KEZIAH PRABA BONZO

Kpbonzo88@gmail.com

This project is a machine learning model to predict if a client will subscribe to the product, given their demographic and marketing campaign-related information.

1. Initial Findings about Data

1. The original dataset contains **45,211 observations** and **17 variables** (7 numerical and 10 categorical).
2. There are no explicit missing values, but some categorical variables have 'unknown' values which were treated as missing.
3. The target variable ('y') is imbalanced: approximately **88% 'no'** and **12% 'yes'**, indicating a class imbalance problem.

2. Exploratory Data Analysis (EDA)

A. For Numerical Variables

1. Boxplots were used to analyze each numerical variable with respect to the target variable.
2. Variables such as **balance, duration, campaign, pdays,** and **previous** contain significant outliers.
3. Reasonable ranges were selected for each numerical variable using boxplot insights. Outliers were imputed using the mean value.

B. For Categorical Variables

1. Cross-tabulations were created to understand the relationship between each categorical variable and the target variable.
2. Variables with more than 50% 'unknown' values (**poutcome**) or high imbalance (**default**) or negligible impact (**contact**) were dropped.
3. Categorical variables with fewer 'unknown' values were imputed using the mode.

3. Model Training

1. Multiple classification algorithms were tested with cross-validation:
    - Logistic Regression
    - Linear Discriminant Analysis

- K-Nearest Neighbor
- Decision Tree
- Naive Bayes
- Support Vector Machine
- Random Forest

4. Model Selection & Hyperparameter Tuning

1. **Random Forest Classifier** was the last model selected for its performance and interpretability.
2. **GridSearchCV** was used for hyperparameter tuning. Best parameters:
    - N estimators: 200
    - Max depth: None
    - Min samples split: 2
    - Min_samples_leaf: 1
    - Class_weight: 'balanced'

5. Prediction & Evaluation

1. **Confusion Matrix:**
    - True Negatives: 7520
    - False Positives: 465
    - False Negatives: 456
    - True Positives: 602
2. **Evaluation Metrics:**
    - Accuracy – 0.90
    - Precision – 0.56
    - Recall – 0.57
    - F1 Score – 0.57
3. **Interpretation:**
   The model achieves strong overall accuracy. While the precision and recall for the minority class are moderate, performance improved significantly after applying

SMOTE and tuning. This model is capable of identifying likely subscribers without excessive false alarms.

6. Key Insights & Business Recommendations

- **Call duration** is a strong indicator of subscription interest.
- **Previous outcome of a campaign** positively influences current subscriptions.
- **Job type** and **contact method** correlate with likelihood to subscribe.

Recommendations:

1. Prioritize clients with long previous calls and successful past outcomes.
2. Customize marketing strategies based on job categories.
3. Use the model to score leads and focus campaign efforts on high-probability clients.
4. Consider seasonal or monthly trends when timing outreach.

7. Conclusion

Throughout this project, multiple iterations of model training and tuning were carried out with the specific goal of improving the model's ability to correctly identify clients who would **subscribe to a term deposit** — the **minority class ("yes")**. This was measured using the **F1-score**, which balances both precision and recall.

Despite trying several advanced techniques, including:

- Resampling using SMOTE to address the class imbalance,
- Testing different algorithms (e.g., Logistic Regression, Decision Tree, Random Forest, etc.),
- Hyperparameter tuning using GridSearchCV to find the optimal model settings,
- Adjusting class weights to make the model more sensitive to the minority class,

…the F1-score for the 'yes' class consistently remained in the range of 0.54–0.60.

This happened because:

- The dataset is **heavily imbalanced** (~88% 'no', ~12% 'yes'), making it hard for most models to learn the minority pattern well.
- Some of the most predictive features (like **'duration'**) may not be available before making a marketing decision, which limits real-time usefulness.
- The **'yes' class might be inherently harder to predict**, due to overlapping feature patterns between subscribers and non-subscribers.

Final Takeaway:

While the model achieved high overall accuracy (90%), it still struggled to improve F1 performance for the 'yes' class beyond a certain point. This underscores a critical insight: data quality, feature richness, and business context are just as important as the modelling technique.