Report on Term Deposit Subscription Prediction

1. Summary of Analysis & Key Findings from the Data

The exploratory data analysis (EDA) phase revealed several compelling insights that provide a deeper understanding of customer behaviour and the factors influencing term deposit subscriptions.

One of the most significant observations was the **high class imbalance** in the target variable. Only about **11% of the clients subscribed** to a term deposit (y = yes), while the remaining 89% did not. This imbalance highlighted the need for techniques such as class weighting or resampling to ensure the model does not become biased toward the majority class.

Another crucial finding was the **strong impact of call duration** on subscription rates. Clients who spent more time on the phone with an agent were far more likely to subscribe. This suggests that **engagement during the call plays a vital role** in influencing a customer's decision potentially indicating that longer conversations either build more trust or are reserved for more qualified leads.

We also discovered that **economic context matters**. Variables such as the **Euribor 3-month rate** and the **employment variation rate** were statistically correlated with subscription behaviour. When interest rates were low or employment conditions improved, clients were more open to investing in a term deposit aligning with real-world economic intuition.

The pdays variable, which indicates the number of days since a client was last contacted, revealed that a large portion of clients had **never been contacted before** (indicated by the placeholder value of 999). This suggests that **first-time contacts** still form a large portion of outreach, and re-engagement strategies may not be fully leveraged.

Demographic analysis showed that **retired individuals, students, and the unemployed** were **more likely to subscribe** than working professionals. This could reflect either more available time for consideration, greater financial conservatism, or responsiveness to long-term financial planning.

Seasonality also played a role. Clients contacted in the **months of March and December** had **noticeably higher subscription rates**. These months may coincide with seasonal financial planning periods for instance, end-of-year savings or new-year investment goals and suggest optimal timing for future campaigns.

Outlier Detection and Treatment

During the exploratory data analysis phase, we examined key numerical features for the presence of outliers, particularly in columns such as age, duration, campaign, and pdays. Outliers can have a significant impact on both statistical summaries and model performance, especially for algorithms sensitive to feature scale, like logistic regression.

Visualizations such as boxplots were used to identify unusually high or low values. For example, the duration variable (call duration in seconds) showed a long right tail, with some calls lasting much longer than the typical range. Similarly, the pdays variable, which captures the number of days since the last contact, exhibited a special value of 999 indicating clients who had not been previously contacted, which was handled as a separate category during feature engineering.

Based on this analysis, outliers were not arbitrarily removed. Instead, we ensured that the model was robust to these values through appropriate scaling and by engineering features that captured important outlier-driven signals (e.g., distinguishing never-contacted clients). This approach preserves genuine rare events that may correspond to successful or failed marketing strategies, while minimizing the risk of bias from anomalous data points

2. Important Features Driving Subscription

Feature importance analysis, conducted using both Random Forest (feature importances) and Logistic Regression (model coefficients), highlighted several variables that significantly influence the likelihood of a client subscribing to a term deposit:

- **Duration of the current call**: This emerged as the single most predictive feature. Longer calls typically indicate more engaged clients or that agents are investing more time in persuading or informing the client. A longer conversation offers greater opportunity to build rapport, explain the benefits, and address objections, leading to a higher probability of subscription.
- **Outcome of the previous campaign (poutcome)**: Clients who had a successful outcome from a previous campaign were far more likely to subscribe again. This confirms that previous behaviour is a strong predictor of future actions. It also underscores the importance of maintaining historical campaign records and re-engaging clients who responded positively in the past.
- **Contact method (contact)**: The data showed that clients contacted via **cellular phones** had a significantly higher conversion rate compared to

those contacted via **landlines**. This may reflect increased availability, a greater willingness to engage via mobile, or simply a generational shift in preferred communication channels.

- **Month of contact (month)**: Timing turned out to be a crucial factor. The months of **March and December** showed peak subscription activity. This could correspond with salary bonuses, end-of-year savings plans, or new-year financial resolutions making these months strategically important for scheduling campaigns.
- **Engineered feature (call_efficiency)**: Created by dividing call duration by the number of campaign contacts, this feature revealed how effectively each interaction was being used. Higher call efficiency meaning longer, quality conversations with fewer overall calls was linked to higher subscription rates.
- **Engineered feature (previous_contacted)**: Derived from the pdays field, this binary indicator identified whether the client had been previously contacted. Clients who had been contacted before, especially within a reasonable time window, were more likely to convert. This demonstrates the value of **follow-up campaigns** and **lead nurturing**.
- **Financial burden indicator (debt_load)**: This feature was constructed by checking whether the client had either a housing or personal loan. Clients with **fewer or no loans** were found to be more likely to subscribe to term deposits, which makes sense from a financial planning standpoint. Lower existing liabilities may indicate higher disposable income or a greater ability to invest.

