LOSS RATIO PREDICTION

OF INSURANCE POLICY PORTFOLIOS

ITCS 6100 BIGDATA FINAL PROJECT PRESENTATION

Team 19:
Ramu Sirusanagandla

Manasa Manchineni

Richa Gadagotti

AGENDA

1. Introduction- Scope, Workflow
2. Data Preparation (Preprocessing)
3. Data Exploration
4. Model Evaluation and Results

INTRODUCTION

Problem Definition:

- Auto insurance policy is a contract between the policyholder (the driver) and the insurance company, where the policyholder pays a premium in exchange for damage coverage.

- With the help of a vast amount of data of past customers and different aspects or attributes about them, a key metric can be measured, called the Loss Ratio.

- Loss ratio is the ratio of the claims paid out to the premiums earned.

INTRODUCTION

- OBJECTIVE:

➢ Predict the natural logarithm of portfolio loss ratio for each portfolio in the testing dataset.

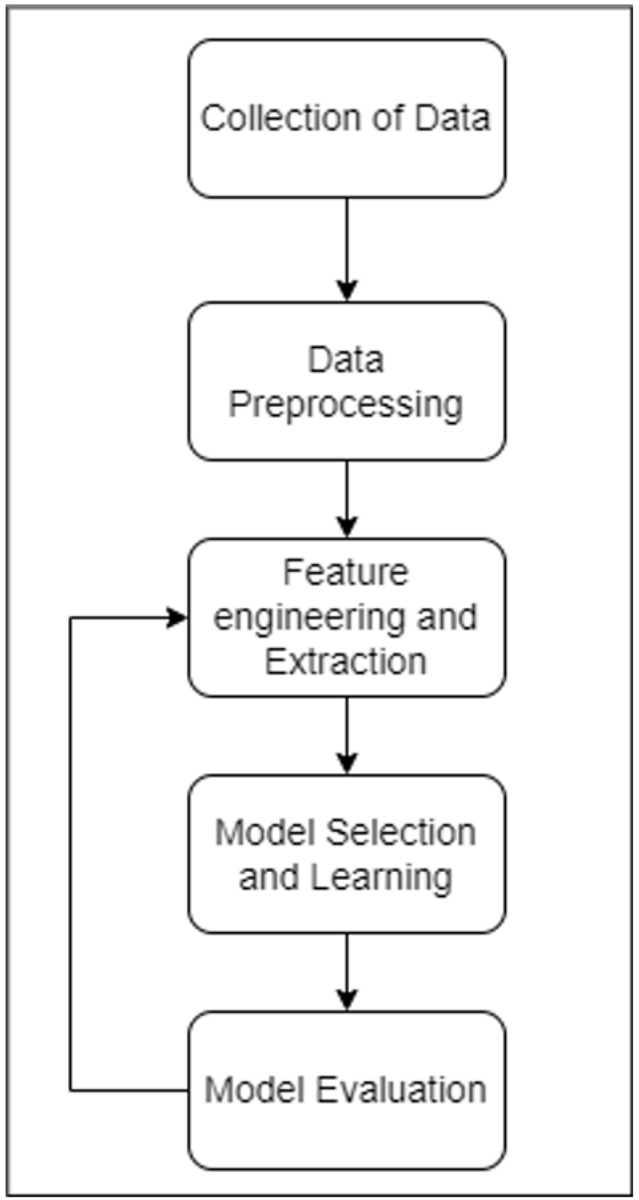
➢ This project aims to develop a model that can accurately as possible to predict the loss ratio given a portfolio.

