DSA5101 Python Machine Learning Project

Li Xiaoli

Problem Statement

- Welcome to our exciting group project!
- In this endeavor, we'll be delving into the fascinating realm of predictive analysis using the Bank Marketing dataset. Our goal is to harness the power of machine learning to predict whether a client will subscribe to a term deposit. The problem revolves around the direct marketing campaigns conducted by a Portuguese banking institution.
- These campaigns revolved around telephone conversations. Notably, multiple interactions with the same client were often necessary to determine whether the client would subscribe to the bank's term deposit product ('yes') or decline it ('no’).
- For our group project, **you will use Bank Marketing data to predict whether a client will subscribe a term deposit**.

Guidelines (1)

- To kick things off, please refer to the Readme.txt file, which contains valuable insights about the Data.
- **Utilization**: For our modeling efforts, we'll exclusively employ the 'traininingdata.txt' dataset. This data will serve as the foundation upon which we'll construct our machine learning models. Additionally, we'll use the 'testdata.txt' dataset to rigorously evaluate the performance of our models.
- **Machine Learning Exploration:** We're in for an exciting journey as we explore a minimum of **three diverse machine learning algorithms for classification**. We're not just limited to the algorithms covered in our classes; feel free to explore new techniques that could potentially provide valuable insights and better performance.
- **Oversampling Strategy**: Should the need arise to perform oversampling to address class imbalance, please be mindful to carry out this process solely on the training dataset. This approach prevents any over-optimistic results that could arise from oversampling the combined training and test data or oversampling the test data itself.

Guidelines (2)

- **Feature Engineering**: We encourage you to unleash your creativity when it comes to feature engineering. This includes the potential for feature generation and feature selection. Specifically, feature generation involves creating new features from the existing data to enhance the representation of the underlying patterns in the dataset. Feature selection is the process of identifying and retaining the most relevant features from the original set of features. By strategically engineering features, we could unlock previously hidden patterns within the data, potentially enhancing the performance of our models.
- **External data**: You can leverage external data if you believe they are useful.
- In a nutshell, our project is centered around making accurate predictions about term deposit subscriptions using the Bank Marketing dataset. By leveraging a variety of machine learning algorithms, responsibly handling oversampling, and applying feature engineering techniques, we're poised to gain deep insights into this intriguing problem space. Let's embark on this journey of discovery and innovation together!

Performance Evaluation (1)

- In the spirit of thorough evaluation, we will report a range of key performance metrics to comprehensively assess the models we'll be building. Specifically, we'll focus on **precision, recall**, and **F-measure** for the '**yes**' class. Moreover, we'll calculate the **macro-average F-measure** and **micro-average F-measure, Accuracy, MCC** to capture overall model performance across classes. Now, let's delve into the significance of these metrics:
- **Precision** for the 'yes' class : This metric illuminates the proportion of positive predictions that were truly correct.
- **Recall** for the 'yes' class : It is referred to as sensitivity or true positive rate, and emphasizes the proportion of actual positives that were successfully predicted by the model.
- **F-measure** for the 'yes' class: It is also known as the F1-score, a harmonic mean of precision and recall. It provides a balanced evaluation of a model's performance by considering both false positives and false negatives.

Performance Evaluation (2)

- **Macro-average F-measure**: This metric calculates the average F-measure across all classes, treating each class equally. It's valuable when we want to ensure that the model performs consistently well across all classes.
- **Micro-average F-measure**: This metric calculates the F-measure by considering the total true positives, false positives, and false negatives across all classes. It's useful when we want to assess the model's overall performance across the entire dataset.
- **Accuracy**: Accuracy measures the overall correctness of predictions by comparing the total number of correct predictions with the total number of predictions.
- **MCC**: It is a metric used to evaluate the performance of binary classification models. It takes into account true positives, true negatives, false positives, and false negatives to provide a balanced assessment of a model's classification accuracy.
- Feel free to expound upon which essential metrics bear greater significance in evaluating our models and hold practical value for our business objectives.

Macro-average F-measure

- It computes the F-measure for each individual class and then takes the average of these F-measures. It gives equal weight to each class, regardless of their sizes.
- To calculate the macro-average F-measure, you perform the following steps:
  1. Calculate the Precision and Recall for each class.
  2. Calculate the F-measure for each class using the formula mentioned above.
  3. Sum up the F-measures for all classes.
  4. Divide the sum by the number of classes to get the macro-average F-measure.
- Mathematically, it can be represented as:

  Macro-average F-measure = (*F-measure_1* + *F-measure_2* + ... + *F-measure_n*)/*n*, where *n* is the number of classes.

Micro-Average F-measure

- Micro-average F-measure considers the overall counts of true positives, false positives, and false negatives across all classes. It's a way to compute a single F-measure that takes into account the performance across all classes while considering the class imbalances.
- The formula for micro-average F-measure can be expressed as:
- Micro-average Precision (MAP) = Total_TP / (Total_TP + Total_FP)
- Micro-average Recall (MAR) = Total_TP / (Total_TP + Total_FN)
- Micro-average F-measure = 2 * (MAP * MAR) / (MAP + MAR)
- Where:
  - Total_TP: Sum of true positives across all classes
  - Total_FP: Sum of false positives across all classes
  - Total_FN: Sum of false negatives across all classes

Group, Submission, and Rating

- **Group Formation**: Each group comprises approximately four students, organized by you. Pls add your group information into https://docs.google.com/document/d/1vpyoGSi6Jsaj4MkRo7FJIsUt6UJHXDYSzaIhf_U_ZM0/edit
- **Submission**: One member from each group should be designated as the representative responsible for submitting the zipped files. These files should encompass:
  - 1. Presentation Slides: These slides should list all group members' names and matriculation numbers.
  - 2. Thoroughly Documented Code: The code should be meticulously documented, offering clarity and ease of understanding.
  - 3. Recorded Audio/Video Presentation: In this recording, any group member can present segments of the slides. The total presentation duration should approximately span **15 minutes**.
- Kindly ensure that the zipped files are uploaded to the designated directory: Files/Projects/Python ML Projects_Submission. The deadline is **October 1 at 11:59 pm**.
- **Rating based on Collaboration Principle**: Within a group, all students collaborate collectively and will receive an identical score, unless fellow group members express concerns about insufficient contributions from any individual.

Proposed Project Presentation Topics for Slides/Videos

- **Introduction and Problem Statement**: A concise depiction of the chosen dataset and problem statement.
- **Dataset Pre-processing**: A comprehensive exploration, visualization, and analysis of the data, along with techniques like feature engineering and selection.
- **Experimental Study and Analysis**: In-depth examination of experimental methods employed, accompanied by a thorough analysis of results and findings.
- **Project Accomplishments Overview**: An encapsulation of project achievements, encompassing significant insights drawn, such as feature importance assessment and the practical application of prediction outcomes for business.
- **Future Enhancement Possibilities**: A glimpse into the future, highlighting avenues for further enhancements and advancements.

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions