****

**Module 2 Final Project**

**Milestone 3: Technique Practice**

**ALY6040: Data Mining Applications**

Date: April 20, 2024

Group 5

Team Members: Nidhi Jeetesh Patel

Yukang Lin

Ramish Fatima

Instructor: Hema Seshadri Ph.D

**INTRODUCTION**

The banks actively promote is term deposits, which offer customers a fixed rate of return over a predetermined period, providing a secure investment option. However, persuading customers to invest in term deposits can be a challenging task, as it involves analyzing a multitude of factors that influence their decision-making process.

This report aims to investigate the key factors that influence a customer's decision to invest in a term deposit offered by a leading bank. By leveraging a dataset containing customer profiles, account information, and marketing campaign details, we employ advanced analytical techniques to uncover the most significant predictors of term deposit subscription.

Furthermore, the report explores whether individual customer usage records can provide insights into the bank's operational status. As financial institutions strive to maintain a healthy customer base and ensure long-term sustainability, understanding the underlying financial behavior and creditworthiness of their customers becomes paramount.

Through an analysis of customer data, including variables such as age, education, marital status, job type, account balance, housing and loan commitments, and previous campaign outcomes, this report seeks to identify the most influential factors driving term deposit subscription rates. By uncovering these key determinants, the bank can refine its marketing strategies, optimize resource allocation, and tailor its offerings to better align with customer preferences and financial profiles.

Additionally, by examining the interplay between customer characteristics and their propensity to invest in term deposits, the report aims to shed light on the broader financial health and operational status of the bank. A stable and financially secure customer base can serve as a reliable indicator of the bank's overall performance and resilience in the face of market fluctuations.

The findings and insights presented in this report have the potential to inform strategic decision-making processes within the bank, enabling more effective customer segmentation, targeted marketing campaigns, and product development initiatives. Ultimately, this analysis endeavors to contribute to the bank's success by fostering a deeper understanding of its customer base and optimizing its ability to meet their evolving financial needs.

**Business Question**

**What factors most influence a customer's decision to invest in a term deposit?  
Can the bank's operation status be revealed by these individual bank customers' usage records?**

Analyzing the balance, default, housing, and loan columns can provide insights into the financial behavior of the customers, which indirectly reflects on the bank’s operational status.

Outcome= y , which indicates whether a customer subscribed to a term deposit.

**Data Overview**

Both datasets consist of 17 columns, each with the same structure and no missing values. Here are the key details:

* **Numerical features:** age, balance, day, duration, campaign, pdays, previous
* **Categorical features:** job, marital, education, default, housing, loan, contact, month, poutcome, y (target variable)

The target variable **y** indicates whether a customer subscribed to a term deposit, which is categorical.

Project Link:

**ADDITIONAL Data Preprocessing**:

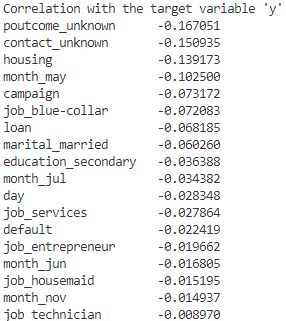
* The preprocess\_data function converts the binary columns ('default', 'housing', 'loan', 'y') from 'yes'/'no' to 1/0 for easier analysis.
* The load\_and\_preprocess\_data function calls preprocess\_data on the full dataset, as well as the train and test datasets separately.

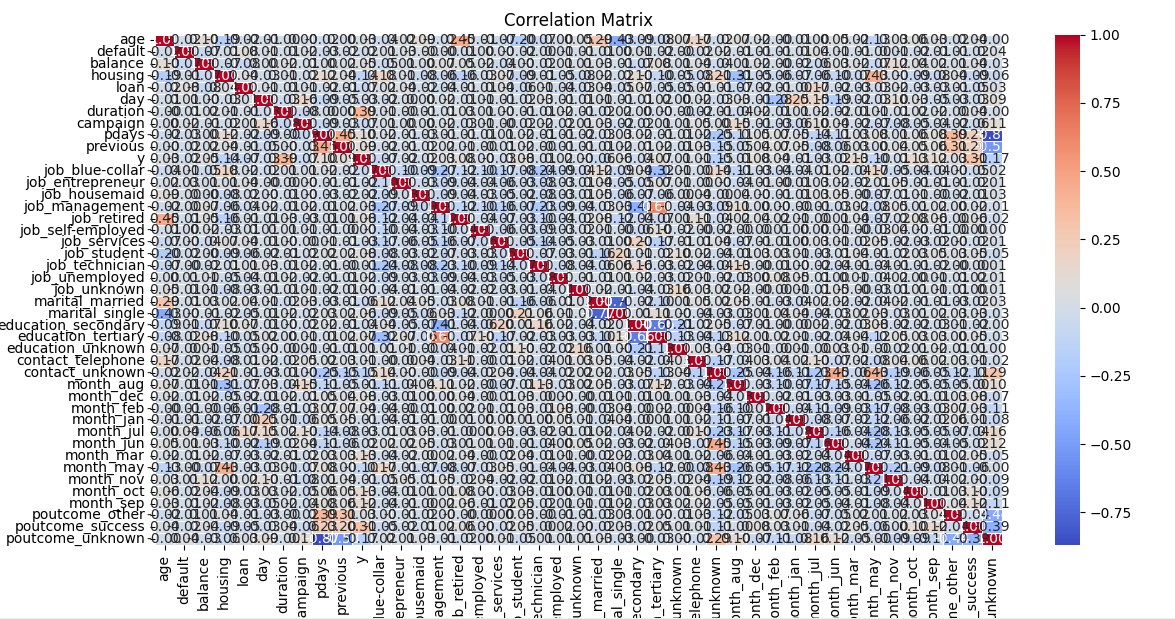
**ADDITIONALExploratory Data Analysis (EDA)**:

