Predicting Customer Contract Propensity

Student ID: 3539054

Methodology & Workflow

Pipeline Strategy

- **Cleaning:**
  - Corrected categorical typos (e.g., `n→no`).
  - Added a new column `never_contacted`.
  - Replaced `-1` days with a large number.
- **Preprocessing:**
  - Applied `OneHotEncoder` for categoricals.
  - Applied `RobustScaler` for numericals to handle extreme outliers.
- **Modelling:** Trained 3 distinct models. Linear (Logistic Regression), Ensemble (Random Forest), and Deep Learning (Multilayer Perception).
- **Tuning:** Used `GridSearchCV` to optimize hyperparameters.

Validation Approach

- **Split:** Used 80/20 Train/Test split.
- **Stratification:** Applied `stratify=y` to lock the target imbalance in both sets.
- **Validation:** Performed k-Fold Cross-Validation on training data to tune models without touching the Test set.

Variable Types & Processing

Variables	Model Treatment	Processing Method
`town, country, job, married, education, arrears, housing, has_tv_package, last_contact, conn_tr, last_contact_this_campaign_month, outcome_previous_campaign`	Categorical (Nominal)	`OneHotEncoder`: Converts categorical data into a numerical format for machine learning, preventing incorrect assumptions about relationships between categories.
`age, current_balance, this_campaign, contacted_during_previous_campaign, days_since_last_contact_previous_campaign, last_contact_this_campaign_day, never_contacted`	Numeric (Continuous)	`RobustScaler`: Centers data and scales based on percentiles to handle extreme outliers (e.g., `current_balance`). Replaced `-1` with (`2×Max`) to preserve the "Recency" order (Recent < Old < Never) in `days_since_last_contact_previous_campaign`

We treated connection type (`conn_tr`) as categorical rather than numeric.

- **Reasoning:** Although the data uses integers (1, 2, 3), these are IDs, not quantities.
- **Impact:** Treating them as numeric would force the model to assume `Type 5` is greater than `Type 1`. By One-Hot Encoding them, the model learns unique patterns for each connection type without assuming a false mathematical relationship.

Hyperparameter Tuning Strategy

Component	Strategy / Execution
Methodology	`GridSearchCV` with **k-Fold Cross-Validation**, optimizing for **F1-Score** to specifically address the target class imbalance (majority ‘No’, few ‘Yes’).
Logistic Regression	Tuned `C` and `penalty` (L1 vs. L2) to find the optimal balance of bias and variance.
Random Forest	Tuned **Forest Size** (`n_estimators: 100, 200`), **Tree Complexity** (`max_depth: 10, 20, None`) and **Leaves** (`min_samples_leaf: 1, 2, 4`) to control overfitting.
Neural Network	Tuned **Architecture** (Wide vs. Deep layers) and **Activation** (relu vs. tanh) to test different feature learning capabilities.
Validation Metric	**F1-Score** is chosen to penalize models that ignore the minority class (`target=‘Yes’`), ensuring the selected hyperparameters prioritize finding actual buyers over simple accuracy.

Parameter grids were chosen to span the **Bias-Variance spectrum**, testing **constrained vs. flexible** architectures (e.g., Shallow vs. Deep Trees) to minimize overfitting.

Final Model Selection

Justification: The **Random Forest** was selected as the production model having achieved the highest **AUC Score (0.8669)** and best balance of Precision/Recall.

- **Vs. Linear:** It outperformed Logistic Regression (`AUC 0.76`) by successfully capturing non-linear customer segments that the linear model missed.
- **Vs. Deep Learning:** It outperformed the Neural Network (`AUC 0.78`), which was too sensitive to the `‘days_since_last_contact_previous_campaign’` attribute. This caused the network to be overly cautious, resulting in too many false negatives (missed sales opportunities).

	Predicted ‘No’ (Do Not Call)	Predicted ‘Yes’ (Call Target)
Actual ‘No’	7,501 *(Correct Rejection)*	652 *(Wasted Calls)*
Actual ‘Yes’	812 *(Missed Sales)*	**1,168** *(Successful Sales)*

The matrix demonstrates high operational efficiency:

- **Success Rate (Precision = 64%) :** Out of 1,820 recommended calls, 1,168 result in a sale.
- **Buyers Found (Recall = 59%) :** The model successfully identifies the majority of buyers (1,168 out of 1,980).

Additional Insights

- **Financials & Age Drive Decisions:** Contrary to the expectation that campaign history is paramount, the Random Forest feature importance analysis ranked `‘current_balance’` and `‘age’` as the top two predictors. This indicates that a customer's financial health and life stage are stronger indicators of their propensity to buy a mobile contract than their previous interactions with the marketing team.
- **Deep Learning Brittleness:** The Neural Network demonstrated extreme sensitivity to our replacement strategy for `‘-1’` while the Random Forest handled the placeholder value (`2×Max`) gracefully. This highlights that complex Deep Learning architectures can be less robust than Ensembles when dealing with engineered outliers in tabular data.

Reference: Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. O'Reilly Media.