Challenges: Binary Classification of Insurance Cross & 2024 Boston Marathon Weather and Splits

Jose Guevara

7/11/2024

Binary Classification of Insurance Cross Selling using LightGBM and fine-tuning with GridSearchCV

Binary Classification of Insurance Cross Selling: Goal and Dataset Overview

Goal

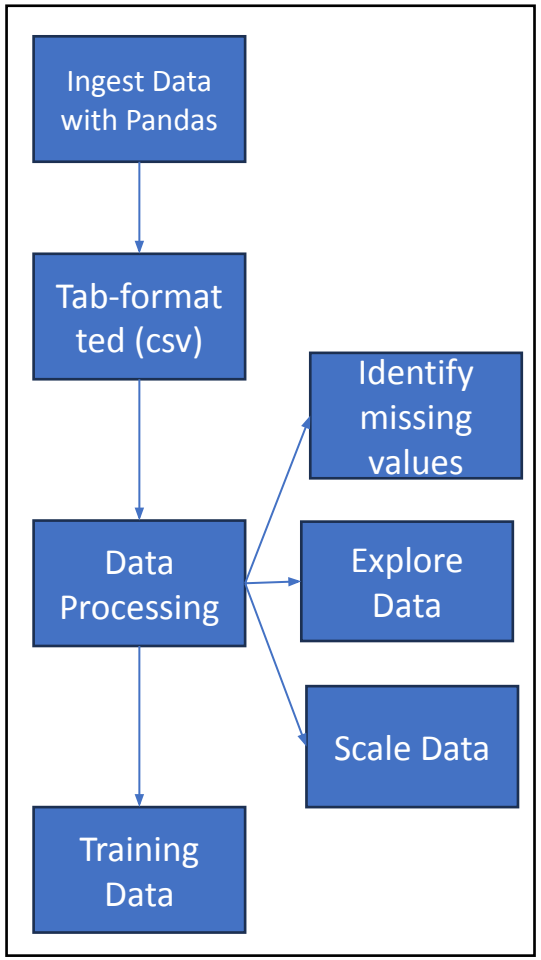
We aim to build a predictive model to determine if health insurance customers from the past year will be interested in purchasing vehicle insurance from the same company.

Features

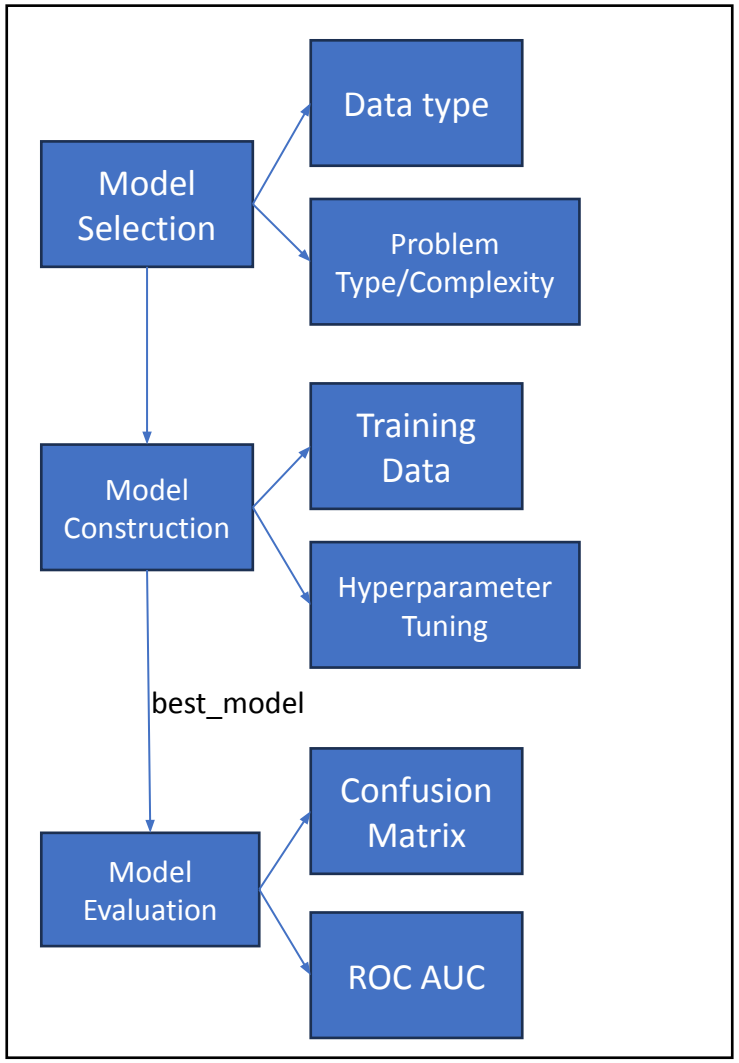
- id -> Unique ID for the customer
- Gender ->Gender of the customer
- Age ->Age of the customer
- Driving_License -> 0 : Customer does not have DL, 1 : Customer already has DL
- Region_Code -> Unique code for the region of the customer
- Previously_Insured -> 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance
- Vehicle_Age ->Age of the Vehicle
- Vehicle_Damage -> 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past.
- Annual_Premium -> The amount customer needs to pay as premium in the year
- Policy_Sales_Channel -> Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc.
- Vintage -> Number of Days, Customer has been associated with the company
- Response -> 1 : Customer is interested, 0 : Customer is not interested

