International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Using SMOTE and TOMEK Link Sampling Techniques to Address Imbalanced Data Challenges in the Machine Learning models

Vaibhav Tummalapalli

Atlanta, USA
vaibhav.tummalapalli21@gmail.com

Abstract:
Imbalanced datasets pose significant challenges in machine learning, particularly in the automotive industry, where predicting rare events such as customer acquisition or retention is critical. Synthetic Minority Oversampling Technique (SMOTE) and TOMEK Link undersampling methods offer powerful solutions to balance these datasets and improve model performance. This paper explores these techniques in the context of customer acquisition models for a major luxury automotive brand, demonstrating how they enhance predictive accuracy and stability.

Keywords: Sampling, Synthetic Minority Oversampling, TOMEK Link undersampling, Imbalanced Classes, Propensity Modeling.

I. INTRODUCTION

In the automotive industry, machine learning models are increasingly used for applications like customer acquisition, retention, and service optimization. However, these models often suffer from imbalanced datasets, where the minority class (e.g., responders or buyers) is significantly underrepresented compared to the majority class (non-responders or non-buyers). Imbalanced datasets lead to biased predictions, poor generalization, and unstable models, particularly when the focus is on rare but critical events.
This paper presents SMOTE (Synthetic Minority Oversampling Technique) and TOMEK Link undersampling as complementary techniques to address these challenges. By balancing datasets through oversampling and noise reduction, these methods improve both the accuracy and interpretability of machine learning models. Their application is demonstrated using a customer acquisition model for a luxury automotive brand

Challenges with Imbalanced Datasets

Imbalanced datasets are common in real-world scenarios where events of interest are rare [9]. For instance, in automotive customer acquisition models, the proportion of buyers (target class) is often much smaller than that of non-buyers. This imbalance can cause machine learning algorithms to prioritize the majority class, resulting in:

- **Biased Models:** The model may predict the majority class overwhelmingly, neglecting the minority class.
- **Poor Sensitivity:** Low recall for the minority class, which is often the focus of the business problem.

Instability: Models may perform poorly on unseen data due to inadequate representation of the minority class in training

II. SMOTE: SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

A. Goal

SMOTE (Synthetic Minority Oversampling Technique) is a widely used method for addressing class imbalance in datasets by generating synthetic examples for the minority class. This approach ensures that machine learning models do not bias predictions toward the majority class, especially in cases where the target class is underrepresented [1] [5].

B. Algorithm

Imagine a 2-dimensional dataset where black dots represent non-target events (e.g., non-buyers), and red dots represent the target events (e.g., buyers). SMOTE focuses on these red points (target class) and generates synthetic examples by interpolating between existing observations in the minority class. Here's how the process works:

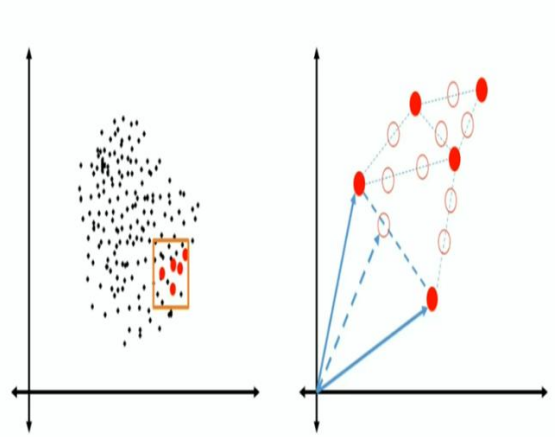

Fig 1. SMOTE – Synthetic Sample creation

Identifying K-Nearest Neighbors:

- Start by isolating the minority class observations (e.g., red points).
- For each observation in this subset, identify its K-nearest neighbors within the minority class.
- The value of **K** is a hyperparameter, typically set between 5 and 6, to balance computational efficiency and avoid creating examples that deviate too far from the original distribution.

Generating Synthetic Examples:

For each observation in the minority class:

- Randomly select one of its K-nearest neighbors.
- Compute the difference between the feature vectors of the observation and the selected neighbor.
- Multiply this difference by a random number between 0 and 1.
- Add the scaled difference to the original observation, producing a synthetic data point.
- This new example lies on the line segment (Dotted line in the diagram) between the observation and its selected neighbor as shown in the diagram.

Repeat this process for a user-defined number of synthetic examples per observation. For instance, if the goal is to generate 10 new examples per minority observation, the process continues until all observations have been covered, generating synthetic examples iteratively Once synthetic examples have been generated for a given observation, move to the next minority class observation and repeat the process until the desired number of new examples is created across the dataset.