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

Target: natural log of portfolio loss ratio

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

PROJECT WORKFLOW



DATA PREPROCESSING

- **Data preprocessing** refers to the techniques and processes used to prepare raw data for analysis.
- This involves cleaning and transforming the data to make it more suitable for machine learning algorithms or other forms of analysis.
- Steps

i. Null value elimination

ii. Removing Trailing and Leading Spaces in categorical columns

iii. Dropping columns based on high missing values

iv. Replacing values of columns with appropriate values

DATA PREPROCESSING

1. Null value elimination

- It refers to the process of removing or imputing missing or null values from a dataset.
- Example columns from dataset:
- `Vehicle_Bodily_Injury_Limit`
- `EEA_Prior_Bodily_Injury_Limit`

2. Removing Trailing and Leading Spaces in categorical columns

- Leading and trailing spaces in categorical columns can lead to incorrect groupings, sorting, or matching of categorical values, which can negatively impact the accuracy of machine learning models or other data analyses.

DATA PREPROCESSING

3. Dropping columns having high missing values

- Filling in large number of missing values can also increase the complexity of the analysis.
- Dropping columns with high missing values can simplify the analysis and reduce the computational complexity of the model.
- Example column from dataset:
  - Vehicle_New_Cost_Amount

DATA PREPROCESSING

4. Replacing values of columns with appropriate values

- Replacing incorrect or missing or NaN values with mean, median or mode of the remaining values of the respective column.
- Examples:EEA_Prior_Bodily_Injury_Limit, Vehicle_Bodily_Injury_Limit,Vehicle_Days_Per_Week_Driven

- Replacing values using data imputation by calculation
- Example: ['Vehicle_Annual_Miles'] = ['Vehicle_Days_Per_Week_Driven'] * ['Vehicle_Miles_To_Work'] * 52

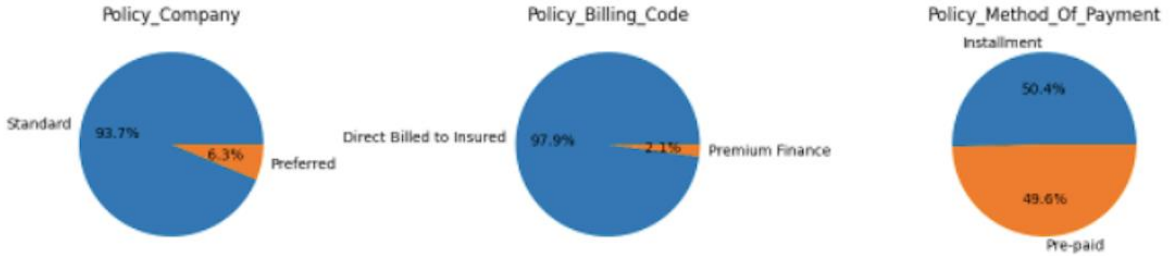
DATA EXPLORATION

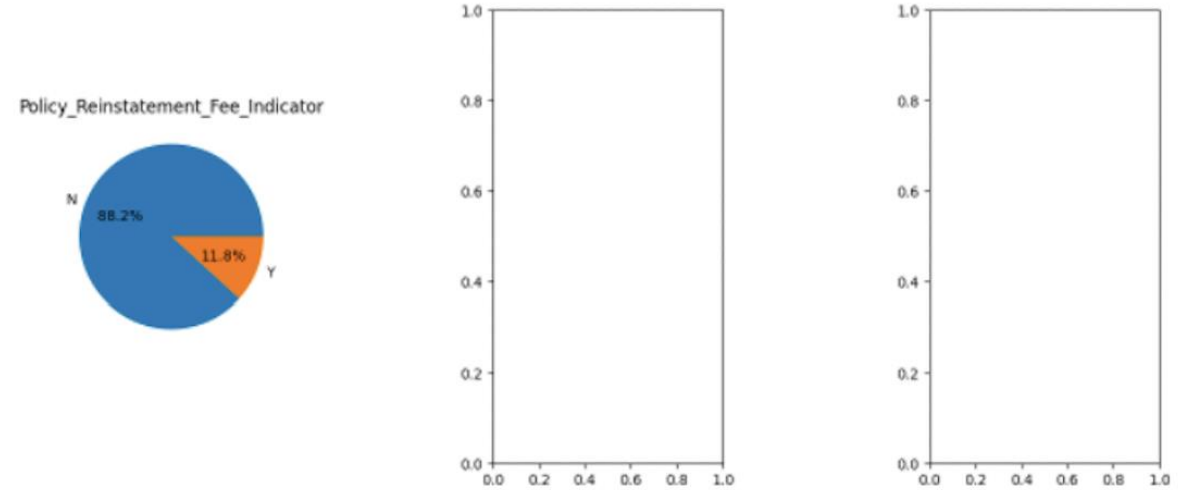
1. Dataframe division into Numerical Dataframe and Categorical dataframe.

