PREDICTIVE MODELING: WALLACE COMMUNICATIONS CAMPAIGN

Tree-Based and Neural Network Classifiers for Contract Prediction

STUDENT NUMBER: 3533826

1. PROJECT METHODOLOGY

Steps and Data Split Strategy (Req. 2)

The project followed a supervised learning workflow.

1. **Data Preprocessing:** Cleansing (string normalization, ID removal), Imputation (Numeric: Median, Categorical: 'missing').
2. **Feature Encoding:** Low-Card (≤ 20 unique values) features used OHE. High-Card features used **Frequency Encoding**.
3. **Data Split: Stratified** 80% Train / 20% Test split to preserve class imbalance ratios.
4. **Tuning/Validation:** Hyperparameter search utilized **3-fold Stratified Cross-Validation** on the Training set.
5. **Evaluation:** Final models were assessed on the independent Test set.

2. VARIABLES & HYPERPARAMETER APPROACH

Feature Treatment (Req. 3)

Variable	Type	Treatment
`new_contract_this_campaign`	Binary Target (1/0)	N/A
`age, duration`	Numeric (Cont.)	Scaled, Median Imp.
`job, marital, contact` `poutcome`	Cat. (≤ 20 unique)	OHE, Constant Imp.
`month, day` (Example)	Cat. (> 20 unique)	Frequency Encoded

Consequence of Choice: Using Frequency Encoding for high-cardinality features greatly **reduces dimensionality** (benefiting MLP) but risks **feature collision** where unrelated categories share the same frequency value.

Tuning Summary (Req. 4)

Method: Randomized Search CV (n_iter $= 2$, $K = 3$, Metric: ROC AUC) Comment: Computationally efficient but **sub-optimal**. The low iteration count sacrifices best possible performance for speed (Elapsed time: 785s).

3. MODEL INSIGHTS & REFERENCES

Insights Gained (Req. 6)

- **Imbalance Impact:** High Accuracy ($\sim 91\%$) is misleading; low Recall (0.6566) on the minority class confirms the need for ROC AUC as the primary metric.
- **Tree Superiority:** Ensemble tree models (RF, XGB, LGBM) drastically outperformed MLP, suggesting the predictive relationship relies on non-linear feature interactions and splits.

References (Req. 7)

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Statistical Learning*. (Ensemble Justification).
- Brownlee, J. (2020). *Frequency Encoding*. (Feature Encoding Justification).
- Jeni, L. A., & Cohn, J. F. (2016). *Class Imbalance*. (Metric Choice Justification).

4. FINAL MODEL AND RESULTS

Model Justification (Req. 5)

Selected Model: Tuned Random Forest

Chosen due to the highest discriminative power (ROC AUC of **0.9365**), demonstrating superior class separation across all tested models.

Key Performance Summary (Test Set)

Metric	Score
ROC AUC	**0.9365**
Accuracy	0.9110
Precision (Class 1)	0.8541
Recall (Class 1)	0.6566

Confusion Matrix

Actual \Predicted	0 (No Contract)	1 (Contract)
0 (No Contract)	**7929** (TN)	224 (FP)
1 (Contract)	680 (FN)	**1300** (TP)

Commentary on Usefulness: Highly effective at identifying true non-converters (High TN rate), resulting in **efficient sales targeting** (avoiding wasted calls). However, the **680** False Negatives represent 34% of valuable positive leads missed.

5. DATA SUMMARY & IMBALANCE

Dataset Profile

- **Total Customers:** 50,662
- **Test Set Size:** 10,133 (20%)
- **Primary Predictive Feature:** `duration` (Highly correlated with success).

Class Imbalance Ratio

Class	Count (Test)	Ratio
0 (No Contract)	8,153	80.46%
1 (Contract)	1,980	19.54%

Conclusion: Significant $\approx$ **4:1 Imbalance**, justifying the use of stratified sampling and AUC-based evaluation metrics.

6. BEST MODEL PARAMETERS

Tuned Random Forest Configuration

The optimal hyperparameters found via Randomized Search (n_iter=2) were:

- **Estimators (`n_estimators`):** 200
- **Max Depth (`max_depth`):** None (Allows full tree growth)
- **Min Samples Split (`min_samples_split`):** 2
- **Oversampling (SMOTE):** Disabled (`use_smote = False`)
- **Criterion:** Gini (Default)

Note: The preference for deeper trees (`max_depth=None`) is typical for Random Forest when aiming for low bias, potentially leading to overfitting if not checked by cross-validation.

7. BUSINESS RECOMMENDATIONS & FUTURE WORK

Recommendations

- **Threshold Adjustment:** Lower the classification threshold for the Random Forest model to increase Recall (TP rate) on the minority class, reducing missed contract opportunities (FN).
- **Focus on Precision:** Utilize the current model's high precision (0.8541) to create a highly curated "High Confidence" list for expensive, high-touch marketing channels, maximizing ROI.

Future Work

- **Leakage Mitigation:** Investigate and potentially remove the `duration` feature, which often leaks target information, and re-train the model to assess true feature importance.
- **Data Balancing:** Re-run the entire pipeline with SMOTE (`use_smote = True`) to directly address the 4:1 class imbalance and attempt to improve the poor Recall for the positive class.