CSC713M / CSC 713D
Midterm Exam
1T2425

Problem #1 (60%)

Deliverables:

- Documentation
- Python / Jupyter Code

You are tasked to join the Loan Approval Prediction Kaggle competition. The details are found in: https://www.kaggle.com/competitions/playground-series-s4e10/data. Note you do not have to have a high score in this competition but a thoroughness of the techniques discussed in the class is expected.

1. You can use any of the following algorithms:
    - Logistic Regression
    - Linear Regression
    - Decision Tree
    - K-Nearest Neighbors
    - SVM
    - Multilayer Perceptron
2. If needed, perform preprocessing techniques to transform the data to the appropriate representation. This may include binning, log transformations, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering, among others. There should be a correct and proper justification for the use of each preprocessing technique used in the project.
3. The project should train and evaluate at least 3 different kinds of machine learning models. The models should not be multiple variations of the same model, e.g., three KNN models with different number of K.
4. Make sure that the data is clean, especially features that are used in the project. This may include checking for misrepresentations, checking the data type, dealing with missing data, dealing with duplicate data, and dealing with outliers, among others. There should be a correct and proper justification for the application (or non-application) of each data cleaning method used in the project. Clean only the variables utilized in the study.
5. Perform hyperparameter tuning to achieve the best model. Make sure to elaborately explain the method of hyperparameter tuning. Explicitly mention the different hyperparameters and their range of values. Show the corresponding performance of each configuration.
6. There should be a correct and proper justification in using the machine learning algorithms. You can combine multiple algorithms and techniques to achieve a higher accuracy.
7. Discuss each algorithm and the best set of values for its hyperparameters. Identify the best model configuration and discuss its advantage over other configurations. Discuss how tuning each model helped in reducing its error in difficult classes and/or instances.
8. Use of GenAI is permitted in this but an explanation on how it is used must be included.
9. The best model in the class (compared via AUC) will get 10% bonus points.
10. Discussion about the concepts discussed in class between peers is allowed but any sharing of code, answers or numerical results is prohibited.

Problem #2 (40%)

Deliverable: Summary

You are tasked to read and summarize the paper "Generalization through Memorization: Nearest Neighbor Language Models" by Stanford and Facebook Research.

- The summary must not exceed 500 words.
- **You cannot use and GenAI in this problem. Use of GenAI will be considered cheating.**
- The summary must include the motivation, purpose, methodology, results, analysis and other relevant concepts. A relation to the concepts discussed in the class is also expected.

For tips on making a summary of a research article you can refer to the following references.

- https://writingcenter.unc.edu/tips-and-tools/summary-using-it-wisely/
- https://prowritingaid.com/write-summary#head3
- https://www.grammarly.com/blog/academic-writing/how-to-summarize-a-research-paper/