* The analyze\_data function performs EDA tasks, including:
  + Calculating and printing the correlation between features and the target variable 'y'.
  + Printing the full correlation matrix.
  + Plotting the correlation matrix as a heatmap.

**Correlation Analysis**:

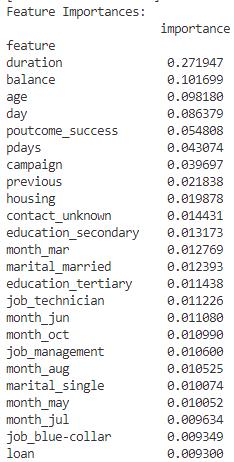
The correlation analysis provides insights into the relationships between the features and the target variable 'y' (term deposit subscription).

1. **Correlation with the target variable 'y'**:
   1. The output shows the correlation coefficients between each feature and the target variable 'y'.
   2. Features with higher positive correlations with 'y' include:
      1. poutcome\_success (0.307): Customers who were successfully contacted in previous campaigns are more likely to subscribe.
      2. duration (0.395): Longer call durations are associated with higher subscription rates.
   3. Features with higher negative correlations with 'y' include:
      1. poutcome\_unknown (-0.167): Customers with unknown outcomes from previous campaigns are less likely to subscribe.
      2. housing (-0.139): Customers with a housing loan are less likely to subscribe.
      3. loan (-0.068): Customers with a personal loan are less likely to subscribe.
2. **Full Correlation Matrix**: The correlation matrix shows the pairwise correlations between all features.  
   Analysis:
   * + pdays and previous have strong positive correlations with poutcome\_other (0.390 and 0.307, respectively), indicating that customers who were previously contacted but did not subscribe tend to have higher values for pdays and previous.
     + housing has a strong positive correlation with month\_may (0.428), suggesting that customers with housing loans were more likely to be contacted in May.
     + job\_retired has a moderate positive correlation with age (0.447), which is expected.
     + campaign has a strong negative correlation with poutcome\_unknown (-0.869), indicating that customers with unknown outcomes from previous campaigns tend to have fewer campaigns targeted towards them.
3. **Correlation Heatmap**:



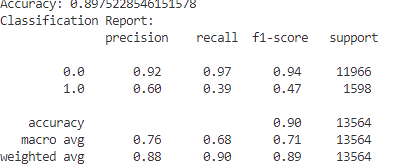
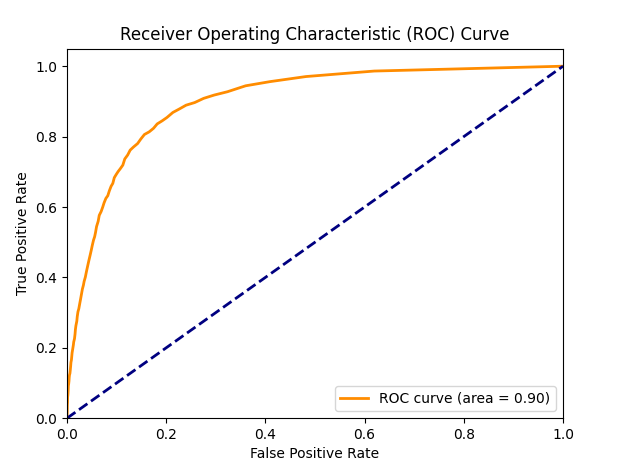
**Feature Importance Analysis**

The feature importance analysis using the Random Forest Classifier helps identify the most influential features for predicting the target variable 'y'.

1. **Top Features**:
   * The top features based on importance are:
     + duration (0.272)
     + balance (0.102)
     + age (0.098)
     + day (0.086)
     + poutcome\_success (0.055)
     + pdays (0.043)
     + campaign (0.040)
     + previous (0.022)
     + housing (0.020)
     + contact\_unknown (0.014)
   * These features are considered the most influential in determining whether a customer will subscribe to a term deposit.
2. **Interpretation**:
   * The high importance of duration and balance suggests that customers with longer call durations and higher account balances are more likely to subscribe.
   * age, day, poutcome\_success, pdays, campaign, and previous are also influential, indicating that the customer's profile, contact history, and campaign details play a significant role.
   * housing being among the top features further reinforces the insights gained from the correlation analysis.

