**Progress Report: Exploratory Data Analysis and Feature Engineering**

**Objective**

The aim of this progress report is to provide insights from the exploratory data analysis (EDA) and feature engineering phases of the project. These findings will guide subsequent predictive modeling and inform marketing strategies to increase term deposit subscriptions.

1. **Exploratory Data Analysis (EDA)**

**Key Findings**

* **Analyzing Patterns:**

**Age Group:** Most campaign participants were between 27–50 years, with peak engagement at age 31. Older individuals (60+ years) had the lowest contact rates.

**Job Titles:** Students showed the longest call durations, suggesting higher engagement levels. Unemployed and unknown categories had the shortest durations.

**Contact Method:** Cellular communication had a higher success rate (14.74%) compared to telephone (11.27%).

**Campaign Days:** Thursdays had the most campaign activity, while Fridays saw the least.

**Campaign Months:** March had the highest subscription success rate (50.55%), and May had the lowest (6.43%).

**Education:** University graduates had the highest subscription rates, while illiterate participants had the lowest.

* **Correlation Analysis:**

**Duration:** The strongest positive correlation (0.405), indicating longer calls are strongly associated with successful subscriptions.

**Previous Contacts:** Moderate positive correlation (0.230), suggesting prior contact increases subscription likelihood.

**Economic Indicators:** Features like emp.var.rate (-0.298), euribor3m (-0.308), and nr.employed (-0.355) negatively correlate with subscriptions, reflecting economic influences.

**Pdays:** Negative correlation (-0.325), indicating shorter intervals between contacts lead to better outcomes.

**Age and Campaign Frequency:** Weak correlations (0.030 and -0.066, respectively), suggesting minimal direct impact.

* **Descriptive Statistics Insights:**

**Age:** Average of 40 years, with most participants between 17 and 98 years.

**Call Duration:** Average of 258 seconds, with highly skewed data indicating most calls were short.

**Campaign Frequency:** Average of 2.57 contacts, with extreme outliers (maximum of 56).

**Pdays:** Majority had no recent contact, reflecting a median value of 999.

**Recommendations for the Marketing Team**

* Focus on individuals aged 27–50, the most engaged age group.
* Prioritize cellular communication for higher success rates.
* Schedule campaigns on Thursdays, the most active day.
* Target campaigns in March, leveraging its high success rate.
* Create tailored messaging for university-educated individuals.
* Encourage longer calls for greater engagement, as duration correlates positively with success.
* Adapt campaigns based on economic conditions, considering their influence on subscription likelihood.

1. **Feature Engineering**

**Correlation Coefficients for Encoded Categorical Features**

|  |  |
| --- | --- |
| Feature | Correlation Coefficient |
| Job Encoded | 0.0251 |
| Marital Encoded | 0.0263 |
| Education Encoded | -0.0358 |
| Housing Encoded | -0.0114 |
| Loan Encoded | 0.0031 |
| Contact Encoded | -0.1448 |
| Month Encoded | 0.0372 |

**Key Observations**

* Month Encoded has the highest positive correlation (0.0372), potentially capturing seasonality effects in client subscription behavior.
* Contact Encoded shows a notable negative correlation (-0.1448), suggesting an inverse relationship with the target variable.
* Correlations for all features are relatively weak, indicating limited linear relationships, but they may hold predictive value when combined or processed using advanced modeling techniques.

This progress report ensures the marketing team are updated on the project’s current status, findings, and actionable recommendations. The next stage will involve predictive modeling, where these insights will be further tested and refined.