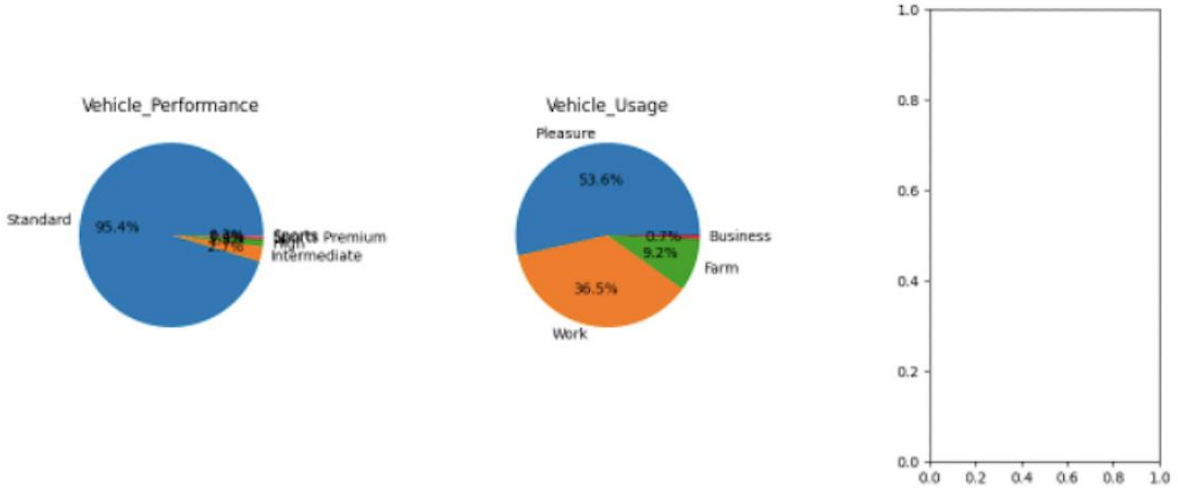
- Advantages:
    - Allows for a quick overview of the types of data present in the DataFrame and calculate statistics
    - Allows for tailored feature engineering for numerical and categorical variables separately.

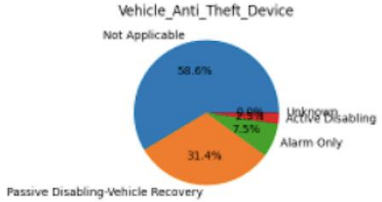
2. Plotting Bar charts of categorical columns
Advantages:

- To explore the distribution of data across categories
- To identify any patterns or trends that existing in the data
- To detect any errors or inconsistencies in the data

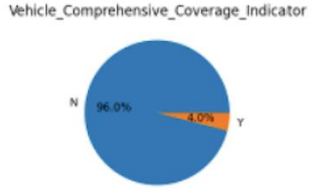


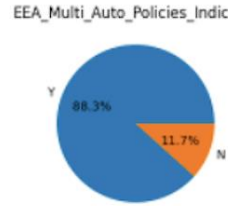


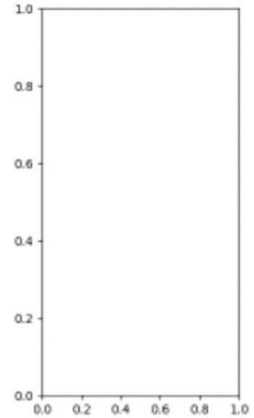


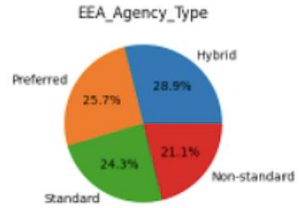

Vehicle_Anti_Theft_Device
Not Applicable
58.6%
Unknown
Active Disabling
7.5%
Alarm Only
31.4%
Passive Disabling-Vehicle Recovery

