Comparing Classifiers: Bank Term Deposit Subscription Prediction

Overview

This project aims to predict whether a client will subscribe to a bank term deposit using a bank marketing dataset. The analysis involves data cleaning, exploratory data analysis (EDA), feature engineering, feature selection, and classifier evaluation. A key challenge addressed in this project is the highly imbalanced target variable.

Data Preprocessing

- **Dataset:**
  The dataset (initially 41,188 rows × 21 columns) contains information such as age, job, marital status, education, and campaign details. After removing 12 duplicate rows, the final cleaned dataset comprises 41,176 rows.

- **Cleaning Steps:**

  - **Duplicates:** Removed duplicate entries.
  - **Handling 'unknown' Values:**
    - Categorical columns with 'unknown' entries were imputed using their respective mode values (e.g., replacing 'unknown' in `job` with "admin.", in `marital` with "married", etc.).
  - **Data Quality Checks:**
    - Verified the `age` column to ensure values fall within a realistic range ([0, 100]).

Exploratory Data Analysis (EDA)

- **Target Variable Distribution:**

  - The target (`y`) indicates whether a client subscribed to a term deposit.
  - The distribution is highly imbalanced, with a majority of clients labeled as "no" and a minority as "yes."
- **Categorical Variable Analysis:**

- Visualizations (using count plots) for variables such as `job`, `marital`, `education`, `contact`, and `poutcome` revealed dominant categories:
    - **Job:** "admin.", "blue-collar"
    - **Marital:** "married"
    - **Contact:** "cellular" is the primary communication channel.
- Insights suggest that first-contact effectiveness is crucial as many customers are new to bank marketing efforts.

- **Continuous Variable Analysis:**

  - Histograms and boxplots for variables (e.g., `age`, `duration`, `campaign`, `pdays`) indicate:
    - **Call Duration:** Longer durations correlate with higher subscription rates.
    - **Euribor3m:** Lower three-month Euro Interbank Offered Rate values tend to be associated with subscriptions.
  - Many clients have few campaign contacts, with numerous cases showing placeholders (like `999`) indicating no previous contact.

Feature Engineering & Preprocessing

- **Engineered Features:**
  Based on EDA insights, several new features were created:

  - `target_demographic`: Flag for key job categories (e.g., "admin.", "blue-collar", "technician") for married clients.
  - `is_cellular`: Indicates if the contact method is cellular.
  - `prev_success`: Flag indicating if a previous campaign was successful.
  - `middle_age`: Flag for customers aged between 30 and 50.
  - `long_duration`: Flag for call durations longer than 200 seconds.
  - `new_contact`: Flag indicating new contacts (using a `pdays` value of 999).
  - `high_education`: Flag for customers with higher education credentials.
  - `job_marital`: Interaction feature combining job and marital status.
- **Data Transformation:**

  - Target variable `y` was label-encoded ("yes" → 1, "no" → 0).
  - Categorical features were one-hot encoded, and numerical features were standardized using a Standard Scaler.
- **Train-Test Split:**
  The data was split into 80% training and 20% testing sets.

Feature Selection

Two methods were used to identify the most informative features:

1. **SelectKBest (Mutual Information):**
   Top features included: `age`, `duration`, `pdays`, `previous`, `emp.var.rate`, `cons.price.idx`, `cons.conf.idx`, `euribor3m`, `nr.employed`, and `is_cellular`.

2. **Random Forest Feature Importances:**
   Key features identified were: `duration`, `euribor3m`, `age`, `nr.employed`, `long_duration`, among others.

Modeling and Evaluation

- **Classifiers Evaluated:**
  The following models were compared:
  - Logistic Regression
  - K-Nearest Neighbors (KNN)
  - Decision Tree
  - Support Vector Machine (SVM)
  - Random Forest
- **Performance Metrics:**
  Models were evaluated using accuracy, confusion matrices, classification reports, and ROC-AUC scores.
  - **Example Findings:**
    - **Logistic Regression:**
      - Accuracy ≈ 90.3%
      - ROC-AUC ≈ 0.9155
      - Notable drop in performance for the minority class (lower recall and F1-score).
    - **Random Forest:**
      - Best ROC-AUC performance (≈ 0.9403), indicating superior discrimination ability.
- **Handling Class Imbalance:**
  To improve performance for the minority class, a RandomUnderSampler was applied to balance the training data. This approach led to enhanced detection metrics (precision, recall, F1-score) for the minority class, although the models still exhibited a trade-off between overall accuracy and minority class detection.

Conclusions

- **Key Insights:**

    - **Longer Calls & Lower Euribor3m Rates:** Clients with longer call durations and lower Euribor3m rates are more likely to subscribe.
    - **Class Imbalance:** A significant imbalance in the target variable affects model performance, with most models favoring the majority class.
    - **Feature Engineering:** The introduction of new features based on domain insights (such as `target_demographic` and `job_marital`) provided additional predictive power.
- **Modeling Outcome:**
  While most classifiers achieved high overall accuracy, the Random Forest model delivered the highest ROC-AUC score, suggesting its robustness in this scenario. However, the imbalance in the data necessitates further techniques—such as oversampling or cost-sensitive learning—to better capture the minority class characteristics.

- **Future Directions:**
  Future work could involve:

    - Experimenting with different resampling techniques (e.g., SMOTE, oversampling) or ensemble methods tailored to imbalanced datasets.
    - Hyperparameter tuning for the Random Forest and other models to further improve performance, especially for the minority class.