ITCS 6100 Big Data Analytics for Competitive Advantage
Project Proposal Report
TEAM - 14

Loss Ratio Prediction for Insurance Policies

Team Members

Harshini Karnati
Navya Alam
Renu Karule
Venkata Satya Sai Praveen Varma Bellamkonda

Problem Statement

The problem which the project resolves is the common issues faced i.e., 50% of all policies can be mispriced by more or less than 10% (up to 50%). This will have a possible implication of profitability of an insurance provider

Introduction

Based on the given training of around 4200000 input samples of auto insurance policies and the testing dataset of 600 portfolios which has ~1000 policies in each. One can use the training dataset to train a model to predict missing loss amounts in the testing dataset. For this, machine learning algorithms such as regression can be used to make these predictions.

Once the missing loss amounts have been predicted, the loss ratio for total policies in the testing dataset can be calculated using the formula:

$$Total\_Premium = \sum_{i=1}^{N} AnnualPremium(i)$$

$$Total\_Losses = \sum_{i=1}^{N} LossAmount(i)$$

Target: natural log of portfolio loss ratio

$$\ln\_LR = \ln\left(\frac{Total_{Losses}}{Total_{Premium}}\right)$$

Where annual premium is the amount the policyholder paid for the insurance policy.

After calculating the loss ratio for each policy in the testing dataset, you can take the natural logarithm (ln) of the loss ratio for each policy. This is often done to transform the data and make it more suitable for analysis, as it can help to stabilize the variance and make the data more normally distributed. This can help to identify the factors that impact the loss ratio and can be used to inform pricing and risk assessment strategies for auto insurance policies.

Data Preparation

Dataset Information:

There are a total 300 portfolios each having at least 1000 policies.

There are 69 total fields in training dataset with data types as shown below:

#	Column	Non-Null Count	Dtype
0	PolicyNo	424431 non-null	int64
1	Policy_Company	424431 non-null	object
2	Policy_Installment_Term	424431 non-null	int64
3	Policy_Billing_Code	424431 non-null	object
4	Policy_Method_Of_Payment	424431 non-null	object
5	Policy_Reinstatement_Fee_Indicator	424431 non-null	object
6	Policy_Zip_Code_Garaging_Location	424431 non-null	object
7	Vehicle_Territory	424431 non-null	int64
8	Vehicle_Make_Year	424431 non-null	int64
9	Vehicle_Make_Description	424431 non-null	object
10	Vehicle_Performance	424431 non-null	object
11	Vehicle_New_Cost_Amount	424431 non-null	int64
12	Vehicle_Symbol	424431 non-null	int64
13	Vehicle_Number_Of_Drivers_Assigned	424431 non-null	int64
14	Vehicle_Usage	424431 non-null	object
15	Vehicle_Miles_To_Work	424431 non-null	int64
16	Vehicle_Days_Per_Week_Driven	424431 non-null	int64
17	Vehicle_Annual_Miles	424431 non-null	object
18	Vehicle_Anti_Theft_Device	424431 non-null	object
19	Vehicle_Passive_Restraint	424431 non-null	object
20	Vehicle_Age_In_Years	424431 non-null	int64
21	Vehicle_Med_Pay_Limit	424431 non-null	int64
22	Vehicle_Bodily_Injury_Limit	407105 non-null	object
23	Vehicle_Physical_Damage_Limit	424431 non-null	int64
24	Vehicle_Comprehensive_Coverage_Indicator	424431 non-null	object
25	Vehicle_Comprehensive_Coverage_Limit	424431 non-null	int64
26	Vehicle_Collision_Coverage_Indicator	424431 non-null	object
27	Vehicle_Collision_Coverage_Deductible	424431 non-null	int64
28	Driver_Total	424431 non-null	int64
29	Driver_Total_Male	424431 non-null	int64
30	Driver_Total_Female	424431 non-null	int64
31	Driver_Total_Single	424431 non-null	int64
32	Driver_Total_Married	424431 non-null	int64
33	Driver_Total_Related_To_Insured_Self	424431 non-null	int64
34	Driver_Total_Related_To_Insured_Spouse	424431 non-null	int64
35	Driver_Total_Related_To_Insured_Child	424431 non-null	int64
36	Driver_Total_Licensed_In_State	424431 non-null	int64
37	Driver_Minimum_Age	424431 non-null	int64
38	Driver_Maximum_Age	424431 non-null	int64
39	Driver_Total_Teenager_Age_15_19	424431 non-null	int64
40	Driver_Total_College_Ages_20_23	424431 non-null	int64
41	Driver_Total_Young_Adult_Ages_24_29	424431 non-null	int64
42	Driver_Total_Low_Middle_Adult_Ages_30_39	424431 non-null	int64
43	Driver_Total_Middle_Adult_Ages_40_49	424431 non-null	int64
44	Driver_Total_Adult_Ages_50_64	424431 non-null	int64
45	Driver_Total_Senior_Ages_65_69	424431 non-null	int64
46	Driver_Total_Upper_Senior_Ages_70_plus	424431 non-null	int64
47	Vehicle_Youthful_Driver_Indicator	424431 non-null	object
48	Vehicle_Youthful_Driver_Training_Code	424431 non-null	object
49	Vehicle_Youthful_Good_Student_Code	424431 non-null	object
50	Vehicle_Driver_Points	424431 non-null	int64
51	Vehicle_Safe_Driver_Discount_Indicator	424431 non-null	object

