Financial Risk Prediction Using Logistic Regression and importance of SMOTE

Nagendar Goud Pagudala

August 7th, 2024

1 Introduction

The goal of this project was to predict obesity levels using a dataset that includes a range of features related to lifestyle, diet, physical activity, and more. Obesity is a significant public health issue, as it is associated with various health conditions, including cardiovascular disease, diabetes, and certain types of cancer. The dataset used in this study includes demographic, behavioral, and health-related features, with the target variable being the obesity status of individuals categorized into different classes.

To address the problem of predicting obesity levels, I applied several exploratory data analysis (EDA) techniques, followed by machine learning modeling. The process involved understanding the relationships between features using heatmaps, distributions, and boxplots, and then utilizing a Logistic Regression (LoR) model for classification. The imbalanced nature of the target variable was addressed using the Synthetic Minority Over-sampling Technique (SMOTE).

The results of this project aim to show how balancing the classes using SMOTE can impact the performance of a Logistic Regression model and provide a comparison of key metrics before and after this technique.

2 Problem Definition

- **Classification:** Predicting whether a loan will be approved or denied based on applicant features.

3 Machine Learning model

For the classification task, I utilized Logistic Regression (LoR), which is a simple yet powerful method for binary and multi-class classification problems. The model predicts the probability that an input belongs to a certain class. In this case, the LoR model will predict the obesity status based on various features.

The primary algorithm used is Logistic Regression (LoR). The pipeline includes:

1. **Data Preprocessing:** Handling missing values using `missingno` and addressing outliers.

2. **Baseline Model:** Logistic Regression on the original (imbalanced) dataset.

3. **Enhanced Model:** Logistic Regression after applying SMOTE.
4. **Evaluation:** Evaluating model performance using metrics such as accuracy, precision, recall, and F1 score.

4 Methodology and Experimental Evaluation

4.1 Data Visualizations:

- Heatmap to identify correlations between features.
- Variable distributions for features with correlations between 0.60 and 0.99.
- Boxplots for detecting and visualizing outliers.
- Missing value matrix using `missingno`.
- Class distribution of the target variable before and after SMOTE.

4.2 Hypothesis Testing:

- Null Hypothesis (H0): Balancing the class distribution with SMOTE will not significantly improve the performance of the Logistic Regression model.
- Alternative Hypothesis (H1): Balancing the class distribution with SMOTE will improve the performance of the Logistic Regression model, particularly for minority classes.

4.3 Evaluation Metrics:

To evaluate the performance of the Logistic Regression model, several metrics were considered, including accuracy, precision, recall, and F1 score. Before addressing class imbalance, I trained and tested the Logistic Regression model using the original dataset and evaluated the metrics.

4.4 Results

4.4.1 Initial Model Performance (Before SMOTE)

- **Accuracy Score:** 0.9985
- **AUC Score:** 0.9998
- **Confusion Matrix:**

$$\begin{bmatrix} 5029 & 1 \\ 9 & 1561 \end{bmatrix}$$

- **Classification Report:**

Class	Precision	Recall	F1-score
0	1.00	1.00	1.00
1	1.00	0.99	1.00

- **TPR, TNR, FPR, FNR:** 0.9943, 0.9998, 0.0002, 0.0057

4.4.2 Enhanced Model Performance (After SMOTE)

- **Accuracy Score:** 0.9993
- **AUC Score:** 0.99997
- **Confusion Matrix:**

$$\begin{bmatrix} 4980 & 4 \\ 3 & 5059 \end{bmatrix}$$

- **Classification Report:**

Class	Precision	Recall	F1-score
0	1.00	1.00	1.00
1	1.00	1.00	1.00

- **TPR, TNR, FPR, FNR:** 0.9994, 0.9992, 0.0008, 0.0006
- **Cross-Validation AUC Scores:** [0.9516, 0.9493, 0.9545, 0.9543, 0.9480]
  Mean AUC: 0.9515

4.5 Discussion

The hypothesis that SMOTE would improve model performance is supported by the results. Specifically, balancing the class distribution allowed the Logistic Regression model to perform better on the minority classes, as seen in the increased recall and F1 score after SMOTE. The original model had lower recall for certain classes, particularly Obesity Type I and Overweight Level II, which were underrepresented in the dataset.

The results suggest that addressing class imbalance is crucial for improving model performance in such classification tasks. Future work may explore the use of other oversampling techniques or more complex models like Random Forests or Neural Networks to further improve predictions.

5 Conclusion

This project demonstrated the importance of addressing class imbalance in financial risk prediction. SMOTE significantly improved the model's ability to generalize, leading to fairer and more reliable predictions. Logistic Regression performed exceptionally well, achieving near-perfect metrics across all evaluation criteria.