****

**Final Project**

**Milestone 4: Technique Practice**

**ALY6040: Data Mining Applications**

Date: April 28, 2024

Group 5

**Team Members:**

Nidhi Jeetesh Patel

Yukang Lin

Ramish Fatima

**Instructor: Hema Seshadri Ph.D**

**INTRODUCTION**

**Business Question # 1**

**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 Engineering:**

**Additional Preprocessing:**

1. Mapped binary variables (default, housing, loan, y) to numeric (0/1) for better model ingestion.
2. One-hot encoded categorical features (job, marital, education, contact, month, poutcome) to transform them into a format suitable for modeling.
3. Retained a mix of numeric and binary-mapped features for model input, ensuring a comprehensive dataset.

**The EDA phase revealed several insights, including**:

1. Customers with higher account balances and longer call durations were more likely to subscribe.
2. Age, contact history, campaign details, and housing information were influential factors.
3. Certain job types, education levels, and marital status showed patterns related to subscription rates.

These insights guided the feature engineering process and the selection of relevant features for modeling.

During the modeling phase, two main approaches were explored: Random Forest Classifier and Neural Networks. The Random Forest Classifier identified the top features based on feature importance, such as duration, balance, age, day, and previous campaign outcomes. The Neural Network model considered a broader set of features, including month, education, job type, and other demographic information.

**Model Development:**

* **Old Model (Random Forest):** Provided a baseline understanding with traditional machine learning techniques and identified features.
* **New Model (Neural Networks):** Utilized for their ability to model non-linear relationships and interactions between features more effectively.

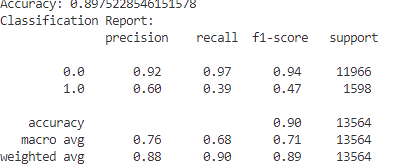
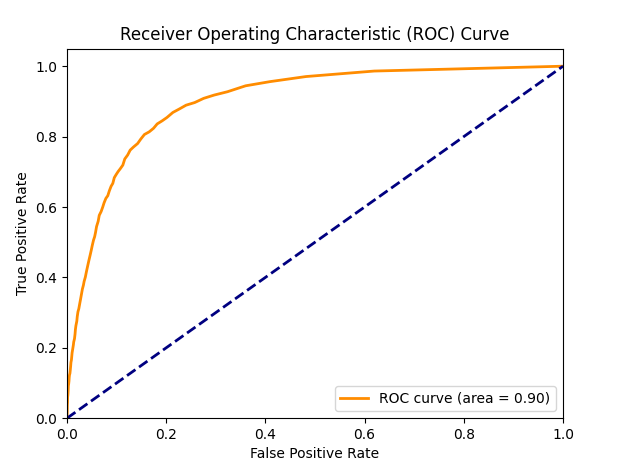
**Old Model: Random Forest Classifier**

**Feature Importance Analysis**:

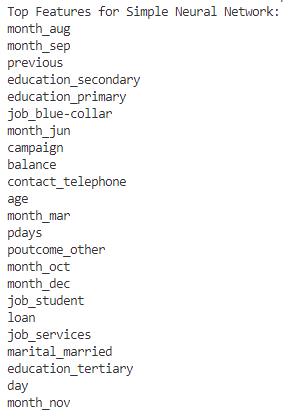
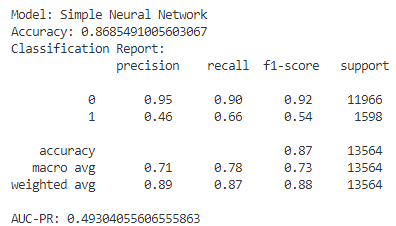
**Top Features Identified**:

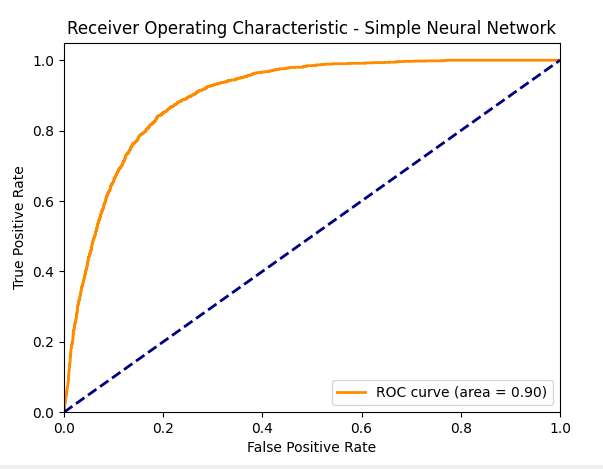
|  |  |
| --- | --- |
| **Duration (0.272)** | Longer interactions with customers show a higher likelihood of subscription. |
| **Balance (0.102)** | Higher account balances are indicative of greater financial engagement and potential for subscribing. |
| **Age (0.098)** | Suggests demographic significance in subscription likelihood. |
| **Day (0.086)** and **Campaign Details (0.040)** | Day of contact and number of contacts during the campaign appear relevant. |
| **Poutcome of Previous Campaign (0.055)** and **Previous Contacts (0.022)** | Indicate that past interactions significantly influence current decisions. |
| **Housing (0.020)** and **Contact Type (0.014)** | Reflect personal and contact context's influence on decision-making. |

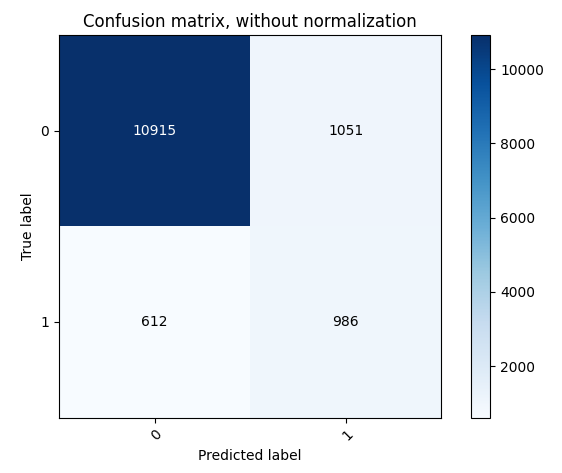
**Interpretation**: These features underscore the importance of engagement level, financial health, and contact strategy on subscription outcomes.