3. Marketing Recommendations Based on Observed Patterns

Based on the insights drawn from the data analysis and modelling, the following marketing strategies are recommended to enhance the effectiveness and efficiency of future campaigns:

- **Prioritize calling clients with previously successful engagements**: The outcome of past marketing efforts (captured in the poutcome feature) is a powerful indicator of future behaviour. Clients who responded positively in earlier campaigns are **far more likely to subscribe again**. Marketing teams should maintain and leverage a list of clients with successful histories and move them to the top of the call queue.
- **Schedule campaigns during high-performing months (March and December)**: The analysis revealed clear **seasonal trends** in client behaviour. March and December saw significantly higher subscription rates, likely due to financial planning around year-end bonuses, tax filing, or new year savings goals. Focusing outreach efforts in these months can **maximize return on marketing investment**.

- **Favour cellular contact over landlines for higher engagement**: Clients contacted via **cellular phones were more responsive** than those reached through landlines. This may be due to the increasing dominance of mobile communication, better accessibility, or convenience. Campaign strategies should shift more heavily toward **mobile-first outreach**.
- **Target less-indebted clients to improve conversion efficiency**: Clients without outstanding housing or personal loans (as represented by the engineered debt_load feature) were **more inclined to invest in a term deposit**. By identifying and prioritizing financially stable clients, marketers can improve **conversion rates** while reducing effort spent on low-probability leads.
- **Allocate resources to longer, quality conversations instead of high-frequency short calls**: The duration of the last call was the **most predictive feature** of subscription. This suggests that successful subscriptions are typically preceded by **longer and more meaningful conversations**. Rather than increasing call volume, teams should invest in **training agents to hold productive discussions** and give them the time to do so.
- **Focus campaigns on retirees, students, and those with prior positive outcomes**: Certain demographic groups notably **retirees, students, and the unemployed** showed **higher subscription tendencies**. These segments may be more receptive to financial products like term deposits. Combined with past campaign outcomes, this demographic focus can significantly improve the precision of targeting.

4. Business-Ready Conclusions

The insights gained from this project and the performance of the developed models provide strong foundations for actionable business strategies. These conclusions can be immediately applied by the marketing and operations teams to enhance campaign effectiveness and improve return on investment:

- **Enhancing client targeting to improve subscription rates**: By using predictive modelling based on historical data, the business can **identify clients with a high likelihood of subscribing**. This allows marketers to focus their energy and resources on the most promising leads rather than casting a wide and inefficient net.
- **Reducing marketing costs by minimizing unsuccessful outreach**: With insights into which features most strongly influence client behaviour such as call duration, prior campaign success, and demographic indicators the team can **avoid unnecessary or low-value interactions**. This translates to **fewer wasted calls, reduced manpower costs, and better campaign ROI**.
- **Providing real-time lead scoring using the deployed model**: Once integrated into the business's customer relationship management (CRM) or campaign systems, the model can be used to **score new leads in real**

time. This scoring allows marketing agents to know ahead of time which clients are more likely to subscribe, leading to smarter decision-making during active campaigns.

- **Prioritizing high-value leads based on engineered features and call metadata**: The inclusion of engineered features like call_efficiency, previous_contacted, and debt_load enables the model to go beyond surface-level attributes. These indicators help the business **recognize behavioural patterns and financial readiness** ultimately leading to **better prioritization and higher success rates**.

5. Final Model Performance

The Logistic Regression model, after feature scaling and class balancing, demonstrated consistent and stable results across both training and test datasets:

- **Train Accuracy**: 86%
- **Test Accuracy**: 86%
- **F1 Score (Class 'yes')**: 0.59 (train), 0.60 (test)
- **Recall (Class 'yes')**: 0.89 (train), 0.91 (test)

The **high recall** values for the positive class ('yes') are particularly important for business objectives. This means the model is able to correctly identify **most of the clients who are likely to subscribe**, making it highly valuable for lead targeting and maximizing campaign success rates. The **minimal drop in F1 score between training and test** indicates that the model is not overfitting and maintains its performance on unseen data.

