Project 2 Report: The Sky's the Limit

Aadya Choudhary, Lambert Li
Wake Forest University

---

1. Introduction

This project focused on developing an effective machine learning model for a highly imbalanced dataset containing **1 million records and 32 features**. The goal was to accurately classify the **target variable (Target_Y)** while addressing the challenges of severe class imbalance.

To achieve this, we explored **three different modeling approaches**:

1. **Synthetic Data Approach**: Using **SMOTE (Synthetic Minority Over-sampling Technique)** to generate synthetic data for balancing.
2. **Non-Synthetic Data Approach**: Using **Random Undersampling (RUS)** to reduce the majority class without introducing synthetic data.
3. **Gradient Boosting Model**: An advanced ensemble technique that sequentially improves weak learners.

After conducting extensive evaluations, we found that **Random Forest with SMOTE performed the best**, achieving the highest **recall and precision balance**, making it the optimal model for deployment.

---

2. Data Preprocessing

2.1 Dataset Overview

The dataset consists of:

- **1,000,000 rows**
- **32 columns (31 features + 1 target variable)**

The **target variable, `Target_Y`, is binary**, where:

- **Class 0 (Negative cases)**: 988,971 samples (98.9%)
- **Class 1 (Positive cases)**: 11,029 samples (1.1%)

Due to this extreme imbalance, traditional machine learning models would struggle to **identify minority class cases accurately**.

2.2 Data Cleaning & Feature Engineering

- **Converted categorical variables** into **dummy variables** to make them compatible with machine learning models.
- **Removed unnecessary columns**, such as ID, that do not contribute to predictions.
- **Handled missing values** by identifying and filling gaps where necessary.

2.3 Handling Class Imbalance

To address the class imbalance, we experimented with two different resampling techniques:

- **SMOTE (Oversampling)**: Generates synthetic samples for the minority class to balance the dataset.
- **Random Undersampling (RUS)**: Reduces the size of the majority class while keeping the minority class unchanged.

After applying these techniques, we **split the dataset into training and testing sets (70%-30%)** for model evaluation.

---

3. Model Selection & Implementation

3.1 Decision Tree Classifier

Model Choice: A Decision Tree was used as a **baseline model** to compare performance against more advanced models.

Model Setup:

- `max_depth = 25`
- `min_samples_leaf = 10`
- `ccp_alpha = 0.001`

Performance Metrics:

Metric	Training Set	Test Set
Accuracy	95.31%	95.27%

Precision	95.78%	95.66%
Recall	83.33%	83.29%
F1-Score	89.12%	89.05%
ROC-AUC	91.22%	91.08%

Observation: The Decision Tree **performed decently** but struggled with **low recall**, making it unsuitable for detecting minority class instances.

3.2 Random Forest Classifier with SMOTE (Best Performing Model)

Model Choice:
Random Forest is a **robust ensemble learning method** that combines multiple decision trees for better accuracy and generalization.

Steps Taken:

1. Applied SMOTE to balance the dataset by generating synthetic samples for the minority class.
2. Split the dataset (70% train, 30% test) after applying SMOTE to avoid data leakage.
3. Hyperparameter tuning using RandomizedSearchCV to optimize the model.

Model Setup:

- `n_estimators = 150`
- `max_features = 6`
- `max_depth = None`
- `min_samples_leaf = 1`

Performance Metrics:

Metric	Training Set	Test Set
Accuracy	100.00%	99.29%
Precision	100.00%	99.62%

Recall	100.00%	96.29%
F1-Score	100.00%	97.93%
ROC-AUC	100.00%	99.67%

Observation: Random Forest with SMOTE was the best performer, achieving high recall and precision, making it the most reliable choice.

3.3 Gradient Boosting Classifier

Model Choice:
Gradient Boosting is an **ensemble learning method** that improves weak learners sequentially.

Performance Metrics:

Metric	Training Set	Test Set
Accuracy	98.08%	98.09%
Precision	98.29%	98.32%
Recall	97.01%	97.02%
F1-Score	97.65%	97.66%
ROC-AUC	96.51%	99.42%

Observation: Gradient Boosting performed well but was slightly weaker than Random Forest in recall.

3.4 Random Forest with Non-Synthetic Data (Random Undersampling)

Performance Metrics:

Metric	Training Set	Test Set
Accuracy	98.41%	**87.50%**
Precision	91.29%	**63.52%**
Recall	100.00%	**58.72%**
F1-Score	95.44%	**61.02%**
ROC-AUC	99.05%	**75.99%**

Key Finding:

- Without SMOTE, model performance dropped significantly.
- Recall dropped from 96.29% to 58.72%, making the non-synthetic model unreliable for minority class detection.

---

4. Key Takeaways & Recommendations

1. **SMOTE significantly improves model performance** – Synthetic oversampling balances the dataset effectively.
2. **Random Forest with SMOTE is the best model** – It consistently achieved the **highest recall and precision balance**.
3. **Non-synthetic methods underperform** – Undersampling **caused major performance drops**, proving that **SMOTE is necessary**.
4. **Gradient Boosting is effective but slightly weaker** – It performed well, but **Random Forest remains the superior choice**.
5. **Hyperparameter tuning is essential** – Fine-tuning **improves recall and prevents overfitting**.

---

5. Conclusion

Best Model for Submission: Random Forest with SMOTE and threshold tuning
Evaluation Focus: Maximizing Recall & Precision to minimize false negatives
Final Decision: Submit Random Forest with SMOTE as the final model for the leaderboard

Final Verdict: The SMOTE-enhanced Random Forest model with optimized thresholds is the most effective and reliable classification model for this dataset.