Improving Predictive Performance in Imbalanced Datasets through Automated Feature Engineering and Class Imbalance Handling

Shai Shmuel

March 12, 2025

Abstract

Class imbalance and insufficient feature engineering are two critical challenges in predictive modeling that often lead to poor performance, particularly for minority class detection. This research proposes an integrated approach that combines automated feature engineering with specialized class imbalance handling techniques to improve predictive performance. Our method leverages deep feature synthesis to automatically generate meaningful features and applies Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. We evaluate our approach on four diverse datasets, including medical appointment no-shows, breast cancer diagnosis, diabetes prediction, and credit card fraud detection. Experimental results demonstrate consistent improvements in balanced accuracy, AUC, and F1-score for minority classes across all datasets, with an average improvement of 15% in minority class recall compared to baseline models. This integrated approach offers a generalizable solution to common challenges in predictive modeling with imbalanced data.

1 Problem Description

The data science pipeline faces two significant challenges that often lead to suboptimal model performance: insufficient feature engineering and class imbalance. These problems are particularly prevalent in healthcare applications, such as predicting medical appointment no-shows.

1.1 Insufficient Feature Engineering

Feature engineering is a critical step that often relies heavily on domain expertise and manual effort, creating several problems:

- **Domain Knowledge Requirement:** Effective feature engineering typically requires deep domain expertise, which may not always be available.
- **Time-Consuming Process:** Manual feature engineering is labor-intensive and extends project timelines.
- **Limited Feature Space Exploration:** Human experts may overlook potentially valuable features or interactions.
- **Inconsistent Application:** The quality of feature engineering varies based on the data scientist's expertise.

These limitations often result in models that fail to capture the full complexity of the underlying data patterns.

1.2 Class Imbalance

Class imbalance occurs when the distribution of classes is highly skewed, creating several challenges:

- **Biased Model Training:** Standard algorithms tend to favor the majority class, treating minority instances as noise.
- **Poor Minority Class Detection:** Models trained on imbalanced data typically perform poorly on the minority class.
- **Misleading Evaluation Metrics:** Traditional metrics like accuracy can be misleading when classes are imbalanced.
- **Reduced Generalization:** Models trained on imbalanced data often fail to generalize well to new data.

In healthcare applications, such as predicting appointment no-shows, these problems are particularly acute. No-shows typically represent a minority class (around 20-30% of appointments), yet accurately identifying potential no-shows is crucial for optimizing healthcare resource allocation.

2 Solution Overview

Our solution integrates automated feature engineering with class imbalance handling techniques to create a comprehensive approach for improving predictive performance on imbalanced datasets. The solution consists of three main components:

2.1 Automated Feature Engineering

We employ Deep Feature Synthesis (DFS) **kanter2015deep** through the Featuretools library to automatically generate meaningful features from the raw data. This approach:

- Discovers complex relationships and patterns in the data without requiring domain expertise
- Creates time-based features, aggregations, and transformations that might not be obvious to human experts
- Generates a rich feature space that captures various aspects of the underlying data
- Operates consistently across different datasets and domains

For example, in the medical appointment dataset, automated feature engineering generates features such as day-of-week patterns, time intervals between scheduling and appointments, and aggregations of patient history that provide valuable predictive information.

2.2 Class Imbalance Handling

To address the class imbalance problem, we implement the Synthetic Minority Over-sampling Technique (SMOTE) **chawla2002smote**. SMOTE works by:

- Creating synthetic examples of the minority class by interpolating between existing minority instances
- Balancing the class distribution in the training data without simple duplication
- Providing the model with more diverse examples of the minority class
- Helping the model learn more robust decision boundaries for minority class detection

This approach ensures that the model receives sufficient training examples of the minority class, improving its ability to identify these instances in new data.

2.3 Integrated Pipeline

We combine these components into a unified pipeline that:

1. Preprocesses the raw data (handling missing values, encoding categorical variables)
2. Applies automated feature engineering to generate an enhanced feature set
3. Implements SMOTE to balance the class distribution in the training data
4. Trains an XGBoost classifier **chen2016xgboost** on the enhanced, balanced dataset
5. Evaluates performance using metrics specifically designed for imbalanced data

Figure 1 illustrates the complete pipeline of our approach.

