Machine Learning Final Project

Syed Hamza Jamil : 29454

Note

In this PROJECT there are three different section with 3 different dataset with 3 different Class imbalance technique including subsection like below

Section 1 A B C D : Section 2 A B C D : Section 3 A B C D

In last complete project summary

Class Imbalance Selected Techniques

- SMOTE (Synthetic Minority Over-sampling Technique): Generates synthetic samples for the minority class to balance the dataset, improving model performance on imbalanced data by providing more instances of the minority class.

- Class Weight Adjustment: Assigns higher weights to minority class samples during model training, allowing the model to pay more attention to these underrepresented classes without altering the dataset.

- Threshold Adjustment: Modifies the decision threshold for class predictions to favor the minority class, helping to balance precision and recall by adjusting the sensitivity of the classifier to different classes

Why SMOTE (Synthetic Minority Over-sampling Technique)

I selected SMOTE to generate synthetic samples for the minority class, which helps balance the dataset and provides the model with more instances of the underrepresented class, leading to improved performance on imbalanced data.

Why Class Weight Adjustment

I chose Class Weight Adjustment to assign higher weights to minority class samples during model training. This technique allows the model to focus more on the underrepresented classes without altering the original dataset, enhancing its ability to learn from these samples.

Why Threshold Adjustment

Threshold Adjustment was selected to modify the decision threshold for class predictions. By adjusting the sensitivity of the classifier, this technique helps balance precision and recall, favoring the minority class and improving overall model performance on imbalanced data.

Instructions

- In this project I create a single master function that have a ability to perfrom complete data science life cycle on different dataset
- There are 3 different section in this projects, in each section,I used different ML technique with different dataset and also compare different ML alogithm accuracy on different dataset, all the selected datasets are imbalance,the technique that we are used are as belows :
- Explanatory Data Analysis
- Data Cleaning
- Data Scaling
- Feature Selection Chi-Square
- Class Imbalance Technique : Smote, Class Weight, Threshold
- Cross Validation with Hyperparameter Tunning
- ML Algorithm : KNN, Logistic, SVM, Random Forest, Naive Bysian
- Model accuracy matrics including classification report and Roc Curve Plot
- Model Accuracy Comparision

Dataset 1 : Bank Marketing

- Dataset URL : https://archive.ics.uci.edu/dataset/222/bank+marketing
- Dataset Information : 45210 rows and 17 columns with Two class

Section 1 A: Summary of Baseline Model Insights

The KNN model is 88% accurate but struggles with the minority class, achieving low precision and recall. The Logistic Regression model is slightly better at 89% accuracy but is biased towards the majority class, with poor performance on the minority class. The SVM model, with 88% accuracy, is highly imbalanced, favoring the majority class almost entirely. The Naive Bayes model has an 84% accuracy and shows some capability in handling class imbalance but still favors the majority class. The Random Forest model, with 90% accuracy, performs best overall, showing a more balanced performance between the majority and minority classes, though it still leans towards the majority.

Section 1 B: Summary of Model Insights with Resampling SMOTE Technique

The K-Nearest Neighbors (KNN) model, using the SMOTE resampling technique, achieved 85% accuracy and a strong AUC of 0.91, indicating balanced performance across both classes. Logistic Regression, with an accuracy of 81%, shows good performance on the minority class and an AUC of 0.89, reflecting robust model behavior. The Support Vector Machine (SVM) also reached 81% accuracy and a similar AUC of 0.89, demonstrating strong discriminative power. The Naive Bayes model, with 70% accuracy and an AUC of 0.82, performed moderately well, particularly for the minority class. The Random Forest model, with an accuracy of 85% and the highest AUC of 0.92, displayed excellent performance and balance between the classes, indicating the effectiveness of the SMOTE technique in addressing class imbalance.

