ITCS 6100: BIG DATA FOR COMPETITIVE ADVANTAGE

Loss Ratio Prediction

PROPOSAL REPORT

Team 19

Richa Shalom Gadagotti(801358993)

Manasa Manchineni (801352328)

Ram Sirusanagandla (801329345)

1. Introduction:

Insurance firms rely on loss ratio as a critical statistic to guide their pricing, risk management, and financial stability. The natural logarithm of the loss ratio is a quantity that has significant influence over the insurance sector, and this project report focuses on the difficult issue of precisely forecasting it. As a key metric for assessing the financial performance of an insurer, this ratio measures the correlation between earned premiums and actual losses.

The training data contains a set of auto policies including a number of policy level attributes as well as Annual Premium and Loss Amount. The testing data contains a set of 330 policy portfolios, each having at least 1,000 auto policies.We have all the necessary components needed to calculate the Loss Ratio, given the properties of the dataset. Loss ratio may be computed by dividing the Annual Premium by the Loss Amount.

Loss amount is the total amount of claims that policyholders have submitted and that the insurer has paid out. It takes into consideration the true financial cost of incidents that are insurable. The annual premium shows how much policyholders have paid in premiums overall over the course of a year. It is the insurer's main source of income.

Overall, the project involves a number of effective modeling algorithms, handling missing values, choosing pertinent input characteristics, and careful data preparation to guarantee the model's effectiveness.
Through the optimization of the model's predictive capability, these technical details hope to enable it to anticipate the amount of loss for policies in the testing dataset with precision.

2.Data preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.

DATA CLEANING:

The practice of correcting or eliminating inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset is known as data cleaning. There are many ways for data to be duplicated or incorrectly categorized when integrating different data sources which leads to the results and algorithms to be unreliable.
Data cleaning consists of removing duplicates or irrelevant observations, filtering unwanted outliers, and handling missing data.

Removing duplicates- The majority of duplicate observations will occur when gathering data. Duplicate data can be created when you scrape data, merge data sets from several sources, or get data from different departments or clients.When you discover observations that do not fit within the particular topic you are seeking to examine, those observations are considered irrelevant. For instance, you may exclude such pointless observations if your dataset include older generations but you wish to study data on millennial clients. This can reduce distraction from your main goal and increase the efficiency of your research while also producing a more manageable and effective dataset.
We can check for duplicate rows using the .duplicated() function in Pandas, and remove them using .drop_duplicates()

Filtering outliers- An outlier is a single data point that deviates noticeably from the average. Outliers can result from mistakes in data gathering as well as random fluctuations in the data. Before doing any statistical analysis on a dataset, it is crucial to identify any outliers that may

have the potential to distort the findings. First we need to identify the outliers and should remove them.