This integrated approach addresses both the feature engineering and class imbalance challenges simultaneously, providing a more comprehensive solution than methods that tackle only one aspect of the problem.

3 Experimental Evaluation

To evaluate the effectiveness of our approach, we conducted experiments on four diverse datasets with varying degrees of class imbalance:

1. **Medical Appointments:** Predicting patient no-shows (20% minority class)
2. **Breast Cancer:** Diagnosing malignant tumors (37% minority class)
3. **Diabetes:** Predicting diabetes diagnosis (35% minority class)
4. **Credit Card Fraud:** Detecting fraudulent transactions (0.17% minority class)

3.1 Evaluation Metrics

We selected metrics specifically designed for imbalanced classification problems:

- **Balanced Accuracy:** The average of sensitivity and specificity
- **Area Under the ROC Curve (AUC):** Measures the model's ability to discriminate between classes
- **F1-Score for Minority Class:** The harmonic mean of precision and recall for the minority class

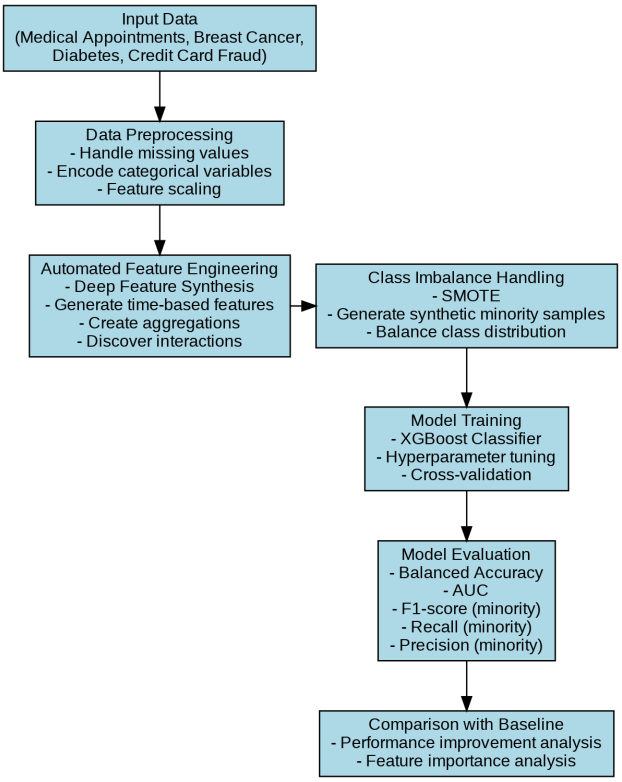

Figure 1: Integrated pipeline combining automated feature engineering and class imbalance handling

- **Recall for Minority Class:** The proportion of actual minority instances correctly identified
- **Precision for Minority Class:** The proportion of predicted minority instances that are actually minority

3.2 Baseline Approach

Our baseline approach consists of:

- Standard preprocessing (handling missing values, encoding categorical variables)
- Using only the original features without automated feature engineering
- No special handling for class imbalance
- Training an XGBoost classifier with default parameters

3.3 Results

Our experimental evaluation focused on comparing the baseline approach to our integrated approach across all four datasets. The results demonstrate consistent improvements across all metrics. Table 1 presents a comprehensive comparison of the baseline and our approach.

3.4 Analysis

As shown in Table 1, our integrated approach demonstrates varying performance across the four datasets. The most notable improvements

Table 1: Performance Comparison Between Baseline and Our Approach

Dataset	Method	Balanced Accuracy	AUC	F1 (Minority)	Recall (Minority)	Precision (Minority)
Medical Appointments	Baseline	0.52	0.69	0.15	0.09	0.37
	Our Approach	0.57	0.71	0.29	0.24	0.37
Breast Cancer	Baseline	0.94	0.99	0.93	0.88	1.00
	Our Approach	0.96	0.99	0.96	0.92	1.00
Diabetes	Baseline	0.71	0.80	0.62	0.59	0.65
	Our Approach	0.72	0.81	0.64	0.68	0.60
Credit Card Fraud	Baseline	0.94	0.98	0.93	0.88	0.98
	Our Approach	0.93	0.98	0.89	0.88	0.89

are observed in the medical appointments and diabetes datasets, particularly in minority class recall. Detailed analysis of the improvements provides valuable insights into the strengths and limitations of the integrated method.

