**Big Data Assignment No. 3: Project Proposal**

**Business and data understanding:**

The main goal of this project is to predict the loss ratio of auto insurance policies from the given testing portfolios. The loss amount divided by the annual premium is the loss ratio of an insurance policy. 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 testing data contains a set of 330 policy portfolios, each having at least 1,000 auto policies. The training data contains a set of auto insurance policies including several insurance policy attribute. The given problem is a supervised problem. The model that describes a relationship between a set of selected variables and a target variable is called supervised learning. The target is to compute the **natural logarithm of the loss ratio of a portfolio.** We are given various attributes that are responsible for computing the Annual Premium of a policy in the dataset. The given attributes have been defined precisely and they take values over a range depending upon what that particular attribute is dealing with.

**Data preparation:**

A set of training data and testing data is being provided. The training data will be used to create the model and the testing data will be used to evaluate the model and predict the loss ratio of the testing dataset. The testing dataset is a set of 330 portfolios each of them having at least 1000 policies with a large number of attributes. The main objective is to predict the natural logarithm of the loss ratio for each portfolio present in the testing dataset. As per the training data that I have observed, I see that the training data consisted of 420K rows along with some 70 odd attributes. Out of these some of these attributes were unnecessary which were removed as they would not be contributing towards the final target of loss ratio. Also while going through the data I was able to find and remove null values in the Annual premium column i.e. the annual premium was not present. Such rows would be useless as we need an annual premium to calculate the loss ratio. So these rows were removed. There were a few attributes that needed to be one hot encoded to understand the data in a better manner. Creating a heat map of these attributes will help us to find out the correlation of the attributes with the target that is the loss ratio. The Loss ratio will be calculated by diving the loss amount of a policy with its annual premium. In this way, we can calculate the loss ratio and the main objective is to calculate the natural logarithm of this loss ratio.

**Modeling:**

The given problem is a supervised learning problem. The various supervised learning techniques are:

* Linear regression
* Nearest neighbor
* Decision tree algorithm
* Support vector machine algorithm

The above algorithms can be applied on the training data to generate a model to predict the loss ratio of the testing dataset. We can compare the performance of these models to evaluate how these models are performing on the given dataset and then eventually use the model which is giving the best predictions.

**Evaluation and Development:**

The model assessment intends to gauge the speculation exactness of a model on future information. Model performance is quantified using model evaluation metrics. Some of the metrics used in model evaluation, namely:

* Accuracy
* Area under the curve
* Mean absolute error
* Root mean squared error

These metrics can be used to evaluate the performance of a supervised machine learning model. The assessed exhibition of a model discloses to us how well it performs on unseen data.