Data Processing, Model Construction and Predictions

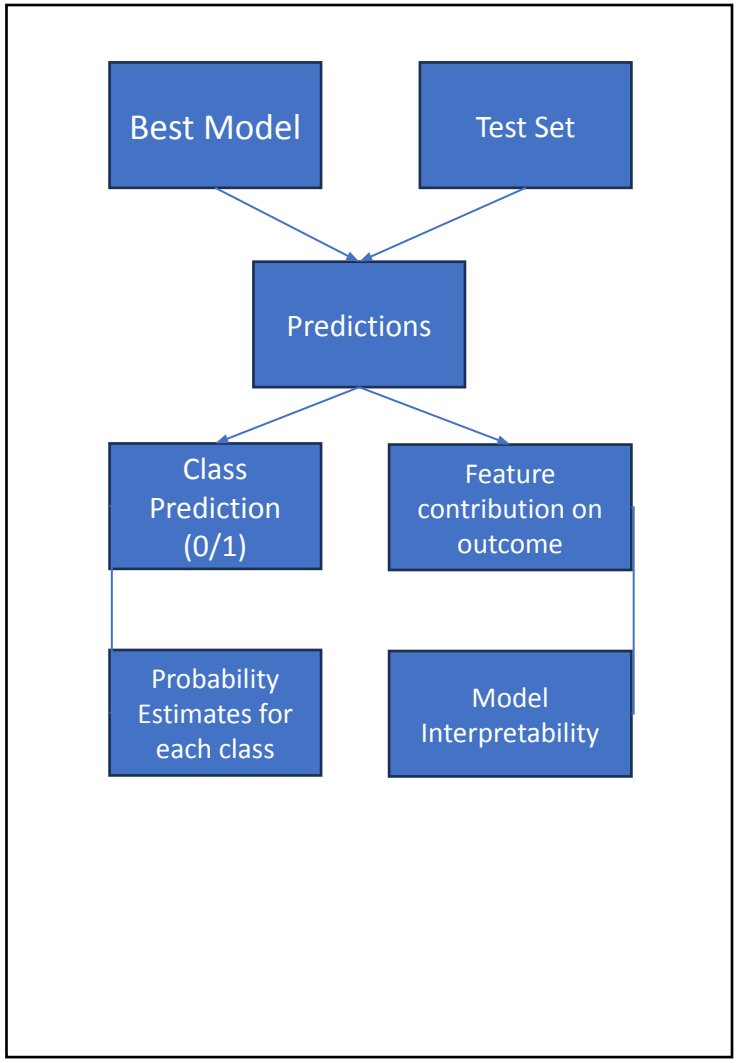
Data



Machine Learning Model



Predictions

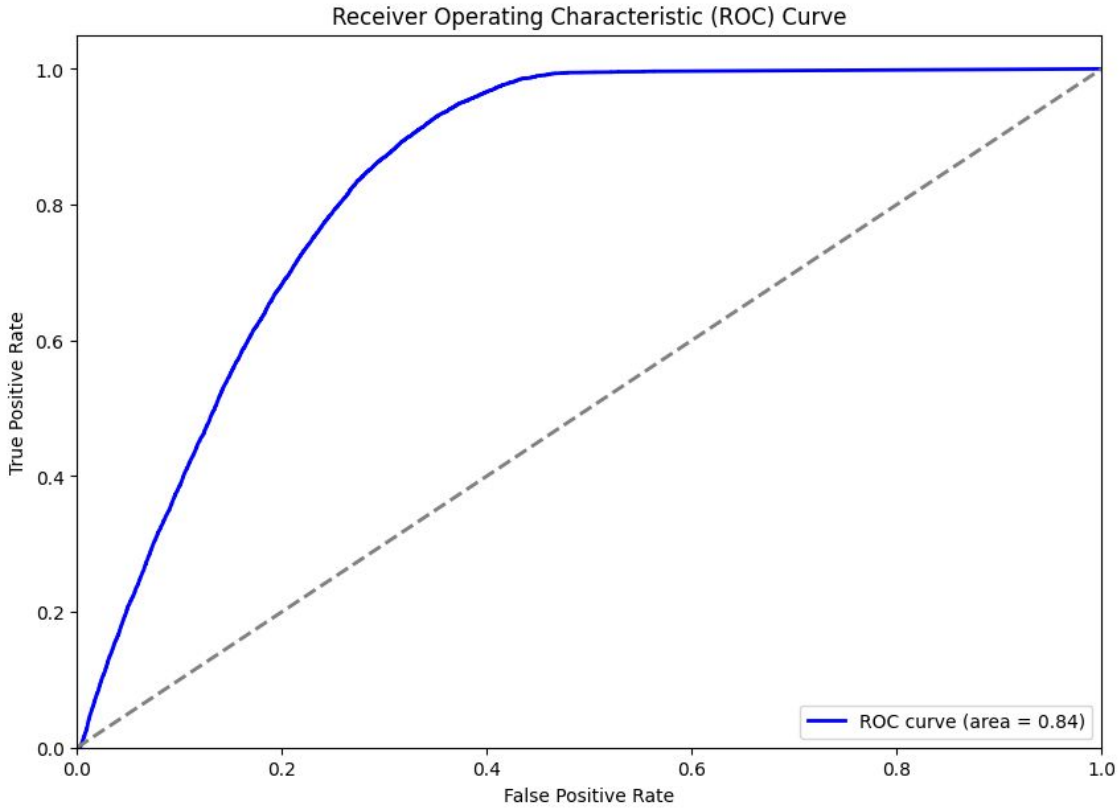