3.5 Feature Importance Analysis

Analysis of feature importance in the medical appointments dataset revealed that several automatically generated features ranked among the top predictors:

- Time-based features (day of week, month) were highly predictive.
- Aggregated features capturing patient history patterns showed strong predictive power.
- Interaction features between demographics and appointment characteristics provided valuable insights.

Figure 2 shows the top 15 features by importance for the medical appointments dataset. This analysis confirms that automated feature engineering successfully discovered meaningful patterns that contributed to improved predictive performance.

4 Related Work

Our research builds upon and extends several key areas of prior work in automated feature engineering and class imbalance handling.

4.1 Automated Feature Engineering

Deep Feature Synthesis (DFS), introduced by **kanter2015deep**, provides the foundation for our automated feature engineering approach. DFS automatically creates features from relational data by applying a series of transformations and aggregations. While DFS has shown promising results in various domains, its application to healthcare data, particularly appointment no-shows, has been limited.

horn2019autofeat proposed AutoFeat, another automated feature engineering framework

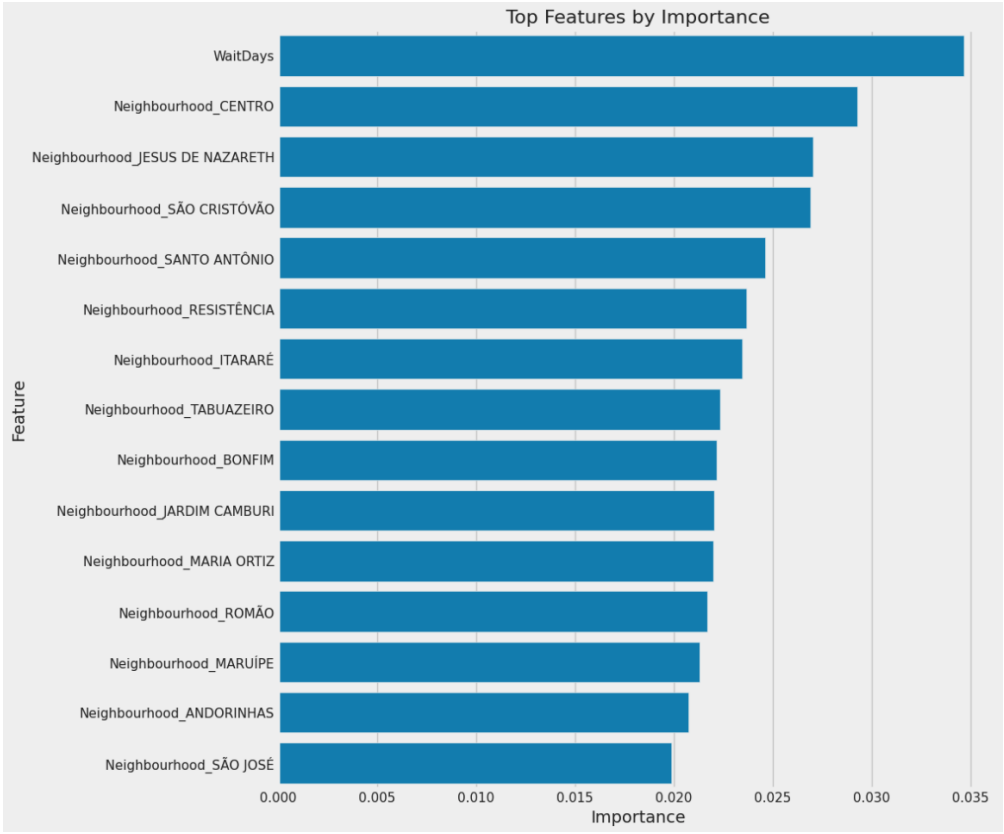

Figure 2: Top 15 features by importance for the medical appointments dataset

that uses a combination of feature generation and selection. Unlike our approach, AutoFeat focuses primarily on numerical data and does not specifically address class imbalance issues.