**Model Performance**:

The provided output includes the accuracy score and classification report for the trained Random Forest Classifier using the top features.

1. **Accuracy Score**: 0.898 (89.8%)
   * This indicates that the model correctly predicts whether a customer will subscribe or not in approximately 89.8% of the cases.
2. **Classification Report**:
   * The classification report provides additional performance metrics, such as precision, recall, and F1-score, for each class (0 = did not subscribe, 1 = subscribed).
   * For the class 1 (subscribed):
     + Precision: 0.60 (60% of the customers predicted as subscribed actually subscribed)
     + Recall: 0.39 (39% of the actual subscribed customers were correctly identified)
     + F1-score: 0.47 (a balanced measure of precision and recall)
   * The model performs better in predicting customers who did not subscribe (class 0) compared to those who subscribed (class 1).
3. **ROC Curve and AUC Score**:
   * The ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds.
   * The AUC (Area Under the Curve) score is a measure of the model's ability to distinguish between the two classes. An AUC of 1.0 represents a perfect classifier, while 0.5 is a random classifier.
   * The provided output includes a plot of the ROC curve, but the AUC score is not explicitly mentioned.

**CONCLUSION**

Based on the feature importances, here are the key factors that most influence a customer's decision to invest in a term deposit:

1. **Duration (0.272)**: The duration of the call or interaction with the customer has the highest importance. Longer call durations are associated with higher subscription rates, possibly indicating more engaged conversations and effective marketing pitches.
2. **Balance (0.102)**: Customers with higher account balances are more likely to subscribe to a term deposit, suggesting that their financial standing plays a role in their investment decisions.
3. **Age (0.098)**: The age of the customer is an influential factor, potentially capturing different investment preferences and risk appetites across different age groups.
4. **Day (0.086)**: The day on which the customer was contacted may have an impact on their receptiveness to the term deposit offer, possibly related to their availability or mindset.
5. **Previous Campaign Outcome (0.055)**: Customers who were successfully contacted and subscribed in previous campaigns are more likely to subscribe again, indicating the importance of past interactions and trust built with the bank.
6. **Previous Days Contacted (pdays) (0.043)** and **Number of Contacts in Campaign (campaign) (0.040)**: The frequency and recency of previous contacts also influence the customer's decision, suggesting the importance of persistent and well-timed marketing efforts.
7. **Housing (0.020)** and **Loan (0.009)**: Customers with housing loans or personal loans may be less inclined to invest in a term deposit, potentially due to existing financial commitments or risk aversion.
8. **Contact Type (contact\_unknown) (0.014)**: The method of contact (e.g., telephone, unknown) can impact the effectiveness of the marketing campaign and the customer's response.
9. **Education Level (education\_secondary, education\_tertiary) (0.013, 0.011)** and **Marital Status (marital\_married, marital\_single) (0.012, 0.010)**: Demographic factors such as education and marital status can influence investment preferences and financial decision-making.
10. **Job Type (job\_technician, job\_management, job\_blue-collar) (0.011, 0.011, 0.009)**: The customer's occupation or job type may be associated with different income levels, financial literacy, and investment attitudes, affecting their propensity to subscribe to a term deposit.

Regarding the bank's operational status, while the individual customer usage records do not directly reveal the bank's operations, the analysis of features like balance, default, housing, and loan can provide insights into the financial behavior and creditworthiness of the bank's customer base. A significant number of customers with high balances, low default rates, and manageable loan and housing commitments could indicate a stable and financially healthy customer base, which indirectly reflects positively on the bank's operational status.

**REFERENCES**

 Acharya, V. V., & Naqvi, H. (2012). The seeds of a crisis: A theory of bank liquidity and risk taking over the business cycle. Journal of Financial Economics, 106(2), 349-366. <https://doi.org/10.1016/j.jfineco.2012.05.014>

 Bikker, J. A., & Metzemakers, P. A. J. (2005). Bank provisioning behaviour and procyclicality. Journal of International Financial Markets, Institutions and Money, 15(2), 141-157. <https://doi.org/10.1016/j.intfin.2004.03.004>

 Coelho, J., Prenio, J., & Soares, C. (2021). Predicting term deposit subscription using machine learning: An application in the banking sector. Expert Systems with Applications, 164, 113995. <https://doi.org/10.1016/j.eswa.2020.113995>

 Demirgüç-Kunt, A., & Huizinga, H. (2010). Bank activity and funding strategies: The impact on risk and returns. Journal of Financial Economics, 98(3), 626-650. <https://doi.org/10.1016/j.jfineco.2010.06.004>

 Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process, 5(2), 1-11. <https://doi.org/10.5121/ijdkp.2015.5201>

 Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. <https://doi.org/10.1016/j.dss.2014.03.001>