**Group Project Description**

**I. Introduction**

This project aims to encourage you to accomplish two data mining tasks which cover the essential steps of data mining workflow from data understanding, data preparation, modeling, evaluation and deployment. In particular, Task-1 requires students to have a thorough understanding of the data,

**II. Task-1 Bank Marketing with Structured Data**

1. **Task Description**

This task is an opportunity for student groups to apply data mining techniques to real-world business problems. It tries to simulate a scenario in which you are supposed to provide recommendations to the management team of a **commercial bank** on the problem of “**Bank marketing**” on a given product (term deposit).

You are provided a training dataset (“**bank\_marketing\_train.csv**”) that is about the direct marketing campaigns of a commercial banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Specifically, the training dataset is composed of 26,246 observations with 20 variables (19 input variables/features and 1 target variable ***y***) as below.

1. age (numeric)
2. job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3. marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5. default: has credit in default? (categorical: 'no','yes','unknown')
6. housing: has housing loan? (categorical: 'no','yes','unknown')
7. loan: has personal loan? (categorical: 'no','yes','unknown')
8. contact: contact communication type (categorical: 'cellular','telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
12. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
13. previous: number of contacts performed before this campaign and for this client (numeric)
14. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
15. emp.var.rate: employment variation rate - quarterly indicator (numeric)
16. cons.price.idx: consumer price index - monthly indicator (numeric)
17. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
18. euribor3m: euribor 3 month rate - daily indicator (numeric)
19. nr.employed: number of employees - quarterly indicator (numeric)
20. ***y*** - has the client subscribed a term deposit? (binary: 'yes','no')

Then you are required to build a (binary) classification model to predict the ranking score of being positive (*y*=’yes’) that achieve the best AUC score of the ROC curve on the training dataset, and then predict the **ranking score of** **being positive(y=’yes’)** for the 8,000 clients in the “**bank\_marketing\_test.csv”,** which has 19 input variables/features without the target variable *y***.**

The adopted model could be one of those models introduced in our course, including 1) Decision Tree, 2) Linear perceptron, 3) Logistic Regression, 4) SVM, 5) Naïve Bayes, 6) KNN, 7) Ensemble models of previous base models.

1. **Provided resources:**
2. " bank\_marketing\_train.csv" which is composed of 26,246 observations with 20 variables (19 input variables/features and 1 target variable ***y)***
3. " bank\_marketing\_test.csv " which is composed of 8,000 observations with 19 input variables/features without the target variable ***y***
4. “bank\_marketing\_test\_scores(example).csv ”provides an example to illustrate the expected content and format of the ranking scores for the 8000 observations (in the same order) by your developed model;

**III. Task-2 Data Mining Model Development & Deployment with Limited Resource**

**TBD**

**IV. Evaluation**

1. **Model performance: 60%**
   * 1. Task-1 Performance: AUC score of the ROC curve of the submitted ranking scores in Task-1 (35%):
     2. Task-2 Performance: (TBD, 25%).
2. **Methodology & Presentation: 20%**

Novelty, practicality, and rationality of proposed methodology in both tasks, respectively.

1. **Peer Evaluation: 20%**

**V. Deliverables and Due Dates**

1. **Delivered models and predicted results (Due: May 5, 2023 at 23:59pm)**
2. Task-1: "bank\_marketing\_test\_scores.csv" which stores the predicted ranking score for the 8,000 clients in the "bank\_marketing\_test.csv". This file is supposed to have 8,000 rows, and each row records corresponding ranking score of being positive(y=’yes’) for the corresponding client in the same order in “bank\_ marketing\_test.csv”. You may refer to the "bank\_marketing\_test\_scores(example).csv" for example.
3. Task-2：TBD

1. **Presentation (5-minute presentation + 3-min Q&A, May 6, 2023)**
2. The main idea of your proposed approach for both tasks and highlight the operations you take to enhance models’ performance under the constraints.
3. The estimated out-of-sample performance for your two models based on the training data.
4. **Slides and peer evaluation (Due: May 7, 2023 at 23:59pm)**
5. The slides used in the presentation but including one additional slide to report the assigned job duties of each member for this project as suggested in the following table.

|  |  |  |
| --- | --- | --- |
| Member Name | Student ID | Roles and Responsibility |
| A | 123 | e.g., Task-1 (feature engineering), Task -2 (data preprocess and model deployment) |
| B | 456 | e.g., Task -1 (data clean and process), Task -2 (model development & deployment) |

1. The peer evaluation form.

**V. Notes**

* **Only electronic submission** will be required.
* If required, you need to make sure that the submitted Python code or docker images is **runnable** and can generate and save the predicted labels that are **consistent with** the results in the files submitted by you for each model accordingly. Otherwise, the grading foe the project will be skipped and a 30% penalty on the grade if the team manage to resubmit the required resources within 24 hours upon receiving the reminder.
* The requirement and deliverables of the two tasks could be summarized as the table below.

|  |  |  |
| --- | --- | --- |
|  | Task-1 | Task-2 (TBD) |
| Provided data | Labeled structured data | Labeled unstructured data |
| Required data mining steps | From data understanding to model evaluation | From data understanding to model evaluation + model deployment |
| Constraints | The adopted model could be one of those models introduced in our course, including 1) Decision Tree, 2) Linear perceptron, 3) Logistic Regression, 4) SVM, 5) Naïve Bayes, 6) KNN, 7) Ensemble models of previous base models. | The maximum runtime memory should not exceed XX GB(TBD). |
| Evaluation  metric | AUC score of ROC curve | F1 score (TBD) |
| Deliverables | The results on the test dataset "bank\_marketing\_test\_scores.csv" | The executable model/docker (TBD) |
| Report | Estimated AUC score on out-of-sample data | Estimated F1 score (TBD) on out-of-sample data |
| Weight | 35% model performance + 10% methodology and presentation | 25% model performance + 10% methodology and presentation |