CampaignInsight: AUC-Optimized Mail Response Prediction Pipeline

Executive Summary

This project presents a robust and scalable machine learning solution for predicting customer responses to large-scale direct mail marketing campaigns. The pipeline leverages advanced modeling techniques to address the challenges of severe class imbalance, improve model interpretability, and optimize predictive performance for real-world business impact.

---

1. Objective

The primary objective is to develop a high-performance classification model that can identify potential responders to marketing campaigns with a high degree of precision and recall particularly in the context of highly imbalanced data. The goal is to support smarter targeting strategies that reduce cost and increase campaign ROI.

---

2. Methodology

2.1 Data Preprocessing

- Removed unnecessary identifiers and cleaned missing or anomalous values.
- Performed one-hot encoding and scaling for compatibility with tree-based classifiers.
- Split the dataset into training, validation, and hold-out test sets to ensure reliable generalization.

2.2 Handling Class Imbalance

- Applied **SMOTE (Synthetic Minority Over-Sampling Technique)** to augment minority class samples.
- Used **XGBoost's scale_pos_weight** parameter and **custom class weighting** to improve minority class sensitivity.
- Ensured balance between model performance and overfitting control through repeated cross-validation.

2.3 Model Training

- Used **XGBoost**, a gradient-boosted tree ensemble, due to its performance with structured data.

- Hyperparameters were tuned using grid search and early stopping based on validation AUC.
- Saved models and pipelines modularly to ensure reproducibility and reusability.

2.4 Threshold Optimization

- Evaluated model probabilities on validation sets using ROC and Precision-Recall curves.
- Dynamically optimized decision thresholds to maximize **F1-score**, **Precision**, and **Recall at fixed False Positive Rate**.
- Achieved ROC AUC above **0.84** and PR AUC around **0.36** on unseen test data.

---

3. Evaluation

3.1 Performance on Held-Out Test Set

- **ROC AUC**: 0.8446
- **PR AUC**: 0.3632
- **F1-score (Positive class)**: 0.42
- **Precision (Positive class)**: 0.55
- **Recall (Positive class)**: 0.34

3.2 Confusion Matrix

	Predicted No	**Predicted Yes**
Actual No	8456	30
Actual Yes	71	36

These results demonstrate that the model captures meaningful patterns in customer behavior, while balancing the risks associated with false positives and negatives.

---

4. Key Outcomes

- Successfully implemented a machine learning pipeline for campaign response prediction.
- Addressed class imbalance challenges using hybrid resampling and cost-sensitive learning.
- Optimized decision thresholds to meet marketing-specific business objectives.
- Delivered a modular and production-ready framework for further experimentation or deployment.

---

5. Future Work

- Integrate time-based features and campaign metadata for enhanced context.
- Test alternate models such as CatBoost or LightGBM for performance benchmarking.
- Deploy the solution within a live marketing automation environment for closed-loop optimization.

---

6. Technologies Used

- Python (pandas, scikit-learn, imbalanced-learn, XGBoost)
- Jupyter Notebooks for development and experimentation
- Git for version control

Conclusion

CampaignInsight offers a flexible and reliable solution for predicting customer engagement in direct mail marketing. With a focus on real-world business application and model generalization, this pipeline can serve as a foundation for data-driven decision-making in customer outreach strategies.