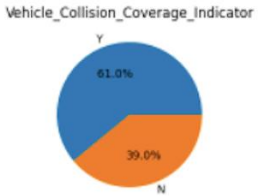

Vehicle_Passive_Restraint
Y
73.5%
0.0%
Unknown
26.5%
N

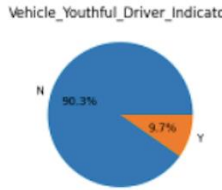

Vehicle_Comprehensive_Coverage_Indicator
N
96.0%
4.0%
Y


EEA_Multi_Auto_Policies_Indicator
Y
88.3%
11.7%
N

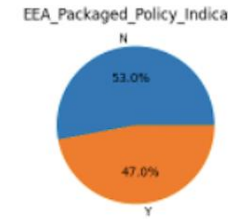

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

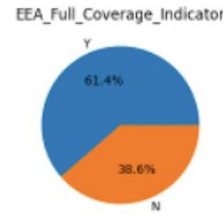

EEA_Agency_Type
Hybrid
28.9%
Preferred
25.7%
24.3%
21.1%
Non-standard
Standard


Vehicle_Collision_Coverage_Indicator
Y
61.0%
39.0%
N

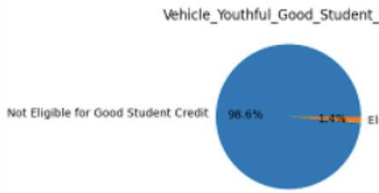

Vehicle_Youthful_Driver_Indicator
N
90.3%
9.7%
Y


Vehicle_Youthful_Driver_Training_Code
Not Applicable
90.3%
With Driver Training
Without Driver Training
4.8%
With or Without Driver Training


EEA_Packaged_Policy_Indicator
N
53.0%
47.0%
Y


EEA_Full_Coverage_Indicator
Y
61.4%
38.6%
N


SYS_Renewed
Y
88.6%
11.4%
N


Vehicle_Youthful_Good_Student_Code
Not Eligible for Good Student Credit
98.6%
1.4%
Eligible for Good Student Credit