- **Random Forest Performance and Overfitting**

On the other hand, the Random Forest model, despite its robustness and ability to model complex interactions, showed clear signs of **overfitting**:

- **Train F1 Score (Class 'yes')**: 1.00
- **Test F1 Score (Class 'yes')**: 0.53
- **Test Recall (Class 'yes')**: 0.43

While it performed perfectly on the training set, its performance dropped significantly on the test data, especially in its ability to recall actual subscribers. This gap suggests that the model had **memorized patterns** in the training set but failed to generalize those patterns to new, unseen data a classic symptom of overfitting.

The **Logistic Regression model strikes the right balance** between predictive performance and reliability, especially when combined with the engineered features. It also offers **interpretability**, allowing the business to understand

which factors contribute to client behaviour key advantage in regulated environments like finance.

Conclusion

In conclusion, we were able to use the power of data-driven decision-making in improving marketing outcomes for financial products. By combining structured data analysis, feature engineering, and predictive modelling, we were able to uncover critical insights about client behaviour, economic influence, and the effectiveness of previous marketing strategies.

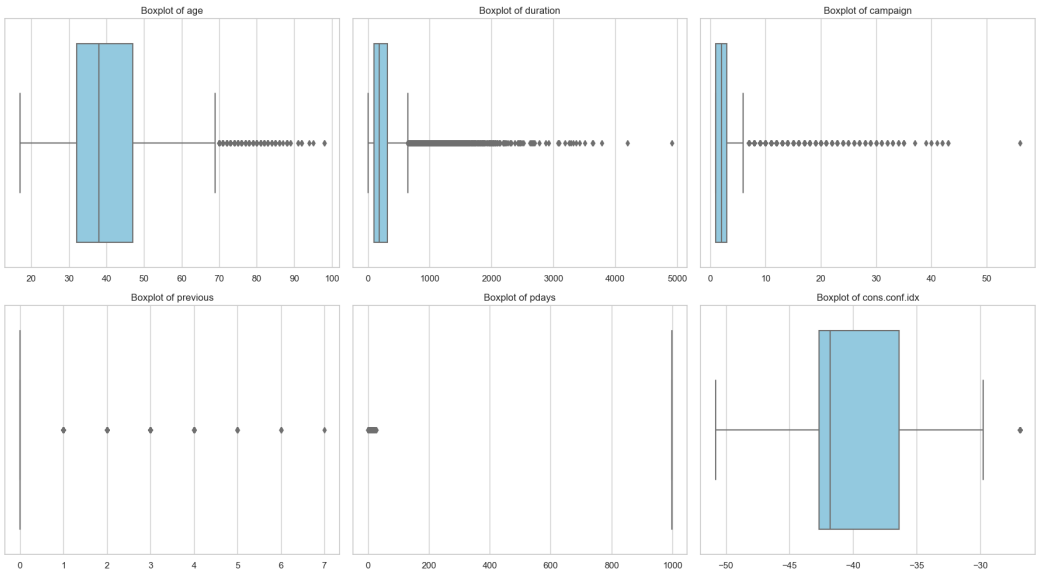
The Logistic Regression model emerged as the most suitable predictive tool for deployment, offering both consistent performance and interpretability two factors that are essential in real-world business applications, particularly in regulated industries like banking. It achieved a strong balance between recall and precision, especially in identifying clients most likely to subscribe, thereby maximizing campaign effectiveness.

Beyond modelling, the insights extracted from the data such as the influence of call duration, timing of campaigns, and client financial status equip the marketing and operations teams with actionable strategies. These include optimizing campaign timing, refining target audiences, and focusing on lead nurturing tactics that improve engagement and conversion.

Overall, this project not only builds a robust foundation for targeted marketing using machine learning but also creates a pathway for continuous improvement through ongoing data collection, model updates, and strategic refinement. The implementation of these insights and tools has the potential to significantly enhance client acquisition, reduce costs, and support data-informed growth.

