CSI5155 - Fall 2024

Assignment 1 – Supervised Learning

<Vrishab Prasanth Davey 300438343>

Overview

A highly Skewed Chocolate Dataset and a balanced Mushroom dataset were analyzed using six machine learning models: Decision Tree, Random Forest, SVM, Gradient Boosting, MLP, and k-NN. In addition to no rebalancing, each model was tested with undersampling, oversampling (SMOTE), and mixed sampling. The main objective was to evaluate each model on imbalanced datasets and how rebalancing affected precision, recall, F1-score, and AUC. A detailed explanation has been provided in the subsequent paragraphs

The Chocolate Dataset(Imbalanced Dataset)

As the Chocolate dataset was skewed the models struggled with class imbalance without rebalancing. Gradient Boosting ranked best with an AUC of 0.75, showing its ability to handle skewed data without rebalancing. SVM and MLP models displayed extremely low AUC values of 0.19 and 0.24, respectively, which is a consequence of class imbalance sensitivity. Despite these AUC disparities, most models had decent precision and recall, resulting in striking F1-scores around 0.99, indicating that they predicted the majority class.

Undersampling balanced class distribution by lowering majority class occurrences, making performance more realistic. SVM and MLP improved their AUC values measuring 0.42 and 0.46, respectively. On one hand, Random Forest improved, but Decision Tree's AUC dropped to 0.58 and Gradient Boosting's AUC dropped to 0.53, respectively. Though the minority class estimation improved the accuracy took a huge tradeoff for the recall value

Most models, performed well, with MLP achieving the highest AUC value of 70. In comparison to Undersampling the models have outperformed them w.r.t the AUC values, and the F1 scores have also significantly improved. While oversampling reduced majority class bias, synthetic data resulted in overfitting likely due to the confusion amongst ensemble models like Random Forest and Gradient Boosting

In the combined sampling strategy, which used oversampling and undersampling, most of the models did take a hit on their AUC values, Decision Tree and Random Forest suffered, indicating that the combined method generated too much noise and reduced their effectiveness. The Precision value of most of the models was really good but the recall value did not meet the mar reducing the F1 score, but MLP did maintain a high with an F1 score of 0.95

Magic Mushroom Dataset(Balanced Dataset)

Models outperformed the Chocolate dataset as the former was more balanced than the latter, using all rebalancing strategies. Both Random Forest and SVM attained AUC values of 0.83 without rebalancing, demonstrating their capacity to handle balanced data. MLP and Gradient Boosting had an AUC value of 0.82 respectively, followed by the Decision Tree with an an AUC of 0.81. The precision, recall, and F1-scores were hovering around 0.65 to 0.70, suggesting they could handle the Mushroom dataset's class distribution better.

Undersampling decreased model performance, however, SVM, Random Forest, and MLP maintained AUC values of 0.82 and 0.83. The recall value of most of the models was commendable, Undersampling didn't really make a difference as the Mushroom dataset was more balanced.

Oversampling made the model unstable as the AUC value took a steer drop. Random Forest, MLP, and k-NN performed well, with AUC values between 0.75 and 0.77. Gradient Boosting dropped to 0.75, suggesting it overfitted on fake samples. Decision Trees and RF showed good recall values. Oversampling somewhat gave a mixed feeling as the synthetic data added noise to the Decision Tree and Gradient Boosting training.

In the combined sampling The DT showed a significant decline in its AUC value reaching a low of 0.64, other models did maintain a good range between 0.80 and 0.82, and RF and SVM show that they can differentiate well enough between the undersampling and oversampling not being susceptible to noise.

Conclusion

The more imbalanced Chocolate dataset improved more with oversampling, while the more balanced Mushroom dataset performed well across all rebalancing strategies. SVM and MLP performed best across both the datasets and resampling methods. Due to synthetic data overfitting, Random Forest and Gradient Boosting performed well but had trouble oversampling and combination sampling. Decision Tree was extremely sensitive to rebalancing approaches, especially coupled sampling, where noise degraded performance. k-NN performed well with all the resampling techniques. As discussed in the lectures the concept of 'Free Lunch' can be applied here, It is not a good approach to choose a particular method and call that the 'Ideal' method as the distribution of the dataset and the model's tolerance to noise should be considered to determine the best approach.