Name: Shukurov Shamil

Project title: Exploring Class Imbalance Solutions: Investigating the Effectiveness of Data Balancing Techniques on Model Performance.

What are you going to do?

In this research project my goal is to investigate the data balancing techniques to overcome imbalanced data problem in ML. I aim to find how data balancing influences evaluation metrics, model calibration and generalization capability of model? I will investigate different types of balancing techniques on multiple datasets with different degrees of imbalance. Some algorithms like LightGBM, XGBoost deals with class imbalance problem internally. I will investigate their effectiveness as well. Furthermore, I will explore how different ml algorithms reacts to balancing techniques, i.e. how their performance change.

How is it done today? Current Limitations?

Currently, data balancing techniques are used extensively to address class imbalance in machine learning. There is, however, no one-size-fits-all solution, and the efficiency of various methods can vary across datasets and algorithms. Moreover, the interaction between data balancing and algorithm-specific techniques, as well as the interpretability aspects, require additional research.

What is your idea to do something better?

My research aims to go beyond existing studies by conducting a comprehensive analysis of data balancing techniques in imbalanced classification tasks. I will examine the effect of various balancing strategies on the efficacy of various machine learning algorithms. In addition, we will investigate how data balancing impacts model calibration, thereby addressing some of the field's current limitations.

Who will benefit from your work? Why?

Both researchers and practitioners in the field of machine learning will benefit from this study's findings. Researchers will acquire an understanding of the advantages and disadvantages of various data balancing techniques, enabling them to make informed decisions regarding class imbalance in their work. Practitioners will gain a deeper comprehension of the impact of data balancing on model performance, allowing them to create more reliable and accurate machine learning models for practical applications.

What risks do you anticipate?

It can be difficult to select metrics that accurately reflect the performance enhancements resulting from data balancing. Another risk is the introduction of possible biases during data balancing, particularly when employing particular oversampling or undersampling techniques. It is essential to thoroughly address these risks and guarantee the dependability and generalizability of research results.

Out of pocket costs? Complete within 11 weeks?

I am confident that this project can be completed within a relatively short timeframe, like 11 weeks. The models I intend to train for experimental purposes will be traditional machine learning models. As a result, there will be no requirement for extra hardware support or extensive computing power. Additionally, obtaining labeled sample datasets for conducting the research will be a straightforward task and it will not be time consuming.

Midterm results?

By the midterm of this research project, I aim to:

1. Conduct a thorough literature review on data balancing techniques, imbalanced classification, and related evaluation metrics.
2. Collect and preprocess benchmark datasets with varying degrees of class imbalance.
3. Conduct at least half of the experiments that should be done:
    a. Implement and evaluate different data balancing methods, including oversampling, undersampling, and hybrid approaches, on the benchmark datasets.

b. Assess the impact of data balancing on model performance using evaluation metrics such as F1 score, AUC-ROC.
c. Analyze the effect of data balancing on model calibration.

Final Demonstration?

I aim to achieve the following final results:

1. Comparative Analysis of Data Balancing Techniques: Build upon the initial evaluation of data balancing methods conducted during the midterm stage and finalize. Identify trends and patterns in the performance of different data balancing techniques and algorithms.
2. Algorithm Specific investigation: Investigate the interplay between data balancing strategies and particular machine learning algorithms. Analyze how distinct algorithms react to data balancing techniques and identify algorithmic preferences or limitations in addressing class imbalance. This analysis will shed light on algorithmic performance in imbalanced scenarios and assist the selection of suitable algorithms for various data balancing strategies.
3. Practical Guidelines and Recommendations: Summarize the findings of the exhaustive analysis and provide practical recommendations and guidelines for implementing data balancing techniques in classification tasks involving an imbalance. Discuss the advantages, disadvantages, and best practices of various data balancing methods in light of the research findings. When addressing class imbalance in real-world applications, provide guidance on selecting appropriate evaluation metrics, algorithm selection, and potential hazards to avoid.