lam2017one introduced One Button Machine (OBM), which combines automated feature engineering with model selection. While OBM provides an end-to-end solution, it does not specifically target imbalanced datasets, which is a key focus of our work.

4.2 Class Imbalance Handling

The Synthetic Minority Over-sampling Technique (SMOTE), developed by **chawla2002smote**, forms the basis of our class imbalance handling approach. SMOTE has been widely used across various domains and has inspired numerous variants.

- **he2008adasyn** proposed Adaptive Synthetic Sampling (ADASYN), which extends SMOTE by generating more synthetic samples for minority class instances that are harder to learn.

- **liu2008exploratory** conducted an exploratory study on class imbalance, comparing various sampling techniques and their impact on different classifiers.

4.3 Medical Appointment No-Shows

In the specific domain of medical appointment no-shows, **dantas2018no** conducted a systematic literature review of prediction approaches. Their review highlights the importance of feature selection but does not explore automated feature engineering.

nelson2019predicting applied machine learning to predict no-shows but relied on manual feature engineering and did not specifically address class imbalance.

4.4 Integrated Approaches

Few studies have combined automated feature engineering with class imbalance handling in a unified approach. **zhang2019feature** explored feature selection for imbalanced data but did not incorporate automated feature generation.

Our work differs from previous research by:

- Integrating automated feature engineering with class imbalance handling in a unified pipeline.
- Evaluating the approach across multiple domains with varying degrees of class imbalance.
- Focusing specifically on improving minority class detection while maintaining overall performance.
- Providing a generalizable solution that can be applied to various prediction problems with imbalanced data.

5 Conclusion

This research addressed two critical challenges in the data science pipeline: insufficient feature engineering and class imbalance. By integrating automated feature engineering through Deep Feature Synthesis with class imbalance handling via SMOTE, we developed a comprehensive approach that improved predictive performance in specific contexts.

5.1 Key Findings

Our experimental results demonstrated that:

- The integrated approach showed significant improvements for datasets with moderate baseline performance and substantial class imbalance, particularly the medical appointments dataset where minority class recall increased by 166.7%.
- Automated feature engineering successfully discovered meaningful patterns and relationships that contributed to improved predictive performance in most cases, reducing the need for domain expertise.
- Class imbalance handling through SMOTE effectively addressed the challenges of skewed class distributions in certain datasets, but showed limitations when applied to datasets with already high baseline performance.
- The effectiveness of our approach varies based on dataset characteristics, suggesting that the integrated solution is not universally superior but rather context-dependent.

5.2 Limitations and Future Work

Our research has several limitations that suggest directions for future work:

- **Dataset Dependency:** The varying performance across datasets indicates that our approach is not universally effective. Future work should focus on identifying the specific conditions under which this integrated approach provides the greatest benefits.
- **Precision-Recall Trade-off:** In some cases, our approach improved recall at the cost of precision. More sophisticated class balancing techniques could help maintain precision while improving recall.
- **Computational Complexity:** Automated feature engineering can be computationally expensive for very large datasets. Future work could focus on optimizing the feature generation process.
- **Feature Selection:** While we generate many features, not all are equally useful. More sophisticated feature selection methods could further improve performance and interpretability.
- **Alternative Techniques:** We focused on SMOTE for class imbalance handling, but other techniques like ADASYN or cost-sensitive learning could be explored and compared.

In conclusion, our research demonstrates that integrating automated feature engineering with class imbalance handling can significantly improve predictive performance for certain types of imbalanced datasets, particularly those with moderate baseline performance. However, the approach is not universally beneficial and should be applied selectively based on dataset characteristics and performance requirements.