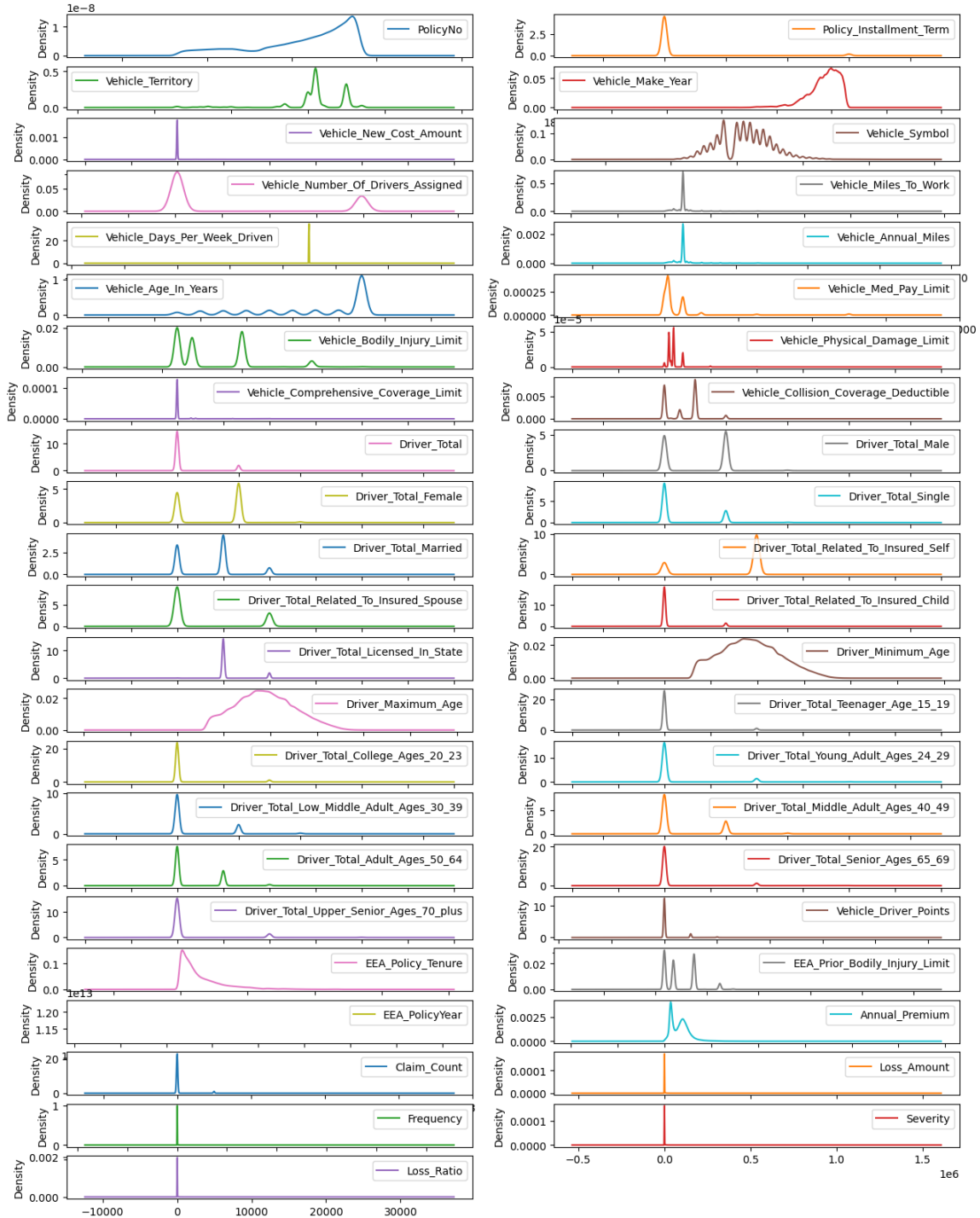

Vehicle_Safe_Driver_Discount_Indicator
Y
60.2%
39.8%
N


EEA_Liability_Coverage_Only_Indicator
N
72.1%
27.9%
Y


SYS_New_Business
N
90.0%
10.0%
Y

