DATA*6100 Project-2 Car Cooperators Insurance prediction
Under the guidance of Prof. Mihai Nica
Student: Atharva Vichare & Prathamesh Walawalkar

Executive Summary:

The project aims to predict whether the client will buy Insurance or not i.e. predicting 'IS_BOUND" variable in our dataset. To help us with our predictions we will train two models Logistic Regression and Random Forest and predict them on test dataset.
After comparing, Random Forest outperforms Logistic Regression in several key areas, including higher accuracy (78% vs. 51%) and AUC score (0.65 vs. 0.64), indicating better overall classification and differentiation between the classes. While Random Forest has a higher false negative rate (39% vs. 36%), it compensates with a lower false positive rate (40% vs. 44%) and generates more advertising revenue (16 vs. 15). Despite its recall being zero, its stronger metrics make Random Forest the superior choice, particularly when prioritizing overall accuracy and AUC score over minimizing false negatives.

Data Pre-processing:

Renaming Columns: After importing both the datasets we noticed that the column names in the both the training and testing datasets are not similar. So, we changed their column names to make them uniform so we can join both the datasets together and perform data preprocessing together.

Visualizing dataset: We used bar plot to visualize our target variable and noticed that it is heavily imbalanced. We also visualized missing data columns by plotting horizontal bar graph and the remaining columns in our dataset using histogram where we noticed most of our data contains categorical values which we need to deal with, as well as most of our data is skewed heavily.

Dealing with missing values: There are some features that had very high percentage of missing values (99.9%) we decided to drop them. As for other features, we filled them with mean/median imputation.

Feature Engineering:
Created new columns like "VEHICLE_CLASS", "VEHICLE_BODY" & "VEHICLE_TYPE" so that our models can capture underlying relationships and improve model performance. Splitting features like "POSTAL_CODE" to avoid increasing feature space after one hot encoding since this column has many unique values.

Feature Selection
For selecting the most important features, we used a technique called Lasso (L1) Regularization. L1 regularization works by adding penalty to the loss function which forces some of model's coefficients to shrink to zero. This process helps in reducing overfitting and helps in selecting the features that have positive impact on our model.

Model Selection and Training

We have used two different models for prediction: Logistic Regression & Random Forest. Logistic Regression is a classification algorithm rather than a regression technique. It is specifically designed

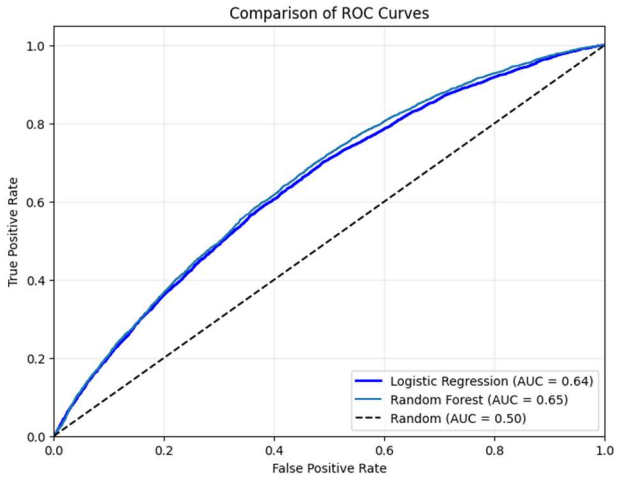
for solving binary classification problems where the target variable consists of two distinct classes, such as 0 and 1 in our case. The fundamental principle of Logistic Regression is to model the relationship between input features and the probability of a particular class. The sigmoid function, commonly used in Logistic Regression, is mathematically defined as: $\sigma(z) = 1 / 1 + e^{()}$

The output is probability value between 1 & 0. If value is close to 1 then client is likely to take insurance or else not. Based on these we can make predictions on our target variable.

A Random Forest is an ensemble learning method that builds multiple decision trees during training and then combines their predictions. Each tree is trained on a random subset of the data. When predicting whether the client will take insurance or not, each tree votes on whether the customer is likely tot take the policy. The final prediction is determined by majority voting among the trees. Then refined it by selecting important features, and optimized parameters using RandomizedSearchCV. To maximize the revenue adjusted threshold value to 0.23.

Model Evaluation and ROC Curves comparison :

We compared both the models on validation set after training looking at metrics like accuracy, precision, recall and f1 score.



To compare the performance, we plot ROC-AUC curves for both the graphs. A higher AUC generally indicated better performance. Based on the metrics, Random Forest is better because it achieves a significantly higher accuracy (78%) and a better AUC (0.65) compared to Logistic Regression's accuracy (51%) and AUC (0.64). These metrics indicate that Random Forest has a stronger overall ability to classify correctly and distinguish between classes. While its recall is zero, the high accuracy and AUC demonstrate that Random Forest performs better when the focus is on overall correct classification and distinguishing between categories, making it the superior choice in this comparison.

Final Model and Result

After evaluating the models, we predicted on the test dataset to predict the target variable. The results are as follows:

Model	Accuracy	False Negative Rate	False Positive Rate	Revenue (Cents)
Logistic Regression	56	36	44	15
Random Forest	59	39	40	16