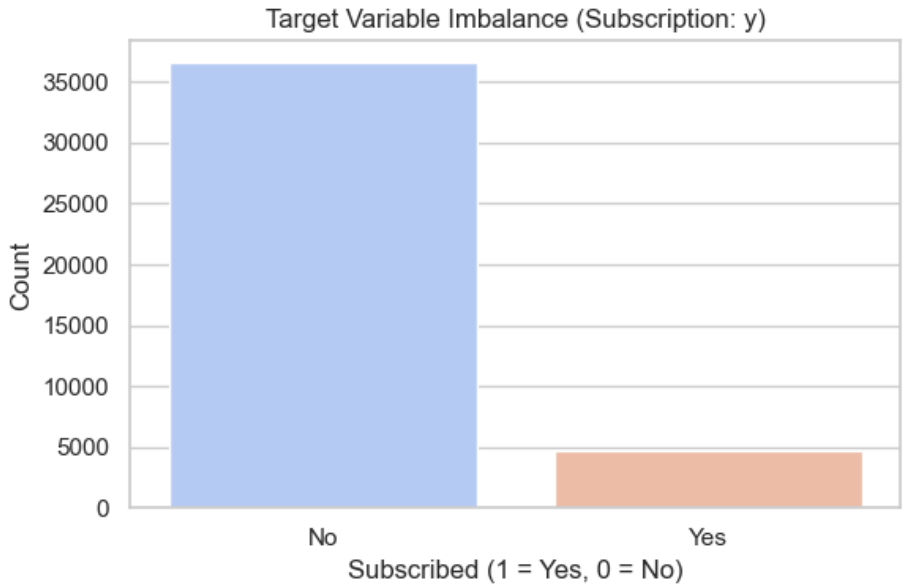
Figure Appendix

Figure 1: Boxplots showing the distribution and outliers for selected numeric features (age, duration, campaign, previous, pdays, and cons.conf.idx). Outliers are indicated by points beyond the whiskers.



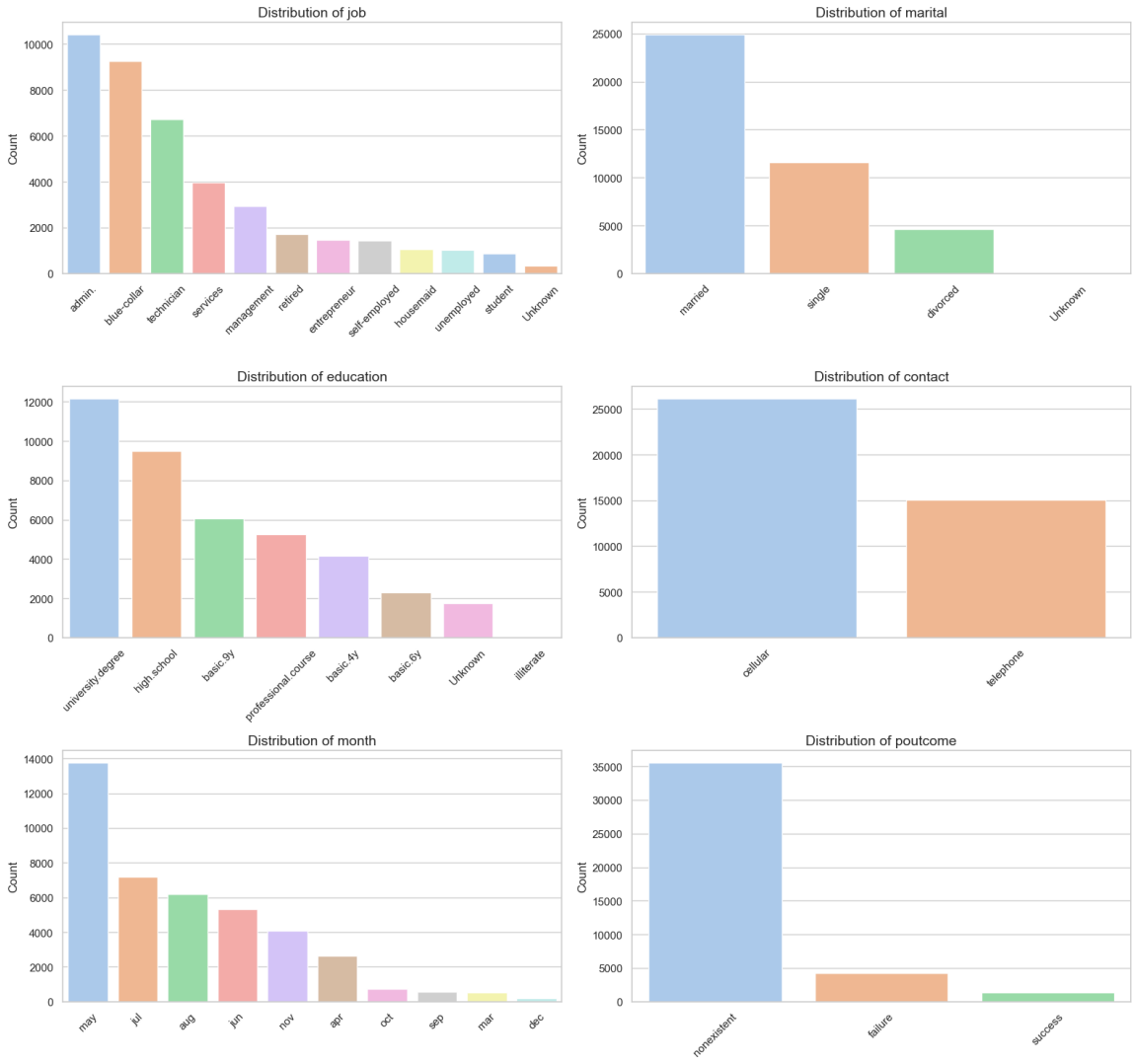
Outlier analysis was conducted for all numerical variables

