ADS 542 Statistical Learning | Classification

Author: Dr. Hakan Emekci

Created: 25 March 2025

Version: v23.01.07

Final Project Assignment

Objective

The objective of this project is to build a machine learning model to predict an outcome based on a selected dataset. Students can choose between the following two dataset options:

1. **Bank Marketing Data Set**: Predict whether a client of a bank will subscribe to a term deposit or not. The dataset can be found at https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. Use the **"bank-additional.csv"** with 10% of the examples (4119), randomly selected from the full dataset, and 20 inputs.
    - The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).
2. **Freely Chosen Dataset**: Students can select their own dataset for classification/regression problems. The dataset must be publicly available, contain at least 5000 instances, and include both numerical and categorical features. The chosen dataset must be approved by the instructor before proceeding with the project.

Requirements

Your project should meet the following requirements:

1. **Data cleaning**: Perform necessary data cleaning operations to make sure the data is in a suitable format for analysis.
2. **Data preprocessing**: Perform necessary data preprocessing operations such as feature scaling, encoding categorical variables, etc.
3. **Feature selection**: Use feature selection techniques to select the most relevant features for the model.
4. **Model selection**: Compare the performance of at least three different models (e.g., logistic regression, random forest, neural network) and choose the best one based on evaluation metrics.
5. **Hyperparameter tuning**: Tune the hyperparameters of the selected model to improve its performance.
6. **Evaluation**: Evaluate the performance of the final model using appropriate evaluation metrics.

7. **Deployment**: Deploy the final model using Streamlit and create a web interface for the model.

Grading

The project will be graded based on the following criteria:

- **Data Cleaning (10%)**: The dataset should be thoroughly cleaned, and any data quality issues should be addressed appropriately.
- **Data Preprocessing (10%)**: The categorical variables should be appropriately encoded, and numerical variables should be scaled if necessary.
- **Feature Engineering (10%)**: New features should be created where appropriate.
- **Model Selection (10%)**: Several models should be trained, and the best-performing model should be selected based on appropriate metrics. Students should evaluate the performance of their model using appropriate metrics and compare it with other models. The selected model's hyperparameters should be tuned using appropriate techniques.
- **Creating Pipeline (15%)**: Students should create a pipeline covering all processes.
- **Presentation (15%)**: A well-structured presentation is needed.
- **Deployment (25%)**: The selected model should be deployed using the Streamlit framework, and the deployed model should be usable by end-users.

Submissions

Submit a Jupyter Notebook containing the code for the project. Make sure to include sufficient documentation and comments in your code. Also, provide a separate document with instructions on how to run and interact with the deployed model in your report file. Each student should submit three files zipped in a single file as "Group_0XX.zip" on the LMS system:

- **Jupyter Notebook** (`project.ipynb`)
- **PowerPoint Presentation** (`presentation.ppt`) (Max 5 slides)
- **A report** (`report.pdf`) summarizing your findings and recommendations. (Max 2 pages). Give the Streamlit cloud address of your project.

The deadline for submission is **12 May 2025 at 11:59 PM**.

Instructions

- The project is open-book and open-internet. You are free to use any resources available to you, but you are not allowed to collaborate with other groups or to copy from other sources.
- This is an **individual project**.
- Each student has to implement the required steps and come up with an optimal solution.

- You are allowed to use any Python libraries for data analysis and modeling.
- You have to provide brief explanations of each step in the notebook and presentation.
- The notebook should be well documented and easy to follow.
- Your performance will be evaluated based on the grading criteria mentioned above.
- Academic honesty is expected from each student. Any act of plagiarism or cheating will not be tolerated and will be reported to the university.
- All submissions must be made on or before the specified deadline. Late submissions will not be accepted.
- Each group has to prepare a short presentation for their project. Selected projects will be discussed in the class as a demo.

Academic Honesty

- This is a group project, and each group is expected to work independently, with each member contributing equally to the project.
- Collaboration between groups is not allowed, and any instance of academic dishonesty will be reported to the university authorities.
- You may use online resources and libraries, but you must cite them properly in your code and presentation.
- Your code and presentation must be original and free of plagiarism.

Deadline

The deadline for submission is **12 May 2025 at 11:59 PM**. **Late submissions will not be accepted.**

Resources

You may find the following resources helpful:

- **Our Lecture Notes and Notebooks on Teams**
- **Pandas documentation**
- **Scikit-learn documentation**
- **Seaborn documentation**
- **Matplotlib documentation**
- **Streamlit documentation**
- **Hyperparameter tuning with GridSearchCV**
- **Understanding Precision-Recall in Scikit-Learn**
- **Data Cleaning and Preprocessing in Python**
- **Machine Learning Pipeline: What is it and how to build one**
- **How to Deploy Machine Learning Models with Streamlit**
- **Machine Learning Project Checklist**