Text

Description automatically generated

**Project Analysis Report**

**Submitted by – Yogesh Siyyadri**

**Student ID - 101164082**

**Course Name – Data Mining for Managers**

**Course Code – DSCI 724**

**Instructor: Dr. Thomas Tiahrt**

**Introduction**:  
This project focuses on developing a predictive deep learning model to forecast client subscription to term deposits, leveraging data from direct marketing campaigns conducted by a Portuguese bank. The primary objective is to analyze client demographics, transactional history, and campaign interactions to predict the likelihood of subscription. By utilizing a deep learning approach, this project aims to identify critical factors influencing client decisions, enabling financial institutions to implement more targeted and effective marketing strategies. The predictive model serves as a tool to understand the intricate relationships between client attributes and their propensity to subscribe, offering actionable insights to optimize resource allocation and campaign effectiveness.

**Problem Description**:  
In the competitive banking industry, the success of direct marketing campaigns plays a crucial role in acquiring new customers and selling financial products such as term deposits. However, predicting which clients are most likely to subscribe remains a significant challenge due to the diverse range of client profiles and behaviors. Banks need a robust mechanism to identify potential subscribers and tailor their marketing efforts accordingly.

This project examines which demographic, and transactional characteristics significantly influence a client’s likelihood to subscribe to a term deposit. Additionally, it explores how predictive modeling can enhance the precision and efficiency of marketing campaigns by identifying and targeting clients with a higher probability of subscription.

The dataset includes various attributes such as client age, job type, marital status, education level, credit history, and interaction details like contact duration and campaign frequency. Each of these factors contributes to the decision-making process of potential subscribers. Through comprehensive analysis, including visual representations like correlation matrices and box plots, the findings of this project will empower financial institutions to maximize customer engagement, improve conversion rates, and allocate marketing resources effectively. The insights derived will serve as a foundation for designing personalized and impactful marketing campaigns, ultimately driving business growth and client satisfaction.

**Dataset Description**:  
The dataset [*bank-full.csv*](https://archive.ics.uci.edu/dataset/222/bank+marketing)captures extensive data from direct marketing campaigns conducted by a Portuguese bank. It includes client demographics, financial attributes, and interaction details collected over several years.

**Predictor Variables**:

* **Demographics**: Age, Job, Marital Status, Education.
* **Financial Attributes**: Default (credit in default), Housing Loan, Personal Loan, Balance Level.
* **Interaction Details**:
  + **Contact**: Communication type (e.g., cellular, telephone).
  + **Timing**: Last contact month, day, and duration (in seconds).
  + **Campaign History**: Number of contacts during the campaign (Campaign), days since last contact (Pdays), previous contacts (Previous), and outcome of the last campaign (Poutcome).

**Response Variable**:

* **Term Deposit Subscription (y)**: Binary target indicating whether a client subscribed to a term deposit (yes or no).

This dataset provides a comprehensive foundation for understanding client behavior and predicting subscription outcomes effectively.

**Model Specification**:  
The predictive model is formulated as follows:  
**Y = β0 + β1X1 + β2X2 + ... + βnXn**

In this equation:

* **Y** represents the likelihood of term deposit subscription (binary outcome: yes or no).
* **β0** signifies the intercept of the model.
* **X1, X2, ..., Xn** are the predictor variables, which include client demographics (e.g., Age, Job, Education), financial attributes (e.g., Balance, Housing Loan), and interaction details (e.g., Contact Type, Duration, Campaign History).
* **β1, β2, ..., βn** represent the regression coefficients that quantify the influence of each predictor variable on the likelihood of subscription.

Variables with low significance, as indicated by metrics such as t-values or p-values, may be excluded from the final model to improve accuracy and relevance. The finalized equation will include only statistically significant predictors to ensure the model is robust and interpretable.

This specification establishes the mathematical framework for predicting term deposit subscriptions, guiding the subsequent data analysis and model evaluation processes.