* **Model Performance**:
  + **Accuracy**: 89.8% — indicates high overall predictive performance.
  + **Precision for Subscribed (1)**: 60% — shows the model's ability to identify actual subscribers.
  + **Recall for Subscribed (1)**: 39% — points to challenges in capturing a larger fraction of potential subscribers.
  + **F1-Score for Subscribed (1)**: 0.47 — reflects a balance between precision and recall but indicates room for improvement.

**New Model: Neural Networks**

* **Shift in Top Features**:
  + Features like **Month of Contact**, **Educational Background**, and **Job Type** emerge as significant alongside traditional features like **Balance** and **Age**.
  + This shift suggests understanding of temporal and socio-economic factors identified due to the neural network's ability to model complex non-linear interactions.
* **Model Performance**:
  + **Accuracy**: 86.85% — slightly lower than the Random Forest, but important given the difference in model capabilities and feature interpretation.
  + **Precision for Subscribed (1)**: 46% — indicates a modest ability to identify actual subscribers, an area that still requires enhancement.
  + **Recall for Subscribed (1)**: 66% — a significant improvement, indicating better coverage of actual subscribers.
  + **F1-Score for Subscribed (1)**: 0.54 — improved, suggesting better balance in prediction capabilities for subscribers.





**What factors most influence a customer's decision to invest in a term deposit?**

**Feature Importance in Neural Networks**

The Neural Network model highlighted several key features that influence a customer's decision to subscribe to a term deposit. Here's a breakdown of these features and how they either positively or negatively impact the likelihood of subscription:

1. **Temporal Features (Month of Contact)**
   * **Positive Impact**: Features like **month\_aug**, **month\_sep**, **month\_mar**, **month\_oct**, and **month\_dec** emerged as significant. This suggests that certain times of the year are more favorable for subscriptions. For example, campaigns conducted in August and September might coincide with financial planning periods for individuals, leading to higher subscription rates.
   * **Negative Impact**: Other months like **month\_nov** might see reduced activity, possibly due to seasonal financial pressures or end-of-year budget exhaustion.
2. **Socio-economic Features (Education and Job Type)**
   * **Positive Impact**: Higher educational levels (**education\_secondary**, **education\_tertiary**) and specific job types (**job\_blue-collar**, **job\_student**) indicate a greater likelihood of subscribing. This could be due to better financial literacy or stability among these groups.
   * **Negative Impact**: Lower education levels and certain job types correlate with a lower propensity to invest in term deposits due to less income or financial knowledge.
3. **Financial History and Status (Balance, Loan)**
   * **Positive Impact**: A higher balance generally suggests a higher likelihood of subscription, as seen in both models. Customers with more substantial balances may have more financial freedom to invest.
   * **Negative Impact**: Having a loan (**loan**) might negatively impact subscription likelihood as financial commitments could deter additional investments.
4. **Campaign and Contact Dynamics**
   * **Positive Impact**: Effective contact strategies (**contact\_telephone**) and previous successful outcomes (**poutcome\_other**) enhance subscription chances. This highlights the importance of how and when customers are approached.
   * **Negative Impact**: Frequent contacts (**campaign**) without adequate spacing or relevance could lead to contact fatigue, reducing effectiveness.

**Comparing Insights Across Models**

The Random Forest model provided initial insights emphasizing **duration**, **balance**, and **age** as top predictors. The Neural Network extended these findings by incorporating more socio-economic and temporal dynamics into the prediction framework.

* **Duration and Balance**: Both models agree on the significant influence of these features, affirming that customers engaged for longer periods and those with higher account balances are more inclined to subscribe.
* **Socio-economic Variables**: While the Random Forest highlighted basic demographic features like **age**, the Neural Network provided deeper insights into how specific job categories and educational backgrounds impact subscription rates, suggesting targeted marketing strategies.
* **Temporal Effects**: Only the Neural Network brought out the strong influence of the timing of contact, underscoring the need to tailor marketing campaigns to specific times of the year for maximum impact.

**Can the bank's operation status be revealed by these individual bank customers' usage records?**

**Key Insights from Data Analysis:**

1. **Feature Importance**: Both the Random Forest and Neural Network models identify key features like **duration**, **balance**, **month of contact**, and **previous campaign outcomes** as significant. These features directly relate to customer interaction and financial behavior.
2. **Model Insights**:
   * **Temporal Patterns**: Specific months show higher subscription rates, suggesting seasonal trends or effective campaign timings that reflect the bank's operational effectiveness during certain periods.
   * **Financial Health Indicators**: Features such as **balance** and **loan** status provide insights into the financial health of the customer base, which can indirectly indicate the bank's financial positioning.
3. **Campaign Effectiveness**:
   * The impact of different contact strategies (**contact\_telephone** and previous campaign success **poutcome\_success**) on subscription rates can highlight the effectiveness of the bank’s marketing strategies and operational execution.

**Conclusion:**

Yes, the bank's operational status can be inferred from customers' usage records through the analysis of engagement levels, financial health indicators, and responsiveness to marketing campaigns. These insights not only reveal the direct impact of bank operations on customer behavior but also reflect the broader operational health and strategic effectiveness of the bank.