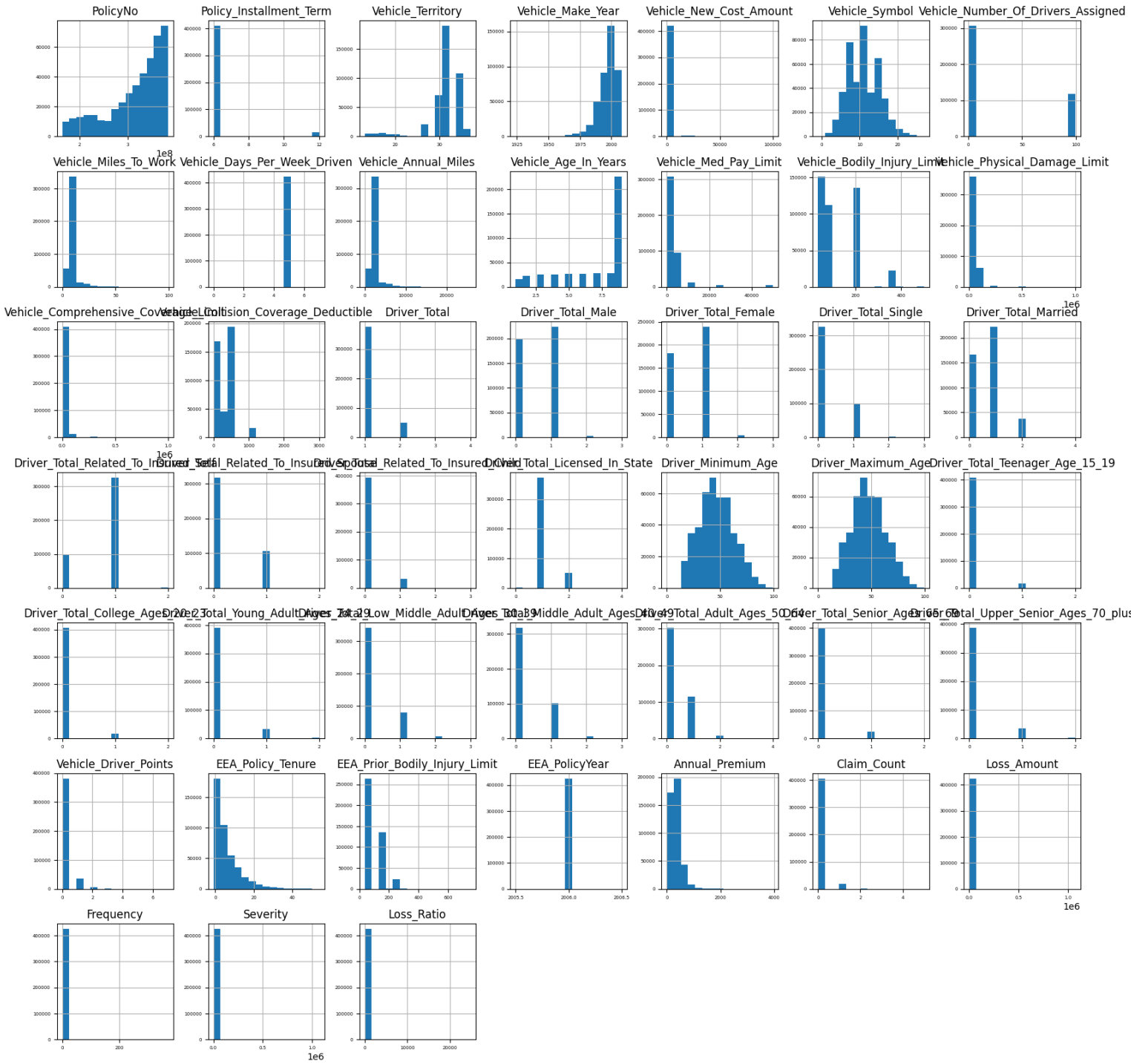
KDE PLOTS FOR NUMERICAL COLUMNS:

- The density plots reveal a diverse range of values across various variables, reflecting the heterogeneous nature of the insurance dataset.