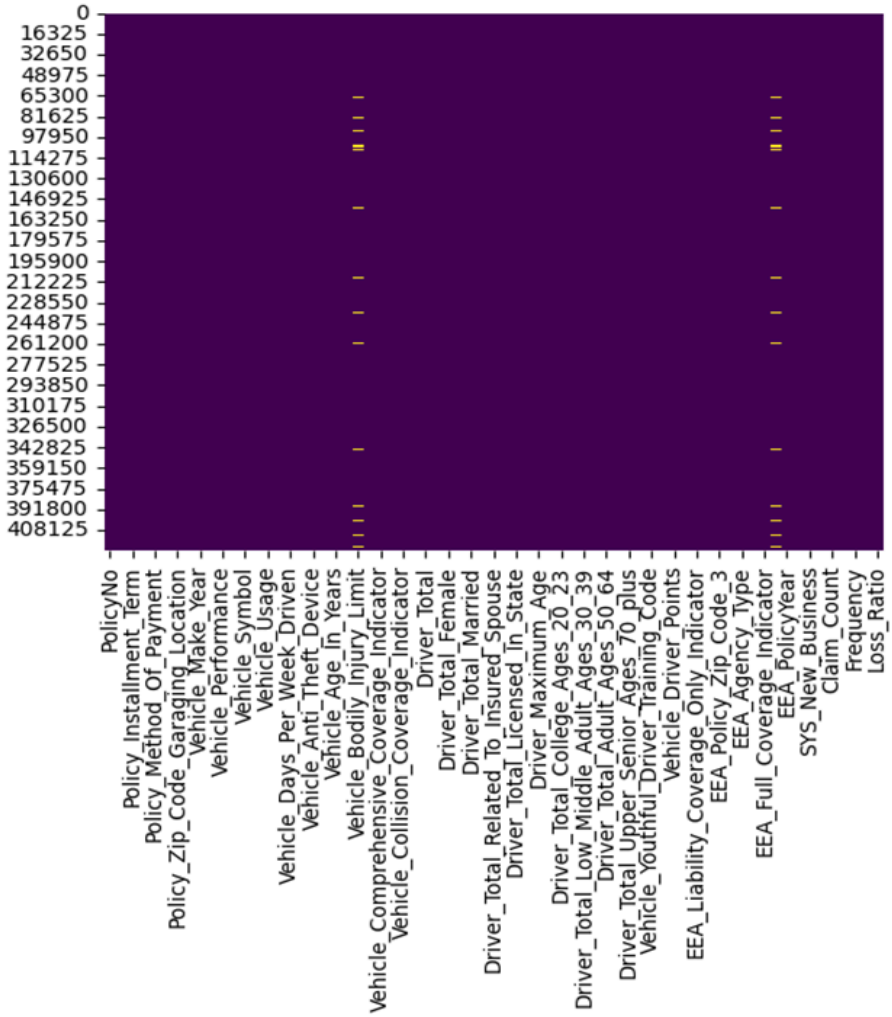
Identify the outliers- To identify outliers we can compute the interquartile range and we can identify any data points that deviate more than two standard deviations. We can also make a graphical representation of the data using box plots or graphing the data points and searching for those that are separated from the rest of the data.

Handle the outliers- And to handle the outliers, we have two options, either delete them or change them. For example, we may use a log transformation to lessen the influence of outliers or remove severe outliers by defining a threshold.

Handling missing data- Missing values can be handled by first identifying the missing values. Pandas libraries like df.isnull().sum() can be used. After identifying the missing values, we can impute the values. Techniques like mean, median, mode, or more advanced methods like regression imputation can be used to impute the values. For instance, to replace a value with the mean, we can use df['Column_Name'].fillna(df['Column_Name'].mean(), inplace=True).

For any column, if we have any domain specific knowledge, we can replace it with the appropriate value.

Heatmap showing Missing values



DATA TRANSFORMATION:

Data transformation is the act of transforming unprocessed data into a format appropriate for analysis and modeling.

Conversion of categorical data- First we need to identify the categorical variables in the dataset like “Policy Billing code”, “Vehicle usage” etc. and select a suitable encoding technique depending on the kind of variable. One-hot encoding can be used for nominal variables and label encoding can be used for ordinal variables. This converts the categorical variables into numerical representations.

Standardization and Normalization: Standardization like Z-score scaling makes the data have a mean of 0 and a standard deviation of 1 whereas data is scaled to a specified range, often between 0 and 1, by normalization (min-max scaling). Depending on the features, we determine which one to choose and based on that we apply the selected scaling technique.

Feature Engineering: New features are designed to better portray relationships within the data or to get new insights.

