Banking Campaign Output Prediction

Ahmad Kamal Baig

Abstract

This report presents a comprehensive analysis of a predictive modeling project aimed at forecasting client responses to bank marketing campaigns. Utilizing a dataset comprising various client attributes, contact information, and economic indicators, we employ machine learning algorithms to predict the likelihood of clients subscribing to a term deposit. The study encompasses data preprocessing, including handling missing values, duplicates, and outliers, feature engineering, and the application of logistic regression and random forest classifiers. Through rigorous evaluation, we demonstrate the effectiveness of these methodologies in enhancing prediction accuracy, contributing valuable insights for optimizing marketing strategies in the banking sector.

1 Introduction

The banking industry increasingly relies on data-driven decision-making to enhance customer engagement and service delivery. In order to anticipate client behavior and tailor marketing efforts accordingly, the industry can leverage the power of predictive modeling to foresee customer's behaviour. In this report, we use machine learning predictors to uncover patterns of customer engagement and predict their responses to term deposit marketing campaigns. This information can act as actionable intelligence to refine marketing initiatives and bolster deposit subscription rates.

2 Data Processing

2.1 Dataset

The dataset used in this project is centered around the direct marketing campaigns of a banking institution, primarily conducted through phone calls. It consists of **40k rows** and **21 columns**. The columns are a mix of categorical and numerical attributes. In particular, they correspond to:

1. **Bank Client Data**
2. **Last contact data in regards to campaign**
3. **Social and economic attributes**
4. **Client subscription to a term deposit** (target variable)

2.2 Missing Values

The dataset contained **10,700 rows** having one of the column value as "**unknown**". Since the total count of dataset was relatively big, we decided to move forward with removal of such entries.

2.3 Duplicate Values

The dataset was relatively low on number of duplicates, **10** to be particular. Hence, we decided to remove them as well.

2.4 Outliers

Utilizing the **IQR method**, values lying outside the first and third quartiles were considered outliers and removed respectively. This method prevents extreme values from skewing the models' training and predictions.Around **7000 outliers** were removed in the process.

2.5 Label and One-hot Encoding

Categorical variables within the dataset are encoded to transform them into a format that can be provided to machine learning algorithms.

Label encoding is applied to binary categorical variables (e.g., "yes" or "no" responses), converting them into 0s and 1s.

One-hot encoding is applied to non-binary categorical variables, creating a separate binary column for each category.

2.6 Feature Pruning

Certain features not relevant to the predictive model or that could introduce bias are removed. For example, the 'duration' column is dropped as it should not be used in a realistic predictive model (it's not known before a call is made).

2.7 Dataset Balancing

The dataset was initially unbalanced with roughly 80-20 ratio of target variable. The Synthetic Minority Over-sampling Technique (**SMOTE**) is used to address such class imbalance in the dataset. This technique generates synthetic samples for the minority class, aiming to balance the distribution between classes. As a result, only **2312 entries** of class 0 were raised to match **21462 entries** of class 1.

2.8 Standardization

Numerical features are standardized to have a mean of 0 and a standard deviation of 1. This is important for models that are sensitive to the scale of the data, such as logistic regression and support vector machines.

2.9 Correlation Heatmap

Correlation matrix showed us that almost all features have moderate relationship with the target variable, therefore we decided not to exclude any of the feature.

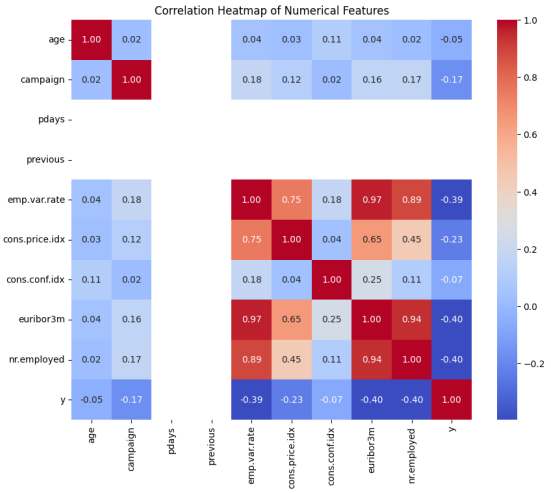

Figure 1: Correlation Heatmap

3 Modeling

3.1 Logistic Regression

- **Overview:** Logistic Regression is a fundamental statistical approach used for binary classification problems. It models the probability of a binary outcome based on one or more predictor variables.
- **Reason for choosing:** Logistic regression has been chosen because of its Interpretability, efficiency and performance.
- **Accuracy**: The logistic regression model achieved an overall accuracy of **86%**, indicating correct predictions for the majority of cases.

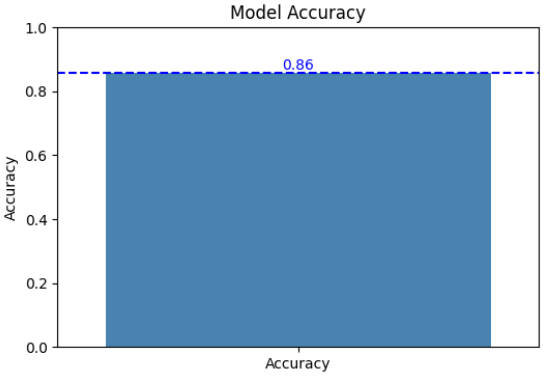

Figure 2: Logistic Regression Accuracy

- **Confusion Matrix**: The model correctly predicted non-subscription and subscription outcomes for 3796 and 3565 cases respectively, while incorrectly predicting 559 non-subscriptions and 665 subscriptions. This shows some room for improvement.