Resulting Performance of LightGBM

The resulting model led to a ROC AUC of 0.84

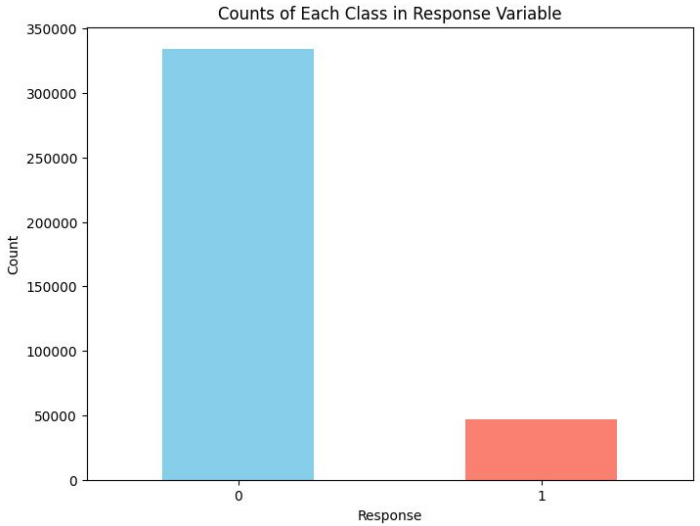
With potential more tuning this metric could be improved

