Big Data for competitive advantage: Spring 2020 Project Proposal

Abdul Rahman G
agulammo@uncc.edu

1 ABOUT THE DATA:

Kaggle is an online community of data scientists and machine learning practitioners, which allows user to build and publish data sets and enter competitions to solve data science challenges.

2 BUSINESS AND DATA UNDERSTANDING

The business problem I am trying to solve is to build a model that predicts the natural logarithm loss ratio of a portfolio of policies. Also, a natural log transformation would help us to reduce the influence of outliers in the data and also helps to eliminate heteroscedasticity. Further log transformation will make the given data from a skewed distribution, if any, to a normal distribution.

This is a supervised machine learning problem and we are going to use predictive modelling approach. Here the target variable is Loss_Ratio and takes numerical values between 0 and 24787, each value representing different policy number.

This prediction model provides a precise loss ratio that helps the auto-insurance companies to address the challenge of mispricing of auto-insurance. It will help them to establish a price technique that differentiates low or high-risk customers based on multiple attributes. Detecting appropriate loss-Ratio can also be used to reduce or increase the premium rates and determine the estimated loss. As I am not aware of the actual business value attached to the dataset, I wish to provide the expected value approach for framing the problem in the later phase.

3 DATA PREPARATION

The dataset is from Kaggle. The dataset contains training and testing sets. There are appropriately 400000+ observations with 69 attributes in the training set. There are 330 sets of policy portfolio each having approximately 1000 policies with 64 attributes.

Also, it is practically possible to get the values for attributes and create vectors and put them into a data frame by using the Panda's packages in python.

Further, the model is a supervised machine learning problem and the target variable (Loss_Ratio) is well defined for training data. For test data, the target variable (Loss_Ratio) is not given. I will use the NumPy package to calculate the natural log and put them into a data frame by using panda's packages in python.

Since the data are from a similar population i.e. the dataset contains policies for one year there is no selection biases.

4 MODELLING

Since we need multiple variables to forecast the possible outcome, I prefer using the generalized linear model (GLM) which is a multivariate regression model

I believe GLM will help me to quantify the relationship between several independent variables and a dependent variable. It is useful to understand how each independent variable relates to the target variable (Loss Ratio). Further, GLM is less susceptible to over-fitting than other predictive modelling technique.

Also, I will incorporate Tweedie distribution, which is a form of GLM, that has the ability to model both discrete and continuous distribution.

Further, during my Initial EDA (Exploratory Data Analysis) I didn't find any missing values in the given data.

5 EVALUATION AND DEPLOYMENT

To come up with optimal loss ratio and evaluation metric I need to have an extensive domain knowledge validation and cost-benefit understanding. Based on which I can formulate a confusion matrix, which helps by indicating correct predictions and develop an optimal loss ratio threshold to evaluate the model. To determine the appropriate Independent variables influencing dependent variable I will either select the predictive attributes, that is the attributes that have a non-zero coefficient or use the attribute/feature selection algorithms for the training set and remove the least corelated attributes in the testing data. Cross-Validation technique will be employed as well to improve the performance.

6 REFERENCE

1. Abraham, J., Ozaksoy, I., Alfieri, C., Ketterer, T., Obasare R., Thiha, K. (2012). Predictive Loss Ratio Modeling with Credit Scores, For Insurance Purposes. Retrieved from https://web.wpi.edu/Pubs/E-project/Available/E-project-042612 150045/unrestricted/Predictive_Loss_Ratio_Modeling.pdf