Auto Insurance Data Mining Proposal - Loss Ratio Prediction

BUSINESS AND DATA UNDERSTANDING:

The business problem we are handling is corresponding to Auto Insurance business entity. Given a large dataset consisting of auto policies for one year (2006) and 50% of policies are mispriced from ±10% up to 50%. Our goal is to find out the natural logarithm of Loss ratio in order to reduce or increase rates and thereby reduce the losses and maximize the profits to the organization. There are two datasets, training and testing. We model on training dataset and will use that model to predict the target variable in the testing dataset. The training dataset comprises a set of attributes of policies, such as Annual premium, claim count, loss amount, frequency, severity and various other attributes. The target variable is ln_LR (natural logarithm of portfolio Loss ratio) which is obtained by dividing total losses by total premium amount for each portfolio and taking natural logarithm of that. It is a continuous variable where the output is real-valued, both negative and positive values. As the target variable is defined precisely, this is a supervised problem. The other attributes are also defined precisely. By modeling the target variable ln_LR, the business problem is completely addressed to maximize the profits. The subtasks that need to be solved is to find out the missing attributes in the testing data. We cannot use expected value framework to structure the sub task, since those attributes are numeric.

DATA PREPROCESSING:

We will develop a predictive model using modern data-mining technology. The training dataset is divided into 330 portfolios containing all the combinations of parameters as following:

- number of policies per portfolio: 1000, 3000, 5000.
- The percentage of policies that have losses: 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%).

Now these portfolios will be further divided into train and validation sets in 80:20 ratio using Cross Validation. Cross-validation will be used for yielding better results and to select the best regularization parameter. We will look for any missing values in each portfolio of the training dataset by checking for the null values. Outliers can be seen using visualization techniques like histograms. We will handle them by imputation techniques or by dropping unnecessary values.

For each training portfolio, we will do feature engineering on the set of data to summarize the data in that portfolio i.e. mean driver age, mean miles to work, total premium, total losses, mean/mode for claim count, mean for severity, sum,mean,median for other attributes etc. With all the new values in hand, we will create a new dataset with these new features and ln_LR of each portfolio which is calculated from the total premium and total loss.

The given attributes in the portfolios would suffice to calculate ln_LR attribute. There is not much cost involved in calculating ln_LR as all the attributes are available in the training dataset portfolio.

Next we will check the correlation between attributes using Heatmaps to see the relationship between attributes and magnitude of the association between them. We will use PCA (Principle Component analysis) for dimensionality reduction.

The attributes consist of both categorical and numeric attributes which needs to be handled during preprocessing. Categorical attributes need to be encoded and then all the numeric attributes need to be normalized and standardized.

MODELING:

Then using various techniques like linear and polynomial regression models, we will train various models iteratively on these new features and ln_LR and validate using validation dataset. These regression techniques are appropriate in this business case as the target variable is continuous valued. The performance and accuracy of these models has to be compared by calculating R-square to choose the best model.

Then we will use this model to predict the ln_LR in the testing portfolios. We can compute ln_LR for the training dataset easily as we have all the required data available in the training dataset. But in the given testing dataset, 4 important attributes (claim count, loss amount, frequency, severity) are missing, making it unclear to predict the target variable. In order to predict ln_LR, we must find out the attributes on which the Loss amount is dependent on. Loss amount is highly dependent on claim count and severity (Loss Amount= claim_count*severity). This means we need to build separate models for estimating claim count and severity. It involves extra costs pertaining to the prediction of these missing attributes.

Using Gradient Descent with regression technique in order to fit the model to the data will help in increasing the speed of learning and application. Using Lasso Regularization on Linear regression can generalize model performance and performs variable selection by setting some coefficients to zero.

EVALUATION AND DEPLOYMENT:

Model evaluation is done by using cross validation technique. Also, the Evaluation metrics, Explained variance, R-square coefficient and Mean squared error will be calculated. For domain-knowledge validation by stakeholders we use validation curves and learning curves to visualize the generalization and to diagnose bias and variance.

We take the baseline Loss Ratio value as 0.67658. Any predicted value which is greater than this baseline value indicates that the premium rate has to be increased. And any value lesser than this baseline value indicates low or zero loss, hence the premium value can be left as it is unchanged.

FLAWS IN THE PROPOSAL:

In the subtask of predicting missing attributes in the testing dataset, the approach for predicting claim count and severity is not clear. The proposal is just assuming to take mean/mode of claim count and mean of severity. Also, it’s not clear on how the other missing attribute frequency is related to the Loss ratio and it is not taken into account for the prediction.