Group Project Description

I. Introduction

This project aims to facilitate student groups in completing two data mining tasks that encompass the fundamental steps of the data mining workflow: data understanding, data preparation, modeling, evaluation, and deployment. Task-1 requires a comprehensive understanding of the provided structured data, while Task-2 encourages the processing of unstructured data to create a deployable data mining pipeline.

II. Task-1 Bank Marketing with Structured Data

2.1 Task Description

This task offers student groups the opportunity to apply data mining techniques to real-world business problems. You are expected to provide recommendations to a commercial bank's management team regarding a marketing strategy for a specific product (term deposit).

You will work with a **training dataset ("bank_marketing_train.csv")** related to direct marketing campaigns conducted by a commercial banking institution, primarily via phone calls. Often, multiple contacts were required to determine whether a client would subscribe to the product (bank term deposit).

The training dataset consists of 26,246 observations and 25 variables (24 input features and 1 target variable y), detailed as follows:

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) ***feature_1***: feature with unknown meanings (numeric)
21) ***feature_2***: feature with unknown meanings (numeric)
22) ***feature_3***: feature with unknown meanings (categorical)
23) ***feature_4***: feature with unknown meanings (categorical)
24) ***feature_5***: feature with unknown meanings (categorical)
25) ***y***: has the client subscribed to 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 24 input variables/features without the target variable ***y*.**

Your task is to build a binary classification model to predict the likelihood (*ranking score*) of a positive subscription (*y*='yes') that achieves the best AUC score of the ROC curve on the training dataset. Then you will use the above model to predict the subscription likelihood (*ranking scores*) for 8,000 clients in “bank_marketing_test.csv”, which contains 24 input variables/features without the target variable ***y***.

The adopted model may include one of the following introduced in the course:

- Decision Tree
- Logistic Regression
- SVM
- Naïve Bayes
- KNN
  Ensemble models of previous base models including models implemented in XGBoost, CatBoost, and LightGBM.

Please note that NO external resources should be utilized.

2.2 Provided resources

1) **"bank_marketing_train.csv"** comprising 26,246 observations with 25 variables (24 input variables and 1 target variable y)
2) **"bank_marketing_test.csv"** comprising 8,000 observations with 24 input variables without the target variable y
3) **"bank_marketing_test_scores(example).csv"** which illustrates the expected content and format of the ranking scores for the 8,000 observations generated by your model.

III. Task-2 Text Classification with Model Deployment

3.1 Task Description

This project aims to encourage you to upcoming project aimed at developing two separate text mining pipelines for analyzing financial news as below.

- **Subtask 1 - Sentiment Analysis**

  The goal for this subtask is to build and train a pipeline that predicts the sentiment (positive or neutral) of given financial news. Additionally, the pipeline should return the corresponding probability of the predicted sentiment. Generally, positive sentiments indicate optimism regarding market or company performance, while negative sentiments reflect pessimism. Please utilize the training data provided in “training_news-sentiment.xlsx” (as introduced below) for this subtask..

- **Subtask 2 - Topic Classification**

  For this subtask, you will develop a pipeline to classify financial news into one of 18 predefined topic labels (as listed below). Additionally, the pipeline should return the corresponding probability of the predicted topic. The training data for this pipeline can be found in “training_news-topic.xlsx” as introduced below.

1	'上市保荐书',	10	'发行保荐书'
2	'保荐/核查意见'	11	'年度报告全文'
3	'公司章程'	12	'年度报告摘要'
4	'公司章程修订'	13	'独立董事候选人声明'
5	'关联交易'	14	'独立董事提名人声明'
6	'分配方案决议公告'	15	'独立董事述职报告'
7	'分配方案实施'	16	'股东大会决议公告'
8	'分配预案'	17	'诉讼仲裁'
9	'半年度报告全文'	18	'高管人员任职变动'

You are encouraged to explore various models and external resources, while also considering text preprocessing to enhance model performance.

Once you have constructed the optimal pipeline (including text preprocessing, feature extraction, and model selection), please ensure to complete the following deployment tasks:

1) Save and reload/test the pipelines.
2) Deploy the pipelines as API services and conduct local tests.
3) Build and test a Docker image with the necessary environment for deploying the API services.

4) Push the Docker image to your public Alibaba Cloud Container Registry (ACR) account.
5) [Optional] Pull and test the Docker image from your public ACR repository.

In detail, the Docker should provide two API services with the following specifications:

- **Docker Port**: 5724
- **Endpoints**:
  - **Sentiment Analysis**: */predict_sentiment*
    - Input: *news_text* (string)
    - Output: JSON string, e.g., *{"sentiment": "-1", "probability": "0.98"}*.
  - **Topic Classification**: */predict_topic*
    - Input: *news_text* (string)
    - Output: JSON string, e.g., *{"topic": "12", "probability": "0.63"}*.