DATA SPLITTING:

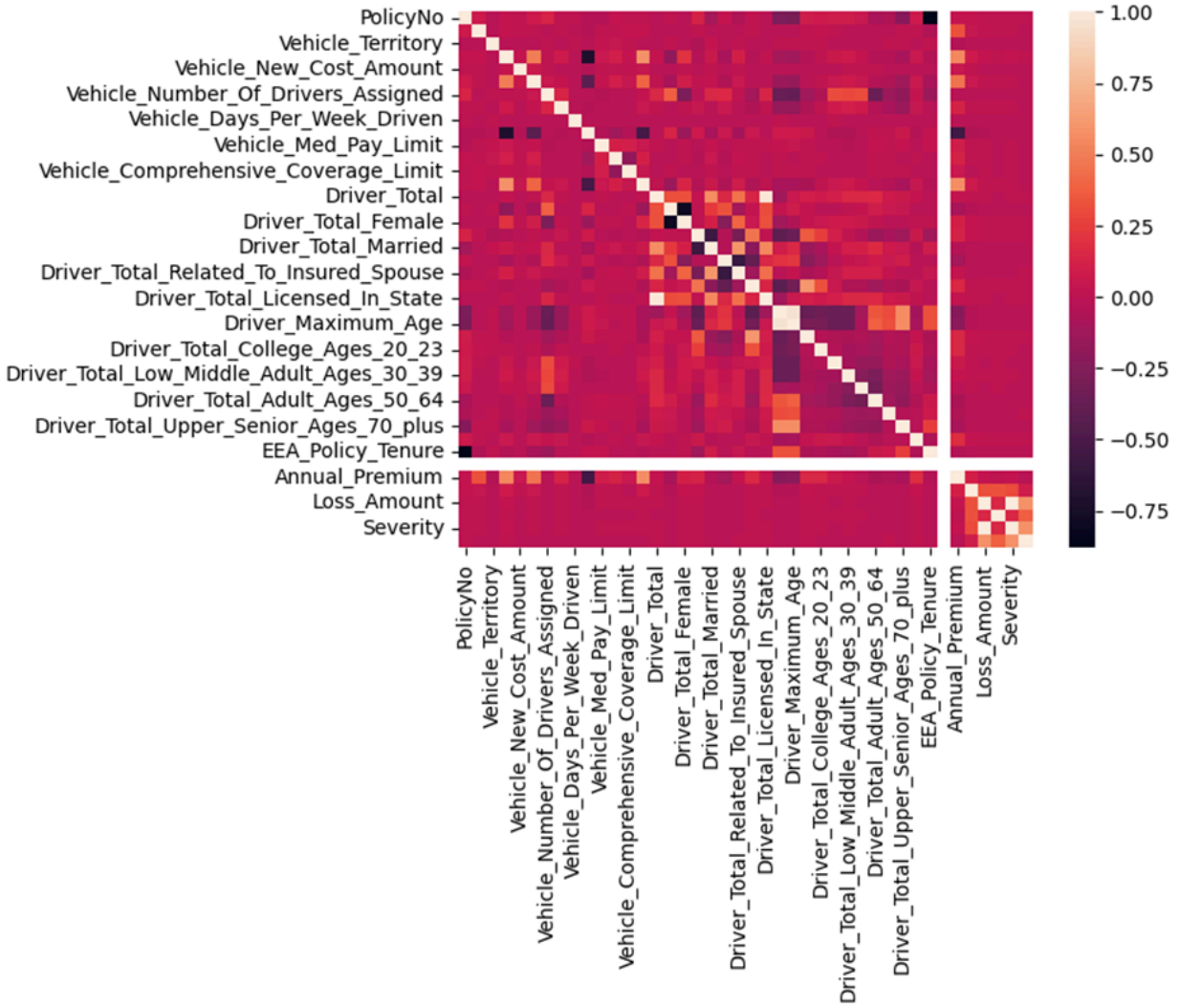
For training and testing, we divided the dataset into several subgroups. The testing set is used to assess the model's performance on unobserved data, whereas the training set is used to train the prediction models. This aids in evaluating the generalization abilities of the model.

The main columns in the training data are policy number,Policy_Company, Policy_Installment_Term, Policy_Billing_Code, Policy_Method_Of_Payment, Policy_Reinstatement_Fee_Indicator, Policy_Zip_Code_Garaging_Location, Vehicle_Territory, Vehicle_Make_Year, Vehicle_Make_Description, Policy_Billing_Code, annual premium, claim count, Loss_Amount, Frequency, Severity, Loss_Ratio whereas the testing data only consists of policy number, the other attributes and annual premium.