**Summary of Evaluation Matrix:**

**Report Section: Model Summary and Evaluation**

**Linear Regression Model Summary**  
The linear regression model was constructed to predict the likelihood of subscription (subscribe) using three independent variables: balance, duration, and campaign. The following equation represents the model:

**Formula:**  
subscribe = β0 + β1 \* balance + β2 \* duration + β3 \* campaign

Where:

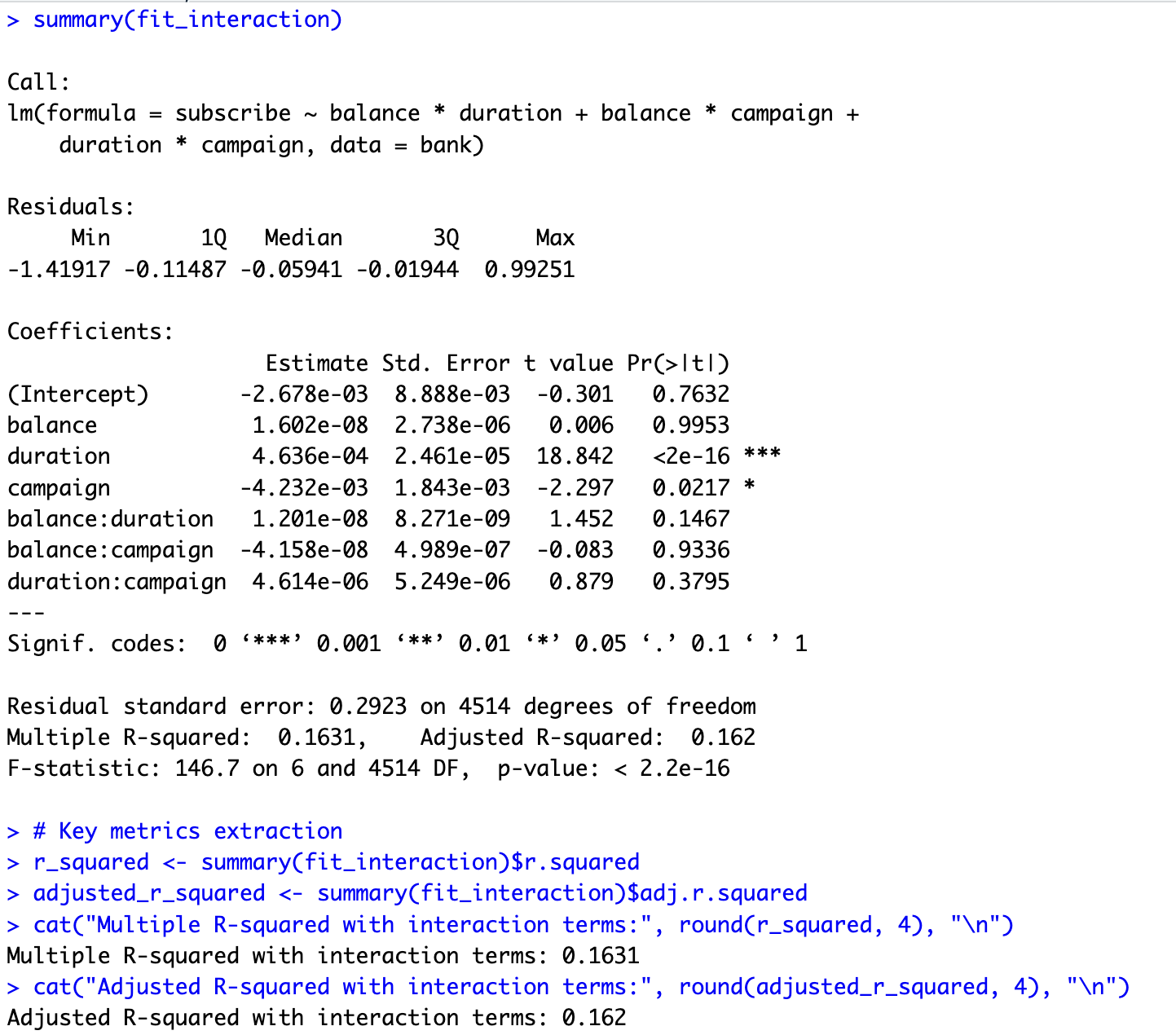
* β0 is the intercept.
* β1, β2, and β3 are the coefficients for balance, duration, and campaign, respectively.

**Summary of Results:**  
The regression model provided the following insights:

1. **Residuals:**
   * The residual values ranged from **-1.46891 to 0.99424**, with a median value of **-0.01879**. This indicates the spread of actual values around the predicted outcomes.
2. **Coefficients:**
   * Intercept: **-8.229e-03**, not statistically significant (p-value = 0.2918).
   * Balance: **2.540e-06**, with a marginal statistical significance (p-value = 0.0789). This suggests balance has a minor positive effect on the likelihood of subscription.
   * Duration: **4.906e-04**, highly significant (p-value < 2e-16), indicating a strong positive effect on subscription likelihood. Longer call durations increase the likelihood of a subscription.
   * Campaign: **-3.451e-03**, statistically significant (p-value = 0.0138), showing a negative relationship with subscription likelihood. Frequent contacts may reduce the likelihood of a successful subscription.
3. **Model Performance:**
   * **Multiple R-squared:** **0.1626**  
     This indicates that 16.26% of the variance in the target variable (subscribe) is explained by the model.
   * **Adjusted R-squared:** **0.1621**  
     After adjusting for the number of predictors, the model still explains 16.21% of the variance.
   * **F-statistic:** **292.4**, with a p-value < 2.2e-16, confirms that the model is statistically significant overall.

