Project Report

Machine Learning Class Project
November 24, 2024
Abdallah Chidjou

Case Study: Maximization of marketing efficiency with predictive modeling.

Main Objective: The key challenge I aim to address is predicting the likelihood of a customer's acceptance of a marketing offer. This prediction can significantly enhance the bank's resource allocation for future campaigns, maximizing marketing efficiency.

The dataset used in the project was fetched from Kaggle. It's from a financial institution that conducted a marketing campaign to promote its banking products. The goal is to predict whether a customer will subscribe to a term deposit after being contacted.

Description of the Dataset: The dataset contains various attributes about customers

Categorical Feature: Include the following; ('job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome', 'y')

Numerical Feature: Include the following; ('age', 'balance', 'day', 'duration', 'campaign', 'pdays', 'previous')

Target Variable:
The target variable here, is ('y'); indicates term deposit subscription (Yes/No).
The dataset shape is of (4521, 17)

Detail description:

Default
Description: Indicates whether the customer has credit in default (yes or no).
Significance: Customers with a history of defaults might be less likely to subscribe to additional financial products

Campaign
Description: Number of contacts performed during this campaign.
Significance: More frequent contacts could increase awareness, but there may also be a saturation point where additional contacts become ineffective.

Previous
Description: Number of contacts made before this campaign.
Significance: Indicates how persistent past attempts were, which may impact a customer's current decision.

Duration
Description: Duration of the last contact in seconds.
Significance: The length of contact has shown to be a strong indicator of whether a customer is interested in the product, with longer calls being more likely to lead to positive outcomes.

Poutcome
Description: Outcome of the previous marketing campaign (e.g., success, failure, unknown).
Significance: Past success or failure could influence a customer's perception of the product and their likelihood of subscribing.

y (Target Variable)
Description: Whether the customer subscribed to a term deposit (yes or no).
Significance: This is the target variable we want to predict using the other features in the dataset.

Data Cleaning Process: The dataset was free of missing values. However, when checking for summary statistics, I discovered some outliers that I was able to correct using transform log and interquartile range. I then checked for imbalance and moved on to label-encoding the categorical features.
After my inspection, the dataset appeared imbalanced. To correct this, I applied the smote method, which was quite efficient in balancing the two labels of my target variable. My train-test split was in the 80-20 range.

My model approach included Logistic Regression, Random Forest, Support Vector Machine, K-nearest neighbors, Gradient Boosting, and Neural Networks. After using Smote to balance the dataset, Random Forest gave me the highest accuracy of 88%. Looking at the confusion matrix for each model, the (yes class) from the target variable is not doing well.

To improve the model, I tried downsampling the majority class (no label) using RandomUnderSampler. This time, the confusion matrix improved for the (yes class) after fitting the models. However, the accuracy level dropped slightly for each model, and the Neural Network model gave the highest accuracy of 81%.

After extraction for feature performance, features like 'duration' (the length of contact) seem highly influential in determining customer response, suggesting that longer contact times correlate with a higher likelihood of success. It means that the customer is showing interest.

I obtained promising results by processing the dataset and using several models, but I have room for improvement. The bank should focus its marketing efforts on customers most likely to subscribe, reducing costs and improving the efficiency of campaigns. The 'duration' feature, which represents the length of contact, is a critical factor in determining customer response, suggesting that longer contact times correlate with a higher likelihood of success. (Customer showing interest)