52	EEA_Liability_Coverage_Only_Indicator	424431 non-null	object
53	EEA_Multi_Auto_Policies_Indicator	424431 non-null	object
54	EEA_Policy_Zip_Code_3	424431 non-null	object
55	EEA_Policy_Tenure	424431 non-null	float64
56	EEA_Agency_Type	424431 non-null	object
57	EEA_Packaged_Policy_Indicator	424431 non-null	object
58	EEA_Full_Coverage_Indicator	424431 non-null	object
59	EEA_Prior_Bodily_Injury_Limit	407105 non-null	object
60	EEA_PolicyYear	424431 non-null	int64
61	SYS_Renewed	424431 non-null	object
62	SYS_New_Business	424431 non-null	object
63	Annual_Premium	424431 non-null	float64
64	Claim_Count	424431 non-null	int64
65	Loss_Amount	424431 non-null	float64
66	Frequency	424431 non-null	float64
67	Severity	424431 non-null	float64
68	Loss_Ratio	424431 non-null	float64

Data Preparation:

All the features in the given datasets might not be useful in training the model. So we have to pre-process the data available to proceed with building a training model. Below are the steps that we intend to follow to clean and transform the dataset

1. Most of the features in the data that has been provided are either categorical or numerical, some of which can be utilized to make useful deductions to predict the target variable.
2. Any null values available in the numerical feature selected should be replaced by zero
3. Any white spaces in the categorical feature selected should be deleted
4. Divide the training dataset into set of portfolios, containing 1000 policies each, with 10% of policies with positive loss amounts
5. Create a dataframe reading the training portfolios, with the useful features such as all age columns, all the gender columns, driver points, vehicle usage, vehicle miles, annual premium, loss amounts etc
6. Categorize the data based on the features extracted. This categorization would be helpful in decision making based on age and gender factor per say.
7. Once the data frame is created capturing useful features, train the model and predict the loss amounts on testing data using the features that has been captured above
8. Furthermore, the loss amounts are predicted for testing dataset, loss ratio can be calculated

Data Exploration

To explore the dataset, some basic data checks and summaries can be performed such as:

1. Check for missing values: Missing values in any column should be handled as per the requirement
2. Check for outliers: Any unusual values or extreme outliers in any of the columns should be handled appropriately as outliers can have a significant impact on analysis results

3. Explore data distributions: The distribution of each variable is to be noted, using summary statistics and visualizations such as histograms or boxplots. This can help to identify any patterns or trends in the data, and can inform decisions about feature engineering and data transformations.
4. Check for correlations: Correlations between pairs of variables should be noticed, using techniques such as scatterplots or correlation matrices. This can help identify any relationships or dependencies between variables, and can inform decisions about feature selection and model building.
5. Explore target variable: Analyze the target variable (Loss_Ratio), using summary statistics and visualizations such as histograms or density plots. It helps in understanding the distribution and variability of the target variable, and can inform decisions about modeling techniques and performance metrics.
6. Explore categorical variables: The distribution and frequency of categorical variables should be considered, using summary statistics and visualizations such as bar charts or pie charts to understand the composition of the data, and can inform decisions about feature engineering and data transformations.

Performing these exploratory analyses, helps gaining insights into the dataset and make informed decisions about how to preprocess and model the data for predictive modeling purposes.

Predictive modeling

To build a predictive model for a given task requires gathering information regarding historical data like claim amount, premium amounts, location and policyholder details. We can use this data to train models that can predict the expected loss ratio for a new portfolio (in our case test portfolios).

In the AI field, there are multiple model techniques for e.g. linear regression, logistic regression, neural networks. Which model type is preferable or suitable for the given task is primarily dependent on the available data. While designing such a model, we have to keep in mind that the prediction might not be accurate and various factors such as unforeseen events can affect the prediction.

We are planning to implement linear regression. Our dataset has fields with linear relationships and that's why it would be simple to start with Linear Regression. Also, Linear regression works better at interpretability problems.

We are going to use Root Mean Squared Error(RMSE) and Mean Absolute Error(MAE) as error metrics to evaluate the performance of the model. MAE is less sensitive to outliers and hence can lead to more accurate results. But this is just the assumption at this initial stage. We will explore both error metrics to decide which one leads to more accurate results.

During the data exploration phase we observed that there is a linear relationship between loss ratio and policy details, vehicle and driver details. For e.g. Vehicle_Make_Year feature can be used for prediction of loss ratio because make year is related with age of year and hence cost of repairs or replacements. Recent car models can have expensive parts which will result in higher claims and consequently higher loss ratio. Therefore, we are planning to go for a Linear regression model.

Now the next step is what features from a given dataset can be used to build the model.

Findings

An ideal step would be to perform data exploration and analysis and perform feature selection for e.g. correlation analysis or feature selection scores. This will help in identifying the important features.

For e.g. does location of policyholder matter? There might be some geographical areas where risk is high.
Does coverage matter? Coverage amount has a direct relationship with loss ratio because having low coverage amount means high deductibles.
Vehicle_Make_Year feature can be used for prediction of loss ratio because make year is related with age of year and hence cost of repairs or replacements. Recent car models can have expensive parts which will result in higher claims and consequently higher loss ratio.

While going through data exploration and a model building plan, our findings are there might be a potential problem of outliers and high leverage points. These types of data points are difficult to handle with Linear Regression. But outlier points are rare in the dataset so we can handle them during the data preparation step as mentioned above.

Conclusion

The idea is to build a powerful predictive model that can forecast the loss ratios of policy portfolios, allowing insurance providers to better assess the total risk of a specific portfolio and decide on premium prices.