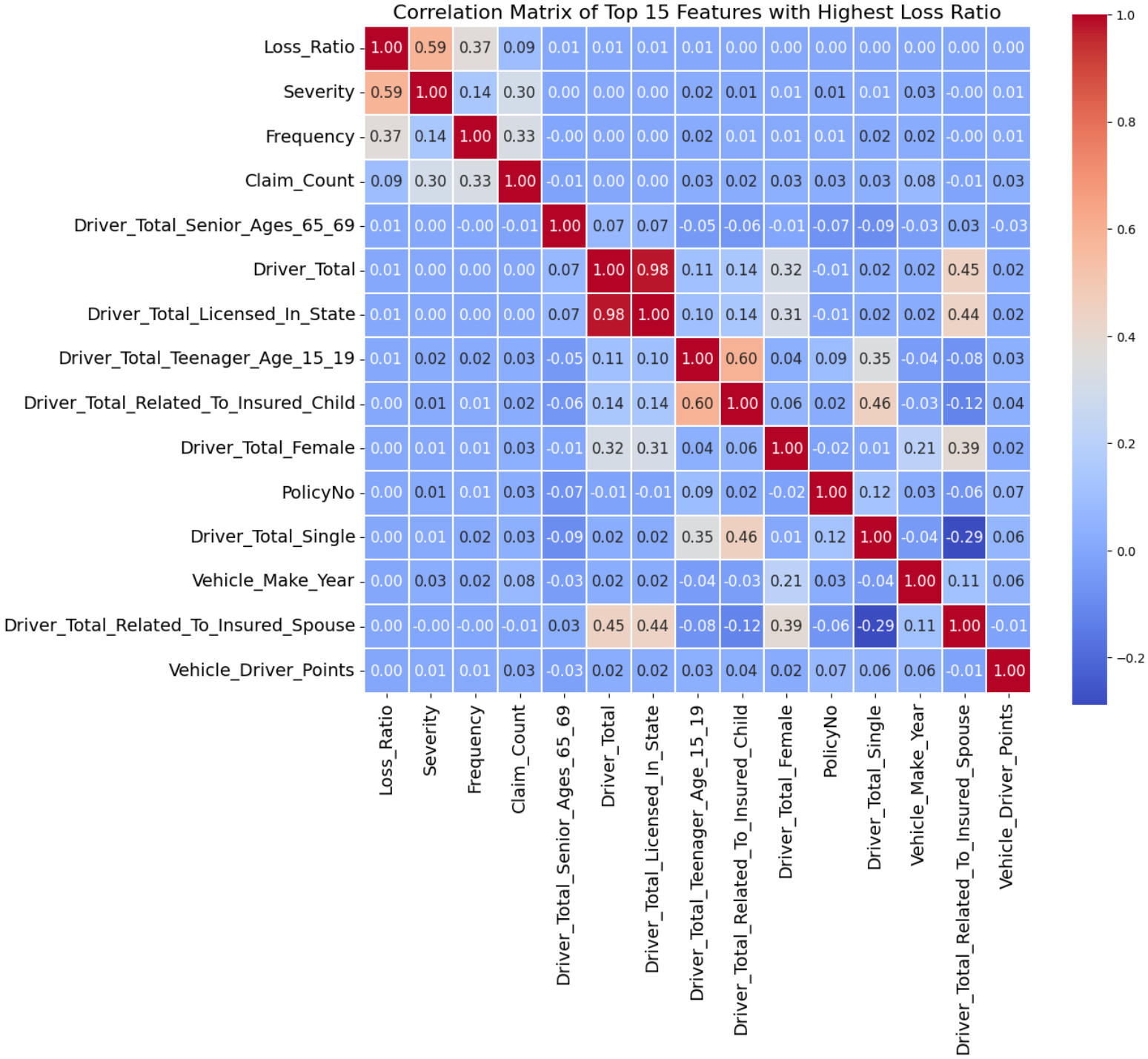


3. Plotting Histogram of each individual numerical column
Advantages:

- To explore the distribution of numerical data and identify the shape of the distribution (e.g. normal, skewed, bimodal, etc.)
- To identify any outliers or anomalies that may have been present
- To determine the appropriate statistical tests to use, and necessary data transformations

