Case Study: SmartML-Opt – A Self-Balancing AutoML Framework for Imbalanced Classification

Contents

1. Problem Statement and Objectives

In many real-world scenarios such as marketing, fraud detection, and healthcare, datasets are often imbalanced — meaning one class heavily outweighs the other. This can lead to machine learning models that perform well on the majority class but poorly on the minority class, which is often the class of interest.

Objective:

To develop a reliable AutoML pipeline that automatically handles class imbalance using oversampling, selects important features, and tunes models to improve performance — especially recall and F1-score — for the minority class.

---

2. Dataset Used

- **Name:** UCI Bank Marketing Dataset
- **Type:** Tabular data
- **Target Variable:** y (client subscribed to term deposit: yes/no)
- **Class Distribution:**
    - Yes: ~11.7%
    - No: ~88.3%

---

3. Data Preprocessing Steps

1. Removed missing values and entries with "unknown" labels.
2. Encoded categorical variables using LabelEncoder.
3. Scaled numerical features using StandardScaler.
4. Balanced the dataset using **SMOTE** (Synthetic Minority Over-sampling Technique).

---

4. Model Development Process

- ## Feature Selection:

  Applied **Recursive Feature Elimination (RFE)** to select top relevant features.

- ### Model Training & Tuning:

  Used **GridSearchCV** to train and tune the following models:

  - Random Forest
  - Gradient Boosting
  - Support Vector Machine (SVM)

- ### Evaluation Metrics:

  Focused on **Recall**, **F1-score**, and **AUC** to assess minority class performance.

- ### Final Selected Model:

  - **Gradient Boosting** with optimized parameters
  - Best performance on validation set in terms of recall and F1

---

5. Visualizations and Insights

- Duration: longer calls correlate with "Yes"
- Age: Peaks around 35-45 and 60+
- Balance: slight trend toward higher balances among "Yes"
- Students & retirees show higher "Yes" rates
- Singles more likely to subscribe
- Tertiary education increases subscription likelihood
- Cellular contact = higher success
- Successful past outcomes => better conversion

Metric	Value
F1 Score	0.79
Recall	0.85
Precision	0.75
AUC Score	0.94

Key Visual Outputs:

- Categorical Features vs Target

- Distribution of Continuous Features by Target
- Confusion Matrix
- ROC Curve
- Architecture Flow

---

6. Recommendations

- Use this framework in any imbalanced classification task across industries.
- It can be integrated into current pipelines for automatic preprocessing, balancing, and model selection.
- Future improvements can include LightAutoML, Optuna, or other AutoML libraries for faster optimization.

---

7. Conclusion

SmartML-Opt offers a complete AutoML solution for imbalanced classification. It combines essential steps — from data balancing (SMOTE) to feature selection (RFE) and model tuning — in an automated and trackable way. This improves model fairness, performance, and usability in real-world applications.