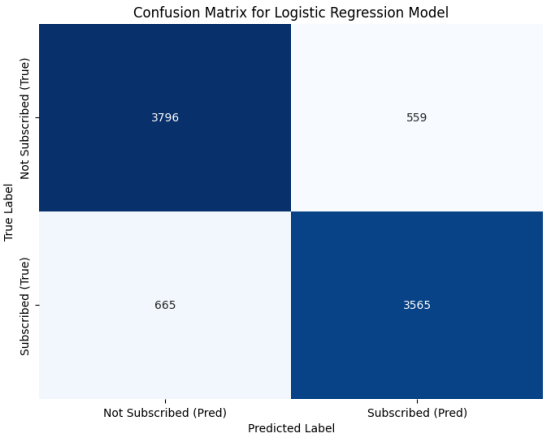

Figure 3: Linear Regression Confusion Matrix

- **Classification Report**: Precision, recall, and F1-score for non-subscriptions are 0.85, 0.87, and 0.86 respectively; for subscriptions, they are 0.86, 0.84, and 0.85 respectively. This indicates that there is a good balance between precision and recall.

3.2 Random Forest

- **Overview:** The Random Forest Classifier is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees.

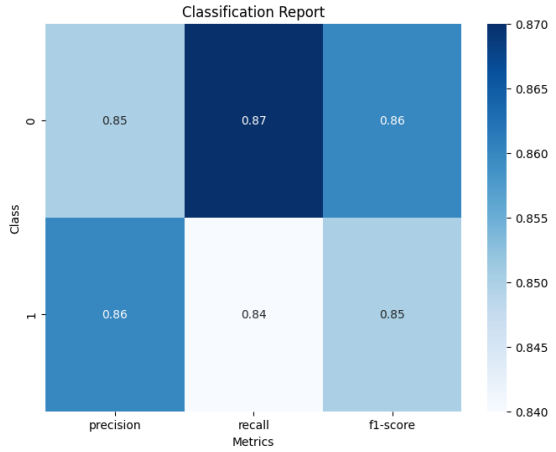

Figure 4: Linear Regression Classification Report

- **Reason for choosing:** Random Forest has been chosen due to its robustness towards overfitting, ability to represent feature importance as well as handling non-linearity.

- **Accuracy**: The model exhibits a high accuracy of **93%**, indicating that it correctly predicted the subscription status for the majority of clients.

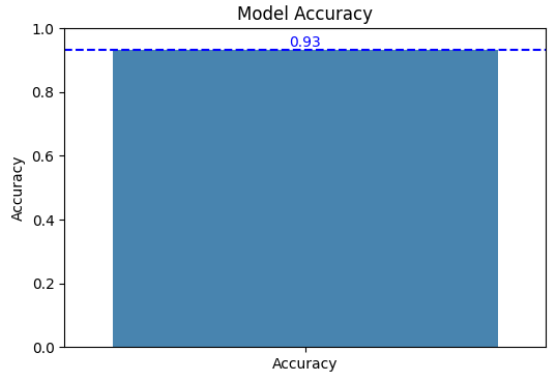

Figure 5: Random Forest Accuracy

- **Confusion Matrix**: The model has successfully predicted 4086 true negatives and 3922 true positives, with a relatively low number of false predictions (269 false positives and 308 false negatives).

- **Classification Report**: Both classes (subscribed and not subscribed) show high precision and recall (0.93 and 0.94 respectively), resulting in a balanced F1-score of 0.93, reflecting the model's robust performance.

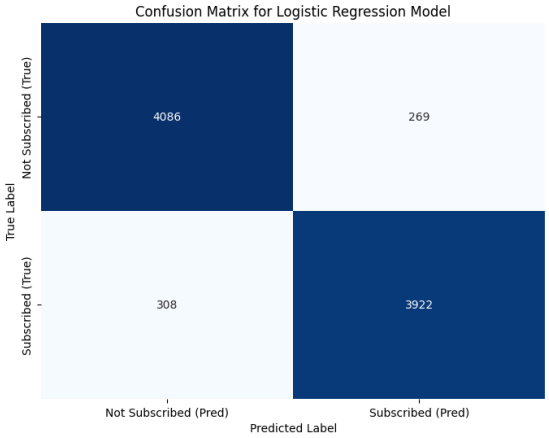

Figure 6: Random Forest Confusion Matrix

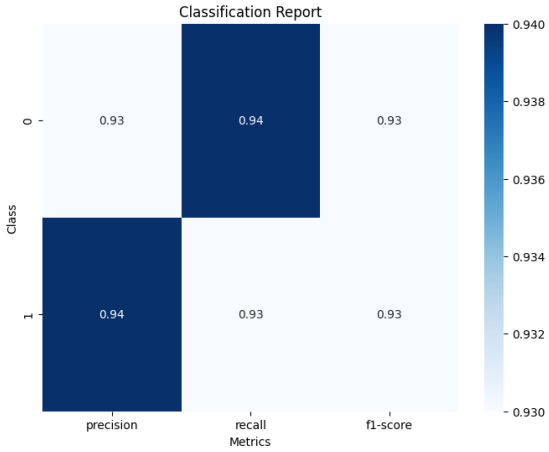

Figure 7: Random Forest Classification Report

4 Conclusion

The main challenge for the project was dealing with raw and unbalanced dataset. Due to this purpose, classification report of both models showed low recall for class 1. After thorough data preprocessing, we were able to balance and clean the dataset successfully.

The Random Forest model outperformed the Logistic Regression model across all metrics. It has a higher accuracy, fewer false predictions, and higher precision and recall scores. These results suggest that the Random Forest model is better suited for this dataset and prediction task, likely due to its ability to capture more complex patterns and interactions between features. It is important to note that while Random Forest shows superior performance, it might also be more computationally intensive and less interpretable than Logistic Regression.