Project Report

Talent Mobility Program - Azubi Africa

by Victor Nyarko Anim

Analysis Overview

The objective was to predict whether a bank customer would subscribe to a term deposit based on data collected during previous campaigns.

The dataset includes attributes such as age, job, marital (marital : marital status), education, default (credit in default), balance (average yearly balance), housing loan, personal loan, contact (contact communication type), day (last contact day of the month), month (last contact month of year), duration (last contact duration), campaign (number of contacts performed during this campaign and for this client), pdays (number of days that passed by after the client was last contacted from a previous campaign), previous (number of contacts performed before this campaign and for this client), poutcome (outcome of the previous marketing campaign)

There were no missing and duplicated values in the dataset.

The following observations were made

1) Majority of bank's customers are between the ages of 20 and 60 years old

2) The Jobs of most bank's customers are in the field of Management, Technicians, blue collar, admin and services, these professions form the bulk of the bank’s customers

3) Most customers are married

4) Most customers have secondary education

5) The majority of customers have no credit in default

6) The average yearly balance is around 1500 euros

7) Most customers have housing loans

8) Most customers don't have personal loans

9) Most customers were last contacted via cellular communication

10) Most customers were last contacted in the month of May and June

11) The duration of last contact is around 250 seconds, which means most customers were contacted within the last 4 minutes

12) The outcome of the previous campaign was mostly unknown

13) Most customers had not subscribed to the previous campaign or term deposit

14) Most of the numerical variables have outliers in their distributions

Methods Used

1. Data Preprocessing:

- Categorical features (e.g., job, marital, education) were encoded using a OneHotEncoder encoder to ensure compatibility with the model, SimpleImputer strategy of most_frequent was used handle any missing categorical values.

- Numerical features (e.g., age, balance) were normalized using StandardScaler

to improve model performance, SimpleImputer strategy of median was used handle any missing numerical values.

- 'unknown' values were retained as categories where necessary to reflect real-world scenarios.

- Most numerical features contained outliers and to ensure better model performance, outliers were treated

- The dependent feature (y) was highly imbalance, so SMOTE was used to eliminate bias in the model

2. Model Selection:

- Multiple machine learning models were evaluated, including XGBoost, Gradient Boosting, Logistics Regression, Support Vector Machine, Decision Tree, K-Nearest Neighbors and Random Forest.

- Pipelines were pre-trained and loaded dynamically to streamline the prediction process.

3. Prediction Framework:

- A web-based user interface built using Streamlit to allow users to input customer data and receive predictions.

- Predictions logged to a history file for review and analysis of model behavior over time.

4. Metrics Evaluation:

- Performance metrics for the models (such as accuracy, precision, recall, F1 score, and AUC) were evaluated during the training and evaluation phase.

- XGBoost emerged as the best-performing model based on the validation dataset, with an F1 score of 0.897976 and an AUC of 0.93.

- Random Forest and Gradient Boosting were next best performing models, with an F1 score of 0.898767 and 0.896514, and an AUC of 0.92 and 0.92 respectively.

Insights Drawn

1.Key Predictors:

- pdays, previous and job are the most important features in the performance of model
- age, day, and age_group are moderately important features.

2. Customer Profiles Likely to Subscribe:

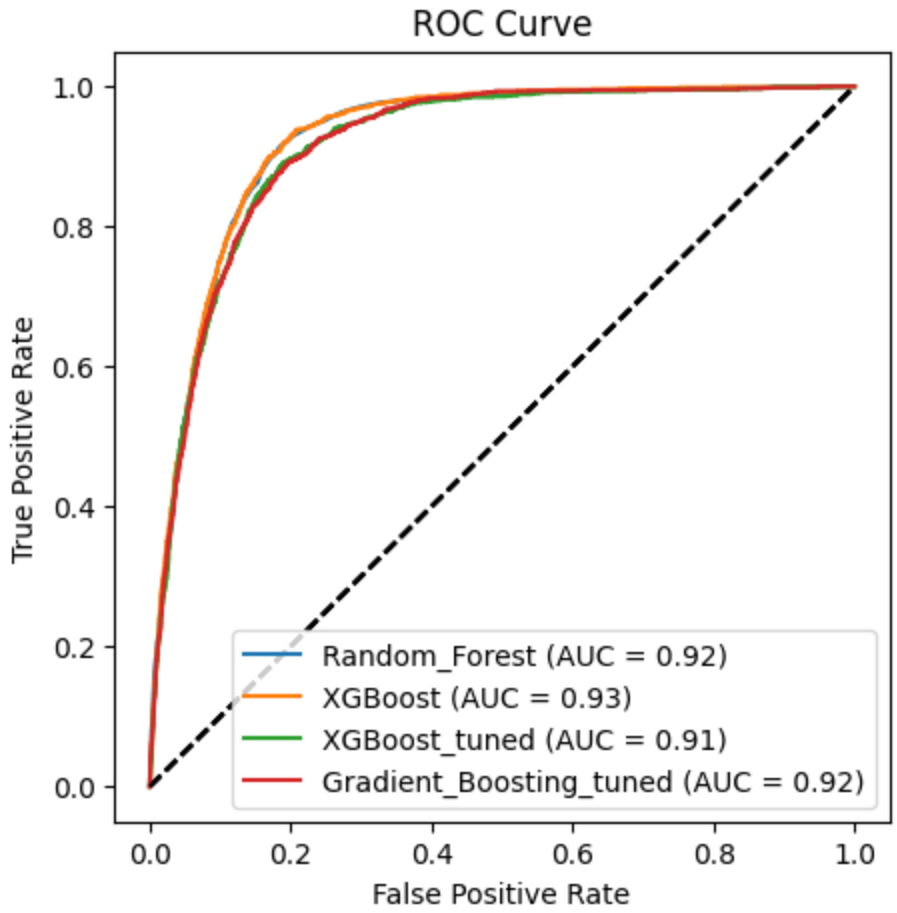
- Customers with higher account balances and longer durations of last contact showed higher subscription probabilities.
- Past successful campaign outcomes strongly correlated with future subscriptions.

3. Business Recommendations:

- Focus marketing efforts on customers who previously responded positively to campaigns.
- Extend the duration of meaningful customer interactions, as it appears to significantly influence outcomes.

Model's Performance Metric

model_name	**accuracy**	**precision**	**recall**	**f1-score**	**AUC**
Random_Forest	0.899850	0.897783	0.899850	0.898767	0.92
XGBoost	0.898876	0.897142	0.898876	0.897976	0.93
Gradient_Boosting_tuned	0.894895	0.898343	0.894895	0.896514	0.92



These results indicate the model is highly effective at distinguishing between subscribers and non-subscribers, making it a reliable tool for targeted marketing efforts.