However, further hyperparameters and tuning would increase training time



Identified Challenges to Improve Performance

The dataset was imbalanced ->



Hyperparameter weights: balanced was used to counter this effect

Further work could be done searching other hyperparameter space:

- Boosting type
- Max depth
- Learning rate

Exploratory Data Analysis and Potential ML Applications of the 2024 Boston Marathon Weather and Splits

2024 Boston Marathon Weather and Splits: Athletes Dataset

Athletes Dataset Features

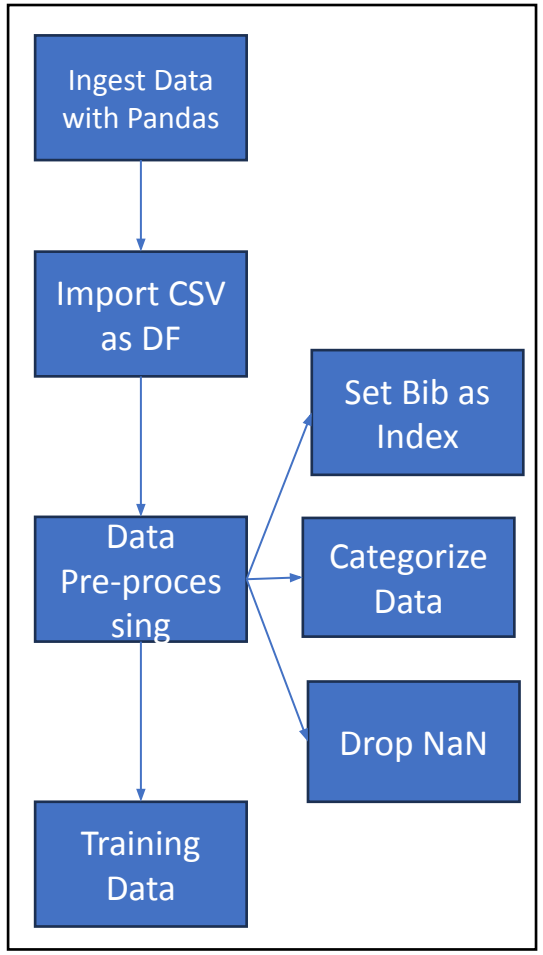
- Individual Bib Number
- Age, Age Group, and Gender
- Zip Code
- First Half Split, Second Half Split, and Overall Finish Time (in Seconds)
- Difference Between Second and First Half Splits (in Seconds and as a Percent)

Project Goal

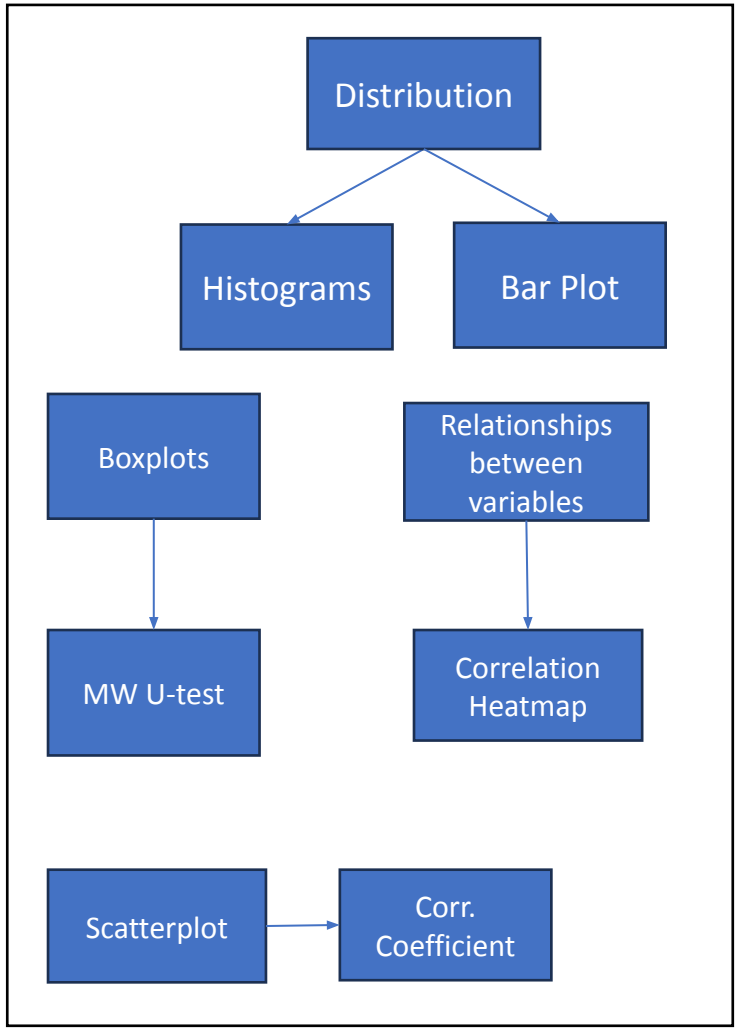
Import the data, explore it, and visualize it to identify potential machine learning approaches that can effectively answer relevant questions about the dataset