Figure 2: Distribution of the target variable (y). This plot visualizes the imbalance between clients who subscribed to a term deposit ("yes") versus those who did not ("no").



The severe class imbalance in the data is visually confirmed in Figure 2

Figure 3: Subscription rate by categorical variables (job, marital, education, contact, month, poutcome). Each subplot shows how a different categorical feature influencing the likelihood of subscription.



As shown in Figure 3, certain job roles, months, and contact methods have a strong influence on subscription rates.

Figure 4: Confusion matrix for the Logistic Regression model (with scaled data), showing performance on the test set.

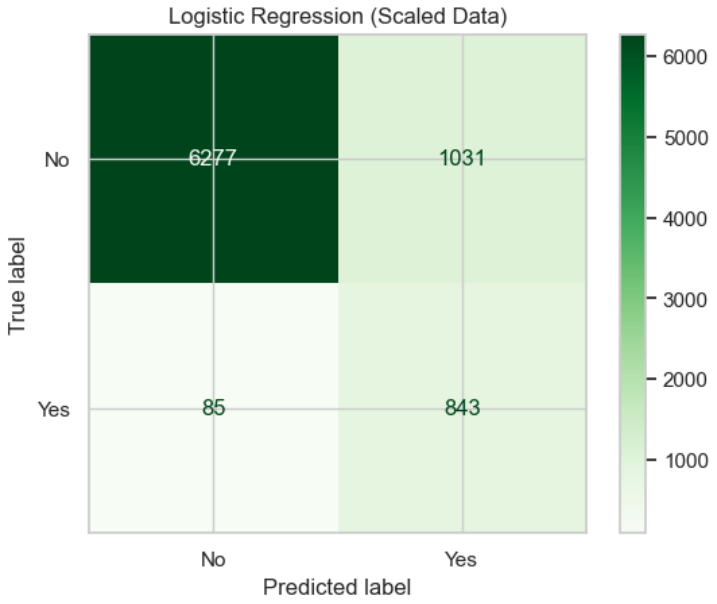


Figure 5: Confusion matrix for the Random Forest model, showing correct and incorrect predictions on the test data.

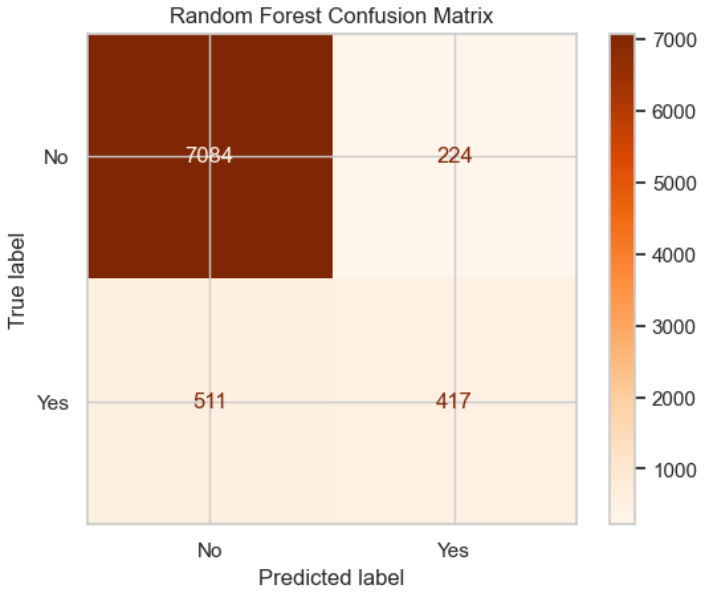


Figure 4 and Figure 5 show the confusion matrices for the Random Forest and Logistic Regression models, respectively.