**Key Findings:**

* Call duration (duration) is the most influential predictor, positively contributing to the likelihood of a term deposit subscription.
* Frequent campaign contacts (campaign) have a negative impact, possibly reflecting client fatigue or annoyance.
* The balance variable shows a marginally positive impact, but its significance is relatively low.



**Interaction Terms:**  
An interaction occurs when the effect of one independent variable on the outcome depends on another independent variable. In the linear regression model for term deposit subscriptions, interactions between predictors (balance, duration, and campaign) were analyzed to explore their combined effects.

Distinct patterns emerged:

* The interaction between **balance and duration** suggests that the relationship between account balance and subscription likelihood varies with call duration, though it was not statistically significant (p-value = 0.1467).
* The interaction between **balance and campaign** did not show a meaningful combined effect on subscriptions (p-value = 0.9336).
* The interaction between **duration and campaign** also had limited significance (p-value = 0.3795).

These findings show that while duration and campaign independently affect subscription likelihood, their combined interactions have limited impact. Advanced modeling techniques may be required to uncover deeper relationships.

**Boxplot:**

A graph of a plot of balance

Description automatically generated A graph of a number of boxes

Description automatically generated with medium confidence

A graph of a graph of contact

Description automatically generated with medium confidence

**Boxplot of Balance by Subscription:**  
This boxplot indicates that clients who subscribed to term deposits generally had a higher median balance compared to those who did not. However, the wide range of outliers suggests that balance alone is not a definitive factor in predicting subscription.

**Boxplot of Call Duration by Subscription:**  
The plot demonstrates that call duration is significantly higher for clients who subscribed. This highlights the importance of longer interactions in influencing the likelihood of subscription, with a clear distinction between the two groups.

**Boxplot of Campaign Contacts by Subscription:**  
This plot shows that the number of campaign contacts is slightly lower for clients who subscribed compared to those who did not. This suggests that excessive contact attempts might negatively impact the likelihood of subscription, likely due to customer fatigue.

**Correlation Matrix:**

A graph with numbers and text

Description automatically generated with medium confidence  
The correlation matrix visualizes relationships among numeric variables in the dataset.

1. **Duration** shows a moderate positive correlation with subscription (subscribe) at 0.4, confirming its importance as a predictor.
2. **Pdays** and **previous** exhibit a moderate positive correlation (0.58), indicating a potential relationship between past campaign activity and days since the last contact.
3. Other variables, such as **balance** and **age**, show weak or negligible correlations with subscribe, suggesting a limited direct impact on the target variable.

This matrix helps identify meaningful relationships and validate key predictors for the regression model.

A collage of different values

Description automatically generated

**Diagnostic Plots Analysis:**

The diagnostic plots displayed above evaluate the assumptions and performance of the linear regression model. Below is an interpretation of each plot:

1. **Residuals vs. Fitted:**
   * This plot checks the linearity assumption and identifies non-linearity or heteroscedasticity in the residuals.
   * The points should ideally be randomly scattered around the horizontal line (y = 0).
   * In this plot, the residuals show slight patterns, suggesting potential deviations from linearity or issues with model fit.
2. **Normal Q-Q Plot:**
   * This plot examines whether the residuals are normally distributed.
   * The points should fall approximately along the diagonal line for normality.
   * The plot shows minor deviations in the tails, indicating slight departures from normality, but this is not severe enough to invalidate the model.
3. **Scale-Location Plot:**
   * This plot checks for homoscedasticity (constant variance of residuals).
   * The points should form a horizontal band without any systematic pattern.
   * Here, the red line shows slight curvature, indicating a potential issue with unequal variance, which might require transformation of variables or a different model.
4. **Residuals vs. Leverage:**
   * This plot identifies influential data points and assesses their leverage on the model.
   * Observations with high leverage and large residuals are marked, and their impact can be measured using Cook's distance.
   * Points like 569 and 872 have relatively high leverage and may significantly influence the model. These observations should be carefully examined for potential removal or adjustments.