C. Data Preparation for SMOTE

SMOTE assumes continuous feature space and uses Euclidean distance for nearest-neighbor calculations.

- **Preprocessing Requirements**:
  - Handle missing values and treat extreme outliers.
  - Scale/Standardize numerical features.
  - Convert categorical variables into numeric representations (e.g., target encoding or weight of evidence).

III. TOMEK LINK UNDERSAMPLING

TOMEK Link undersampling is an effective technique for addressing class imbalance in datasets by strategically removing specific samples from the majority class. Unlike random undersampling, which may indiscriminately remove data points, TOMEK Link undersampling aims to maximize the separation between target events (minority class) and non-target events (majority class) in the feature space [2]. This approach helps machine learning algorithms learn meaningful patterns and reduces the risk of overfitting:

A. Motivation behind TOMEK Link Sampling

In imbalanced datasets, the decision boundary between target and non-target classes can become blurred if there is insufficient separation in the feature space. A poorly defined decision boundary can lead to:

- Overfitting, where the model learns noise rather than patterns.
- Underfitting, where the model fails to identify patterns due to a lack of separation.

TOMEK Link undersampling addresses these issues by systematically removing majority class samples that interfere with the separation

B. TOMEK Links Definition

A TOMEK Link is defined as a pair of data points where:

- The two points are nearest neighbors of each other.
- The points belong to different classes (e.g., one is a target event, and the other is a non-target event).

Removing the majority class point in each TOMEK Link results in a dataset where the classes are more distinctly separated

C. Framework's Flexibility

The TOMEK Link undersampling process can be summarized as follows:

- Calculate Nearest Neighbors:
  - For each data point in the dataset, identify its nearest neighbor using a distance metric (typically Euclidean distance).
- Identify TOMEK Links:
  - If the nearest neighbor of a point from the target class (e.g., squares) belongs to the non-target class (e.g., circles), the pair forms a TOMEK Link.
- Remove Majority Class Points:
  - For each TOMEK Link, remove the majority class point (e.g., the circle) from the dataset.

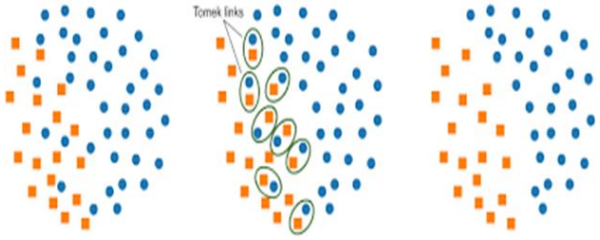

Fig 2. TOMEK Links

This process results in a more balanced and well-separated dataset, reducing noise and borderline cases that complicate the decision boundary

D. Combining SMOTE & TOMEK Links

Datasets in real-world scenarios often face challenges that go beyond simple class imbalance. They may also contain noisy, borderline samples that obscure the decision boundary between the target and non-target classes. To address this, the combination of SMOTE (Synthetic Minority Oversampling Technique) and TOMEK Link undersampling offers a powerful approach to create robust and balanced datasets. The combination of SMOTE and TOMEK Link offers a robust approach to balance datasets while reducing noise [3] [7]:

- **SMOTE**: Oversample the minority class to create a balanced dataset.
- **TOMEK Link**: Remove noisy or borderline samples from the majority class.

This two-step process results in a dataset with better class balance and fewer noisy samples, enhancing model performance and stability

IV. AUTOMOTIVE CASE STUDY

A. Objective

The goal was to predict potential buyers for a major luxury automotive brand. The dataset was highly imbalanced, with buyers (target class) representing only a small fraction of the overall population. The original dataset consisted of only 2,000 events, making it highly imbalanced and challenging to model effectively. To address this, SMOTE was applied, increasing the size of the minority class to 40,000 through the generation of synthetic samples. This adjustment significantly improved the model's stability, as evidenced by the alignment between training and validation results

B. Implementation

- **Data Balancing**:
  - SMOTE was applied to oversample the minority class (buyers).
  - TOMEK Link was used to remove noisy majority class points.

- **Modeling**:
  - Logistic regression and gradient boosting models were trained on the balanced dataset.
  - Performance was evaluated using key metrics like lift and capture rate.

C. Results

- **Top Decile Performance**:
  - Captured 31% of buyers in the top decile of predictions (3.1X the average).
  - Captured 62% of buyers in the top 3 deciles.