Section 1 C: Summary of Model Insights with Class Weight Technique

The K-Nearest Neighbors (KNN) model, using class weighting and optimal parameters, achieved 89% accuracy and an AUC of 0.82, showing high precision for the majority class but lower precision for the minority class. Logistic Regression, with 83% accuracy, has strong discriminative power (AUC of 0.89) but lower precision and recall for the minority class. The Support Vector Machine (SVM), achieving 84% accuracy and an AUC of 0.89, demonstrates balanced precision but lower recall for the minority class. The Naive Bayes model, with 87% accuracy and an AUC of 0.83, performs well for the majority class but has lower recall for the minority class. The Random Forest model, with 86% accuracy and an AUC of 0.79, shows balanced precision but lower recall for the minority class, indicating moderate discriminative ability.

Section 1 D: Summary of Model Insights with Threshold Technique

The K-Nearest Neighbors (KNN) model, using cross-validation and a 0.5 threshold, achieved 89% accuracy and an AUC of 0.80, showing high precision for the majority class but lower precision for the minority class. Logistic Regression, with 90% accuracy and an AUC of 0.88, demonstrates high precision for the majority class and better precision and recall for the minority class compared to KNN. The Support Vector Machine (SVM), also with 90% accuracy and an AUC of 0.83, shows high precision for the majority class but significantly lower recall for the minority class. The Naive Bayes model achieved 87% accuracy and an AUC of 0.83, performing well for the majority class but with lower recall for the minority class. The Random Forest model, with 88% accuracy and an AUC of 0.83, shows balanced precision for both classes but lower recall for the minority class, indicating moderate discriminative ability.

Dataset 2 : Default of Credit Card Clients

- Dataset URL : https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
- Dataset Information : 29999 rows and 24 columns with Two class

Section 2 A: Summary of Baseline Model Insights

The Random Forest model performs the best, with a testing accuracy of 81% and an AUC of 0.76, showing good balance in predicting both classes. Logistic Regression and SVM have moderate accuracy (78%) but fail to predict the minority class. Naive Bayes, with a lower overall accuracy (38%), performs better on the minority class. KNN shows moderate performance with an AUC of 0.61. Overall, Random Forest is the most suitable model for prediction in this scenario.

Section 2 B: Summary of Model Insights with SMOTE Technique

The K-Nearest Neighbors (KNN) model, using SMOTE, achieved a training accuracy of 78% and a testing accuracy of 80%, with strong performance for both classes and an AUC of 0.86. Logistic Regression, with 66% testing accuracy and an AUC of 0.72, shows balanced but lower overall performance. The SVM model, with 70% testing accuracy and an AUC of 0.75, has good precision and recall for both classes. Naive Bayes, with 58% accuracy, performs better for the minority class and has an AUC of 0.73. The Random Forest model stands out with 84% accuracy and an AUC of 0.91, showing high precision and recal.

Section 2 C: Summary of Model Insights with Class Weight Technique

The K-Nearest Neighbors (KNN) model achieved a training accuracy of 80% and a testing accuracy of 79%, performing well for the majority class but showing lower performance for the minority class, with an AUC of 0.70. Logistic Regression achieved 68% accuracy, showing balanced but lower performance overall, with an AUC of 0.72. The Support Vector Machine (SVM) model had 77% training accuracy and 78% testing accuracy, performing well for the majority class and moderately for the minority class, with an AUC of 0.75. The Naive Bayes model achieved 72% accuracy, performing moderately for both classes, with an AUC of 0.73. The Random Forest model had 81% accuracy, showing strong performance for both classes and an AUC of 0.75, indicating good overall performance.

Section 2 D: Summary of Model Insights with Threshold Technique

The K-Nearest Neighbors (KNN) model achieved an average training accuracy of 80% and a testing accuracy of 79%, performing well for the majority class but showing lower performance for the minority class, with a weighted average f1-score of 0.78. Logistic Regression achieved 81% accuracy, showing high precision and recall for the majority class but lower performance for the minority class, with a weighted average f1-score of 0.77. The Support Vector Machine (SVM) model achieved 82% accuracy, performing well for the majority class and moderately for the minority class, with a weighted average f1-score of 0.79. The Naive Bayes model, with 72% accuracy, performed moderately for both classes, with a weighted average f1-score of 0.74. The Random Forest model achieved 82% accuracy, showing strong performance for the majority class and moderate performance for the minority class, with a weighted average f1-score of 0.80, indicating balanced overall performance.