Please adhere to the following constraints for your Docker image/container:

- The file size should not exceed 4 GB.
- The runtime memory must not surpass 900 MB.
- The container is restricted from accessing GPU and internet resources during runtime.

3.2 Provided resources

1) "**Subtask1-sentiment_analysis**" folder: "training_news-sentiment.xlsx" contains 4,200 "news_text" entries alongside their sentiment labels, with "1" indicating positive and "-1" indicating negative sentiment.
2) "**Subtask2-topic_classification**" folder: "training_news-topic.xlsx" contains 20,000 "news_text" entries with topic labels corresponding to the aforementioned table.
3) "**References**" folder: This includes recommended research papers for developing text classification and sentiment analysis models.

IV. Evaluation

1. Model Performance (80% of overall project grade)

1) **Task-1 Performance (30% of overall project grade)**
   The evaluation will be based on the AUC (Area Under the Curve) score of the ROC (Receiver Operating Characteristic) curve. This score will be calculated from the ranking scores submitted for records in "bank_marketing_test.csv" and compared against the true labels (held by the instructor).

2) **Task-2 Performance (50% of overall project grade)**
   The two subtasks, sentiment analysis and topic classification, will hold equal importance. The evaluation for each subtask will be performed using the provided API to predict labels on a holdout news dataset (held by the instructor). These predicted labels will then be compared to the true labels, and performance will be assessed based on the weighted F1-score. Additionally, the total time taken to predict

the labels for the holdout dataset will also be considered. The evaluation metrics and their corresponding weight in the overall project grade are summarized as below.

Subtask	Evaluation Metric	Weight of Overall Project Grade
Sentiment analysis	Total prediction time	5%
	Weighted F1-score	20%
Topic classification	Total prediction time	5%
	Weighted F1-score	20%

2. **Peer Evaluation & Job duty allocation: (20% of overall project grade)**

V. Deliverables and Due Dates

1. **Delivered models and results <u>for each group</u>**

1) **Task-1 (Due: December 2, 2025, 23:59):**
   a) **Ranking Scores File**: Please provide the file "bank_marketing_test_scores.csv," which should include the predicted ranking scores for the 8,000 clients in "bank_marketing_test.csv." Ensure the following:
      - The file must contain exactly 8,000 rows.
      - Each row should correspond to the predicted ranking score that indicates the likelihood of being positive (y='yes') for each respective client.
      - Refer to the provided "bank_marketing_test_scores(example).csv" for formatting reference.
   b) **Mean AUC Score:** You are required to submit the Mean AUC score of the ROC (Receiver Operating Characteristic) curve for your selected (best) model. This score should be reported using 5-fold cross-validation on the training data.

2) **Task-2 (Due: December 14, 2025, 23:59)**
   a) **Docker Image URL**: Submit the URL of the required Docker image hosted in a PUBLIC repository of your own Alibaba Cloud Container Registry (ACR) account.
   b) **Mean Weighted F1-score**: Provide the mean weighted F1-score reported by 5-fold cross-validation on the training data.

2. **Allocation of job duties <u>for each group</u> (Due: December 15, 2025, 23:59)**
   An Excel file summarizing the assigned job duties of each member for this project, structured as in the provided "Job allocation form-template.xlsx".

3. **Peer evaluation for <u>each student</u> (Due: December 15, 2025, 23:59)**
   Online questionnaire for peer evaluation.

VI. Notes

1) Only electronic submissions are required.
2) If needed, you must submit the Python code within 24 hours, ensuring it is runnable and generates predicted results consistent with the deliverables submitted for each model. Failure to do so will incur a 50% penalty on the grade.
3) The requirements and deliverables for the two tasks can be summarized in the table below.

	Task-1	**Task-2**
Provided data	• Labeled structured data for model training. • Unlabeled data for model evaluation	Labeled unstructured text for two separate subtasks as below: 1) Sentiment analysis. 2) Topic classification.
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) Logistic Regression, 4) SVM, 5) Naïve Bayes, 6) KNN, 7) <u>Ensemble models of previous base models including models implemented in XGBoost, CatBoost, and LightGBM</u>.	Any model is acceptable, but the maximum runtime memory of container should not exceed 900 MB (without accessing GPU or Internet). The image size should not exceed 4 GB.
Deliverables	1) The results on the test dataset which should be named as "bank_marketing_test_scores.csv". 2) The 5-fold CV performance on the provided training data	1) The URL of your docker image in ACR 2) The 5-fold CV weighted F1-score on the training data
Use external resources?	Not allowed	Allowed
Evaluation Metrics & Weight	AUC score of ROC curve :30%	1) Sentiment analysis: 25% • Weighted F1-score: 20% • Prediction time: (5%) 2) Topic classification: 25% • Weighted F1-score: 20% • Prediction time: (5%)