modelling this target variable - loss amount would improve the stated business problem. Here the sub problem is first find the loss amount then calculate the loss ratio which is total losses/ total annual premium.

Data Preparation

- Data Preparation is the most important task in any Data Science project. Data cleaning needs to be done for the Auto Insurance Dataset as data is having missing , unknown, outliers and this takes a lot of time of total data science project. After that Feature Selection is done. We can get the values for attributes and select the required features necessary for determining the loss amount.

- The model is supervised and the target variable loss amount is well defined. The values of target can be calculated by taking into consideration the attributes that help to predict it(most correlated attributes). For now I am discarding following irrelevant columns from dataset: Columns to be removed policy_no, Vehicle_make_description, *Claim_count *Frequency *Severity *Loss_ratio. Some of the columns are removed as they are not present in the testing data.
- The loss ratio is to be acquired with the remaining columns. Here the dataset contains categorical as well as numerical data. In order to do furthermore feature selection we need to convert these categorical attributes into numerical in order to plot correlation matrix. One Hot Encoding can be used for it. Also there are -1(irrelevant), unknown, nan values in given dataset preprocessing needs to be done for those attributes. Following are the some of the columns that need to be corrected : Columns to alter:
  *EEA_Policy_Zip_Code_3 : Discard the unknown
  *Vehicle_Bodily_Injury_Limit : Remove nan , remove 1M-1M

Modeling

- After preprocessing the data, right mode must be selected. Linear regressions are among the simplest types of predictive models. Linear models essentially take two variables that are correlated one independent and the other dependent and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points.
- As the dataset is of regression I am planning to choose Linear Regression Model for the target variable. Also other models can be checked.
- Linear Regression meets the other requirements of the task. Also the modeling technique is compatible with the prior knowledge of the problem.
- Other models will be also tried and compare which model gives the best result.

Evaluation and Development

- Domain Knowledge Validation is necessary as all the stakeholders might not be from technical background . For that a detailed report for each step with timeline, cost and outcome of different phases can be prepared.
- Development phase would start as soon as the dataset is cleaned, model is built and other necessary operations are carried out. The (Training)features that are selected as the most related features to the target are being given to Linear Regression(/Finalised model). The model then processes the training data and then work on the testing portfolios to predict the loss amount.
- Model evaluation aims to estimate the generalization accuracy of a model on future (testing data) data. We are taking into account business costs and benefits take. The various business costs such as dataset preprocessing , employing various feature selection techniques to get the best features for predicting the loss amount, trying various. models along with Linear Regression.
- Different metric evaluations can be used like ROC , AUC can be used.
- Cross Validation can be used to avoid overfitting.

For regression, to evaluate the quality of numeric prediction, error metrics enable us to compare regressions against other regressions with different parameters. These metrics are short and useful summaries of the quality of our data. Error metrics like Mean Square Error, Mean Absolute Error ,Root mean squared error. All of the above measures deal directly with the residual produced by our model. For each of them, we use the magnitude of the metric to decide if the model is performing well. Small error metric values point to good predictive ability, while large values suggest otherwise.

The Deployment of the project is planned in the last week of November 2019. The model would successfully able to predict the loss ratio after all the above steps are carried out properly.

Conclusion

Auto Insurance Dataset is used on a Data Science model to calculate the losses occurred over a year. Then this model can be used to increase the profit and thus justifying the expenditure of project.