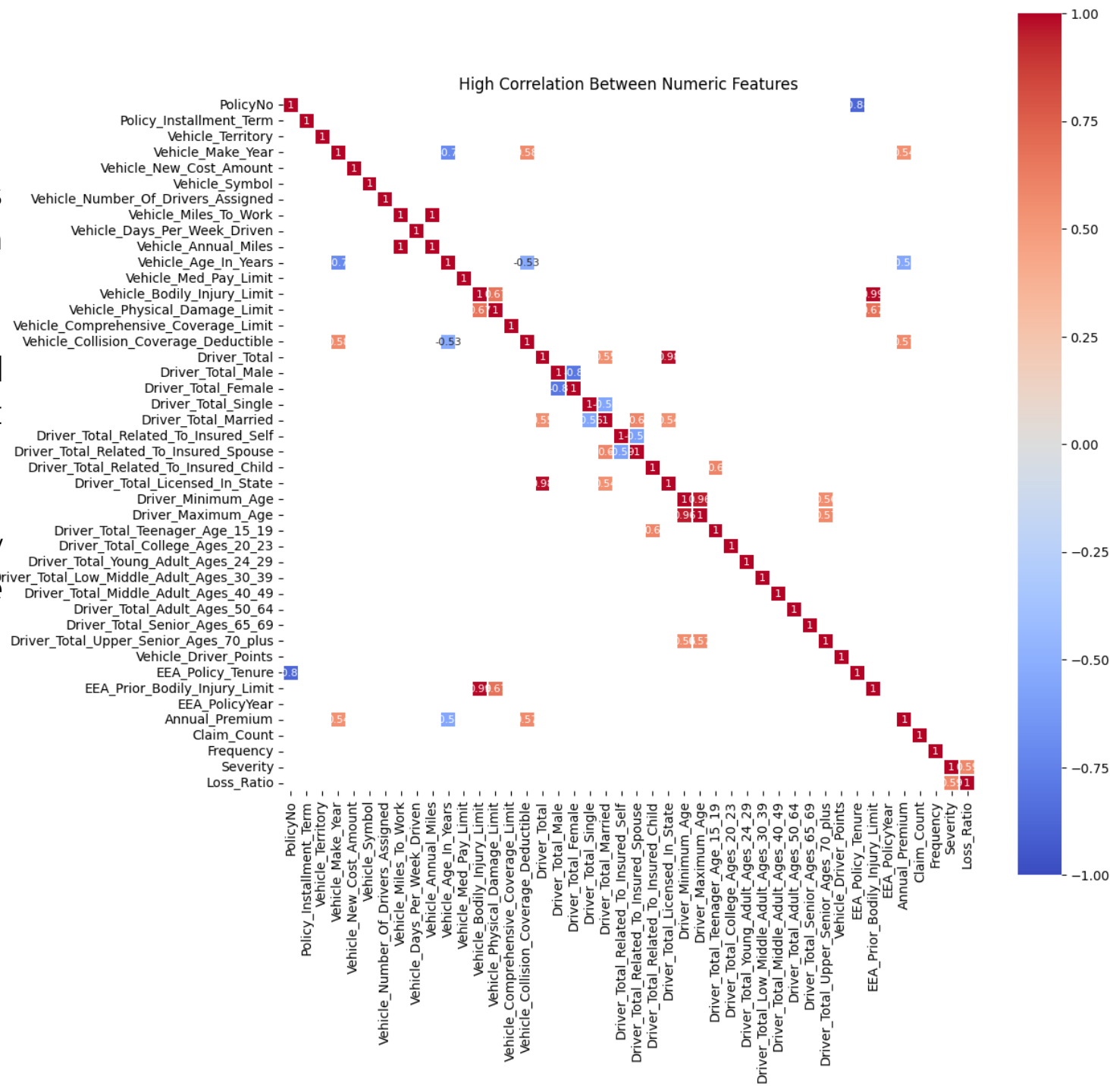


Top 15
correlated features with highest
Loss_Ratio



Advantages:

- To identify which numerical features were most strongly correlated with the target variable
- To select the numerical features had a significant impact on the target variable
- To identify the presence of any **multicollinearity** between the numerical features.



MODELING AND EVALUATION

- For this project we have decided to use Support Vector Regression (SVR) model to predict loss ratio.
- The SVR algorithm works by finding the hyperplane that best separates the data points in the predictor variable space.
- We start by using the given dataset containing historical loss ratio values and a set of predictor variables that are thought to be correlated with loss ratio.
- We used GridSearchCV in combination with a Pipeline to perform hyperparameter tuning on an SVR model for a regression problem.

MODELING AND EVALUATION

- Post performing the grid search cross-validation we determined the following to be the best hyperparameters:

```
Best parameters: {'feature_selection__k': 5, 'svr__C': 1, 'svr__epsilon': 0.1, 'svr__kernel': 'rbf'}
Best score: 0.5746840147117218
```

- The above were used to generate an improved SVR training model, the following are its metrics determined:

➢ RMSE: 0.5182059960394029

➢ MAE: 0.3496050892476213

THANK YOU!