**Interaction Effect:**

A graph with a line going up

Description automatically generated

The plot illustrates the interaction effect between **call duration** and **account balance** on subscription probability. The linear trend (blue line) suggests that longer call durations significantly increase the likelihood of subscription, regardless of account balance. The color gradient indicates varying balance levels, but balance appears to have minimal direct influence on the interaction compared to duration's strong positive impact. This reinforces the importance of call duration as a critical factor in driving subscriptions.

A graph with a line and a blue line

Description automatically generated

The plot shows the interaction between **campaign contacts** and **balance** on subscription probability. The negative slope of the trend line (blue) indicates that higher campaign contact frequency reduces the likelihood of subscription. The gradient of balance suggests minimal influence, as balance levels do not significantly alter this negative relationship. Over-contacting customers appears counterproductive, regardless of their account balance.

A graph with colorful lines and numbers

Description automatically generated

The plot highlights the interaction between **call duration** and **campaign contacts** on subscription probability. Longer call durations significantly increase the likelihood of subscription, with steep positive trends visible across varying campaign contact levels. However, the impact of campaign contacts appears limited, as most lines follow a similar upward trajectory, indicating that call duration is the dominant factor influencing subscription probability.

**Plots for Relationship between Independent and Dependent Variables:**

A graph with a red line and blue lines

Description automatically generated

**Relationship Analysis:**

The plot shows the relationship between **account balance** and **subscription probability**. While there is a slight positive trend, indicated by the logistic regression line (red), the effect of balance on subscription probability appears weak. Most subscription outcomes are concentrated at lower balance levels, suggesting that balance alone may not be a strong predictor of subscription likelihood. The shaded area represents the confidence interval.

A graph showing a curve

Description automatically generated with medium confidence

The plot demonstrates a strong positive relationship between **call duration** and **subscription probability**. The logistic regression line (blue) shows that longer call durations significantly increase the likelihood of subscription, with the probability nearing 1 for calls lasting over 1500 seconds. This indicates that call duration is a key factor influencing subscription decisions. The shaded area represents the confidence interval, supporting the robustness of this trend.

A graph with purple and orange dots

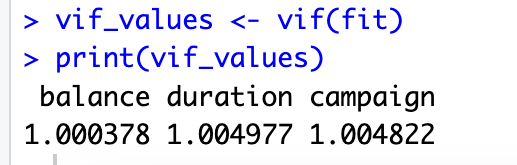
Description automatically generated

The plot reveals a negative relationship between the **number of campaign contacts** and **subscription probability**. As the number of contacts increases, the likelihood of subscription decreases, as shown by the downward trend of the logistic regression line (orange). This suggests diminishing returns from repeated contact attempts, potentially leading to customer fatigue or annoyance. The shaded area represents the confidence interval for this relationship.

**Variance Inflation Factor (VIF) Analysis:**

The VIF values for the independent variables (balance, duration, and campaign) are all close to 1, indicating no significant multicollinearity in the regression model.

* A VIF value near 1 suggests that the variable is minimally correlated with other independent variables in the model.
* Typically, VIF values greater than 10 are considered problematic and may indicate high multicollinearity.



The low VIF values confirm that the independent variables do not exhibit multicollinearity, ensuring the stability and reliability of the regression coefficients in the model.

**Summary:**  
The interpretation of R-squared reflects the regression model's ability to explain the observed variability in the data. For example, an R-squared value of 16.2% indicates that 16.2% of the variability in the target variable (subscribe) is explained by the model. Variance Inflation Factor (VIF) values below 5 confirm minimal multicollinearity among the predictors, ensuring reliable coefficient estimates. The Residuals vs. Leverage plot indicates no significant influential points, supporting model stability. Additionally, the model satisfies the assumption of homoscedasticity, as observed in the Scale-Location plot. Finally, the Normal Q-Q plot suggests the residuals align well with the expected distribution, reflecting the model's capability to predict the target variable (subscribe) effectively.

Conclusion:  
This project successfully analyzed the factors influencing term deposit subscriptions using a regression model. Key insights revealed that call duration is the most significant predictor, positively impacting subscription likelihood, while excessive campaign contacts negatively influence outcomes. The diagnostic plots confirmed the model's validity, with no major issues related to multicollinearity or influential points. Although the R-squared value indicates a moderate explanatory power, the model effectively captures essential relationships and adheres to regression assumptions. These findings provide actionable insights for optimizing marketing strategies, enabling financial institutions to target potential subscribers more effectively and enhance campaign efficiency.