3. Data Exploration

We use exploratory measures to enhance our comprehension of the distribution, features, and effects of variables on our target variable, which is the natural logarithm of the loss ratio in the auto insurance market. To find insights like correlations, possible outliers, and distributions of important variables, exploratory data analysis (EDA) approaches are applied.

Correlation Heatmap



Descriptive statistics: The numerical variables in the dataset that we wish to look at more carefully are first identified, and we then compute summary statistics for these variables. In our data, variables like "Annual Premium," "Vehicle Make Year," and "Driver Minimum Age" can be used as they are quantitative in nature. Important central tendency metrics that aid in understanding the typical value for each variable, such as the mean (average) and median (middle value) can be used. Furthermore, we compute measures of dispersion that provide light on the distribution of data, such as the lowest and maximum values and the standard deviation (variation).

Correlation Analysis: Here, pairs of numerical variables that we think could be related to or have an impact on one another are chosen and depending on the nature of the variables correlation coefficients like Pearson's, Spearman's, and Kendall's are used. For example, the "Annual Premium" and "Vehicle Make Year" exhibit a correlation, suggesting that the age of the vehicle affects the premium amount. A strong positive correlation indicates a tendency for one variable to rise together with another, whereas a strong negative correlation points to the opposite connection.

Data Visualization: We pick important factors for visualization in order to better comprehend the data and its distribution. For instance critical metrics like "Loss Ratio" can be used to visualize and understand its distribution. To graphically display data, histograms, box plots, scatter plots, and bar charts are used.

Frequency Analysis: Categorical variables like "Vehicle Usage" and "Policy Billing Code" can be used to conduct frequency analysis. For each of these variables, we count the instances of each category. The distribution of categories may be understood by frequency analysis.

4. Predictive modeling

We plan to use the whole training dataset to create a model that will forecast loss ratios for the testing portfolios during the predictive modeling phase. Given that various insurance portfolios may have unique characteristics—such as gender, geography, age, and the quantity of claims—that impact their unique loss ratios, we intend to improve the model's accuracy. In order to accomplish this, we will divide the training data into discrete portfolios and create specialized models for each portfolio.These portfolio-specific models will guarantee that each portfolio's distinctive and significant characteristics are appropriately portrayed.These models will then be used to forecast loss percentages for the matching testing portfolios, leading to more precise and customized forecasts.

A range of predictive models like Linear Regression , Random Forest models and decision trees can be used.

Linear regression : Linear regression is used for its simplicity and interpretability. It takes the target variable and input characteristics to have a linear relationship. To avoid overfitting, regularized linear regression methods such as Lasso (L1 regularization) and Ridge (L2 regularization) include penalty terms in the linear regression loss function.

Decision trees : Non-linear models like decision trees are used to discover intricate connections in the data. Random Forest is an ensemble of decision trees.

5. Findings

The use of predictive modeling is essential for accurately calculating the natural logarithm of loss ratios. This accuracy strengthens our data and guarantees accurate projections. Using procedures in data preparation such as feature scaling, categorical data cleaning, and null value elimination, has highlighted the critical importance of thorough data transformation and cleaning. Using visualization techniques to observe the relationship between the variables and finding the correlation analysis have proved crucial in identifying correlations, trends, and possible predictors in the dataset.

Our study highlights the existence of other columns that do not exhibit a direct association with our objective variable, which is the Loss Amount. This information emphasizes how complex the network of links within the auto insurance industry is.

We improve forecast accuracy by focusing on supporting columns.

We have explored a wide variety of predictive modeling techniques, such as Decision Trees, Random Forest, Lasso and Ridge Regression, Logistic Regression, and Linear Regression. These results highlight how flexible these models are in dealing with the particular difficulties that the insurance dataset presents.