BIG DATA ANALYTICS

HOMEWORD #3

MOHAMMAD SAIF

Business proposal for auto insurance

The proposal for this project is to predict the potential loss a customer will invoke on the insurance company by trying to predict what loss he might incur in the future and potentially minimise it by possibly increasing the premium cost of certain criteria of current/future customers.

Using data presently available our goal is to evaluate the trends and features which have an impact on whether a customer will get into an accident (the most frequent cause of loss). We first have a look at what data we have available with us, the features we have with us on hand detail the insurance policy a customer has with features like ‘Policy Type’, ‘Zip code’, ‘Make and Model’, ‘Age of Vehicle’, ‘Age of Drivers’, ‘No. of drivers’, ‘Insurance Coverage’, ‘Premium’, ‘Loss’ and etc. Not all of these necessarily play a part in determining the amount of Loss, so we carefully analyse each feature and determine what plays a part in potential loss.

To get a preliminary view of what factors affect loss amount we perform correlation analysis using all the features and get ranking from most relevant to least relevant features.

0	**Loss_Amount**	**23**	**Driver_Total_Young_Adult_Ages_24_29**
1	**Severity**	**24**	**Vehicle_New_Cost_Amount**
2	**Loss_Ratio**	**25**	**Driver_Total_Senior_Ages_65_69**
3	**Claim_Count**	**26**	**Driver_Total_Middle_Adult_Ages_40_49**
4	**Frequency**	**27**	**Driver_Total_Upper_Senior_Ages_70_plus**
5	**Annual_Premium**	**28**	**Vehicle_Days_Per_Week_Driven**
6	**Vehicle_Make_Year**	**29**	**Driver_Total_Related_To_Insured_Spouse**
7	**Vehicle_Symbol**	**30**	**Vehicle_Number_Of_Drivers_Assigned**
8	**Vehicle_Collision_Coverage_Deductible**	**31**	**Driver_Total_Related_To_Insured_Self**
9	**Driver_Total_Teenager_Age_15_19**	**32**	**Vehicle_Territory**
10	**Policy_Installment_Term**	**33**	**Driver_Total_Male**
11	**Driver_Total_Related_To_Insured_Child**	**34**	**Driver_Total_Married**
12	**Driver_Total_Single**	**35**	**Driver_Total_Adult_Ages_50_64**
13	**Vehicle_Driver_Points**	**36**	**EEA_Policy_Tenure**
14	**Driver_Total_College_Ages_20_23**	**37**	**Driver_Maximum_Age**
15	**Driver_Total_Female**	**38**	**Driver_Minimum_Age**
16	**Vehicle_Med_Pay_Limit**	**39**	**Vehicle_Age_In_Years**
17	**Vehicle_Physical_Damage_Limit**		
18	**Driver_Total_Licensed_In_State**		
19	**Driver_Total**		
20	**Vehicle_Miles_To_Work**		
21	**Vehicle_Comprehensive_Coverage_Limit**		
22	**Driver_Total_Low_Middle_Adult_Ages_30_39**		

Obviously, not all of these are a factor to loss, the top 5 entries contain values that have a value when there is a loss so they aren’t factored in, but from a preliminary look, Young drivers, Premium and Vehicle Make Year are the most detrimental with other factors coming close to determining loss.

Based on these factors we carefully select features detrimental as too many features can cause overfitting and heavily incorrectly predict as only 4.4% of the policies have a loss amount, which is detrimental considering theoretically 95.6% policies should predict 'No Loss' but this turns out bad for the Loss Ratio even if we get an all 'No Loss' 95.6% accuracy.

The data provided to us isn't perfect, missing values, incorrect entries are commonplace (for e.g., the weeks car driven feature has entry as large as 9 when there aren't even 9 days in a week!), to deal with this we either check how many occurrences of incorrect entries there are and suitably decide whether to remove them or insert a weighted mean, removing a few hundred entries in a dataset with more than 400k entries realistically doesn't present a big issue.

Now that all the data has been cleaned and features selected that itself alone isn't enough for our goal, we need a suitable machine learning model to accurately predict loss amount in cases, there are many suitable models out there of which we will test and evaluate them using 10 fold cross validation on the training set and proceed with the best one for predicting our test portfolios, machine learning models such as 'Lasso and Ridge regression', 'ElasticNet regression', 'XGBoost', 'Multiple linear regression', 'Artificial Neural Networks' etc, can prove suitable for the job based on previous research work done, provided that the data is suitable.

Obviously accuracy itself is a bad metric in this case, so we propose to split our data into 10 smaller components (with entries with losses distributed equally and randomly across all 10) and train each model on all of the smaller datasets and rank the model based on the mean RMSE and score we acquire from the dataset as training a huge dataset is highly inefficient, however after scoring and settling on a model to use we will use the complete training data to train a model without cross validation and proceed to get predictions from the loss portfolio.

After all of this has been said and done, some for loops in python can easily automate this task for us (as it will in cleaning the data) and we should get our loss ratios for individual portfolios within a matter of few hours (there are almost 330 portfolios with ~1k entries each after all).

At the end of this proposal we aim to give a formal presentation as to the intricate details of our working, methodology and model selected with intuitive graphs and charts to help make our reasoning easier to understand, with this we aim to successfully predict losses so that the Insurance company can suitably price their premiums to reduce loss because no company wants to run in the red per se.