Data Processing, Visualization and Potential ML Applications

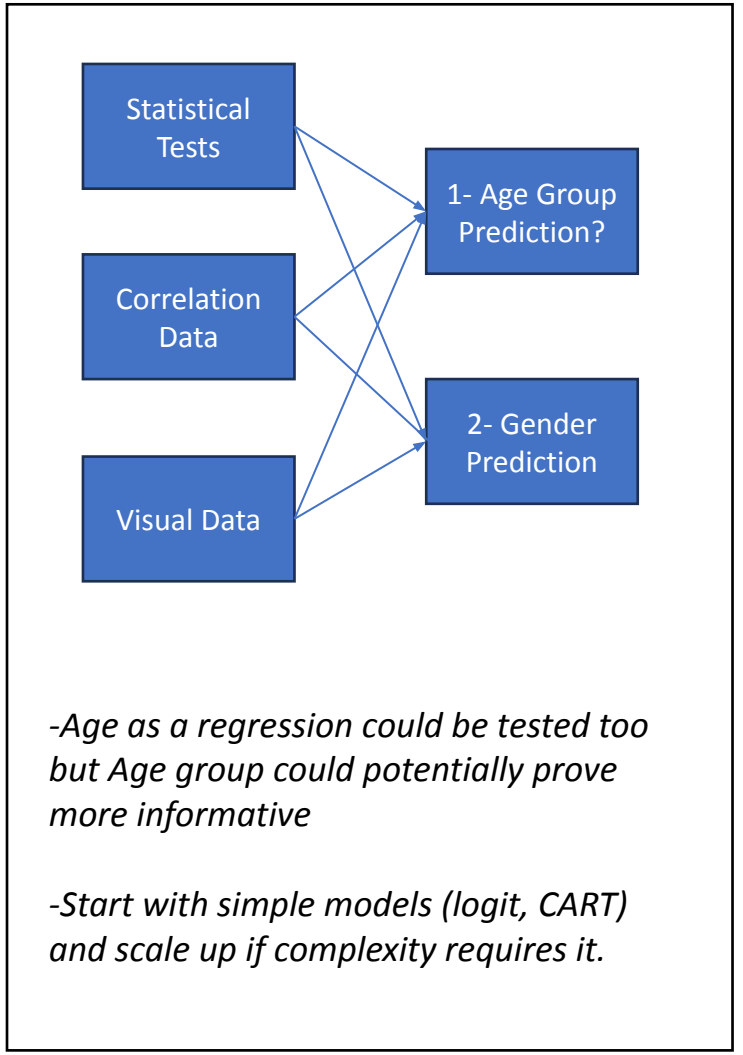
Data Ingestion and Processing



Data Visualization



ML Application



Correlation heatmap reveals levels of association between variables