Dataset 3 : Customer Churn

- Dataset URL : https://www.kaggle.com/datasets/hassanamin/customer-churn
- Dataset Information : 900 rows and 10 columns with Two class

Section 3 A: Summary of Baseline Model Insights

The K-Nearest Neighbors (KNN) model performs well for the majority class but fails completely for the minority class, with an AUC of 0.50 indicating poor overall performance. Logistic Regression and Support Vector Machine (SVM) also perform well for the majority class but poorly for the minority class, with moderate AUCs of 0.69 and 0.59, respectively. Naive Bayes stands out with balanced performance for both classes and a strong AUC of 0.89. Random Forest, while achieving perfect training accuracy, shows strong overall performance with an AUC of 0.86 and reasonable performance for the minority class. Overall, Naive Bayes and Random Forest are the most balanced and effective models in this scenario.

Section 3 B: Summary of Model Insights with SMOTE Technique

The K-Nearest Neighbors (KNN) model, tuned with 5 neighbors and distance-based weighting, achieved high performance with 90% accuracy for both training and testing, and an AUC of 0.96, indicating strong discriminative power. Logistic Regression, with 85% testing accuracy and an AUC of 0.92, shows balanced precision and recall. The Support Vector Machine (SVM) model, using an RBF kernel, achieved 86% testing accuracy and an AUC of 0.94, reflecting robust performance. Naive Bayes, with 84% accuracy and an AUC of 0.92, shows competent classification ability. The Random Forest model stands out with 93% testing accuracy and an AUC of 0.98, demonstrating exceptional discriminative ability and overall performance.

Section 3 C: Summary of Model Insights with Class Weight Technique

The K-Nearest Neighbors (KNN) model shows good performance with 89% testing accuracy and an AUC of 0.89, though it has lower precision for class 1. Logistic Regression, with 82% testing accuracy and an AUC of 0.93, demonstrates strong classification but low precision for class 1. The Support Vector Machine (SVM) model, with 86% testing accuracy and an AUC of 0.93, shows robust discriminative power but moderate recall for class 1. Naive Bayes achieves balanced performance with 90% accuracy and an AUC of 0.93, having reasonable precision for both classes. Random Forest, with 88% accuracy and an AUC of 0.90, reflects solid performance but lower precision for class 1. Overall, Logistic Regression, SVM, and Naive Bayes show strong classification abilities with the class weight technique.

Section 3 D: Summary of Model Insights with Threshold Technique

The K-Nearest Neighbors (KNN) model shows good performance with 89% testing accuracy and an AUC of 0.89, though it has lower precision for class 1. Logistic Regression achieves 91% testing accuracy and an AUC of 0.93, demonstrating high precision for both classes but lower recall for class 1. The Support Vector Machine (SVM) model, with 90% testing accuracy and an AUC of 0.93, shows robust discriminative power and balanced precision. Naive Bayes, with 90% accuracy and an AUC of 0.93, provides balanced performance with moderate precision for class 1. Random Forest, with 88% accuracy and an AUC of 0.90, shows solid performance but lower precision for class 1. Overall, Logistic Regression, SVM, and Naive Bayes exhibit strong classification abilities with the threshold technique.

OVERALL PROJECT SUMMARY

After implementation all different class imbalance technique with different ML algorithm, I analyze that in most of case CI resampling technique work very good, if we compare with different CI technique like class weight, threshold etc and from ML algorithm random forest work very good and KNN also slightly better, both algorithm has also own ability to work or handle class imbalance techniques