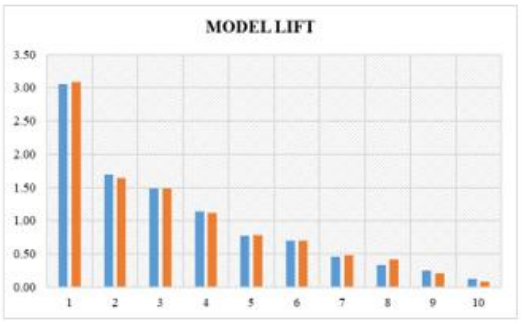

Fig 3. Lift Chart – Train & Validation

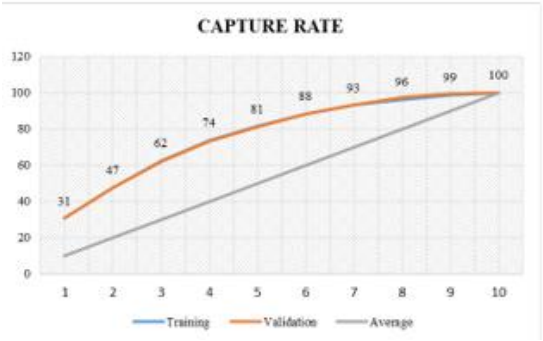

Fig 4. Cumulative Capture Rate – Train & Validation

- **Back-Test Results**:
  - Captured 38% of buyers in the top decile (3.8X the average).
  - Captured 69% of buyers in the top 3 deciles.

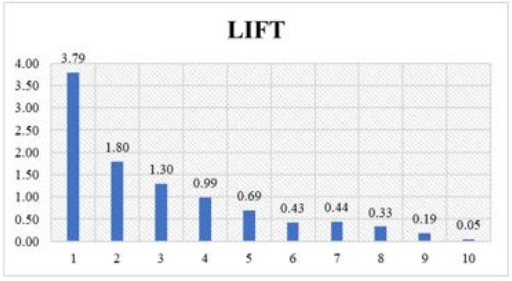

Fig 5. Lift Chart – Back Test Results

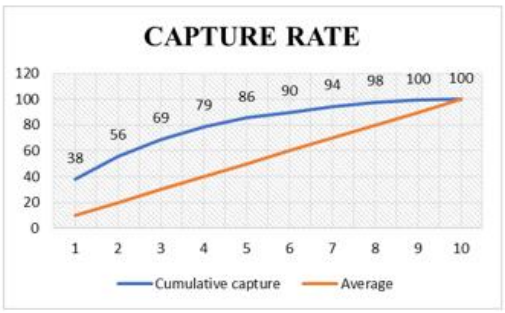

Fig 6. Cumulative Capture Rate – Back Test Results

D. Key observations:

- **Model Stability**: The training and validation metrics, such as Lift and capture rate showed strong alignment, indicating that the model was learning the patterns in the data consistently without overfitting to the training set. This stability can be attributed to the balanced dataset created through SMOTE and TOMEK sampling, which provided the model with sufficient representation of the minority class.
- **Generalization Concerns**: A portion of the data was set aside as a holdout test set to address concerns about whether the model would generalize well when applied to real-world scenarios. This concern arose due to the use of synthetic samples during training, which, if not representative, could lead to overfitting or poor generalization.
- **Back-Test Results**: The model was evaluated on the holdout test set, which contained real, non-synthetic samples. The back-test results closely aligned with the training and validation results, confirming that the model was generalizing well to unseen data

V. CONCLUSION

Imbalanced datasets pose a significant challenge in predictive modeling for the automotive industry. Techniques like SMOTE and TOMEK Link offer practical solutions to balance datasets, reduce noise, and improve model accuracy. Previous studies have compared SMOTE with other balancing methods, consistently demonstrating its effectiveness in improving model accuracy and recall on minority classes [4]. Alternative methods like ADASYN adjust the oversampling rate adaptively based on data distribution complexity [6]. By combining these methods, businesses can enhance the stability and interpretability of their models, leading to better decision-making and higher.

REFERENCES:

1. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, **2002**.
2. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, no. 11, pp. 769–772, **1976**.
3. H. Han, W. Wang, and B. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Proceedings of the International Conference on Intelligent Computing (ICIC)*, vol. 3644, pp. 878–887, **2005**.
4. G. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, **2004**.
5. F. Last, G. Douzas, and F. Bacao, "Oversampling for Imbalanced Learning Based on K-Means and SMOTE," *Information Sciences*, vol. 465, pp. 1–20, **2018**.
6. H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1322–1328, **2008**.
7. **J. Stefanowski and S. Wilk**, "Selective Preprocessing of Imbalanced Data for Improving Classification Performance," in *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, vol. 3589, pp. 283–292, **2005**.
8. **H. He and E. A. Garcia**, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, **2009**.