EEE 485/585 PROJECT PROPOSAL

Group Member Name/Surname: Alp Dursunoğlu **ID:** 22102196

Group Member Name/Surname: Muhammet Melih Çelik **ID:** 22003836

Main Project Task: Credit Card Default Prediction

Introduction

In recent years, according to the financial personal data, there is a clear increase in the credit card defaults. Hence, banks aim to minimize the capital loss by deploying machine learning algorithms to label potential customers that can have credit card default based on several different features of the customers. For this specific problem, a machine learning prediction system will be developed by using three different statistical learning algorithms for the project. In the project, the training and testing will be conducted with the data obtained from:

https://www.kaggle.com/code/gpreda/default-of-credit-card-clients-predictive-models/input

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset contains 23 distinct features related to customers. The features are given below:

1. **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
2. **GENDER**: (1=male, 2=female)
3. **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
4. **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
5. **AGE**: Age in years
6. **PAY_0**: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
7. **PAY_2**: Repayment status in August, 2005 (scale same as PAY_0)
8. **PAY_3**: Repayment status in July, 2005 (scale same as PAY_0)
9. **PAY_4**: Repayment status in June, 2005 (scale same as PAY_0)
10. **PAY_5**: Repayment status in May, 2005 (scale same as PAY_0)
11. **PAY_6**: Repayment status in April, 2005 (scale same as PAY_0)
12. **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)
13. **BILL_AMT2**: Amount of bill statement in August, 2005 (NT dollar)
14. **BILL_AMT3**: Amount of bill statement in July, 2005 (NT dollar)
15. **BILL_AMT4**: Amount of bill statement in June, 2005 (NT dollar)
16. **BILL_AMT5**: Amount of bill statement in May, 2005 (NT dollar)
17. **BILL_AMT6**: Amount of bill statement in April, 2005 (NT dollar)
18. **PAY_AMT1**: Amount of previous payment in September, 2005 (NT dollar)
19. **PAY_AMT2**: Amount of previous payment in August, 2005 (NT dollar)
20. **PAY_AMT3**: Amount of previous payment in July, 2005 (NT dollar)
21. **PAY_AMT4**: Amount of previous payment in June, 2005 (NT dollar)
22. **PAY_AMT5**: Amount of previous payment in May, 2005 (NT dollar)
23. **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)

Algorithm Choices

For the project, the following three statistical learning algorithms are decided to be used.

1. Logistic Regression
2. Shallow Neural Network
3. Support Vector Machine (SVM)

Python 3 programming language will be used throughout the project for the implementations of the algorithms on the dataset.

Challenges

There are some possible challenges specific to the selected algorithms. These challenges especially can be encountered during implementation.

1. Logistic Regression

1.1. Assumption Violation: The main assumption in this model is that presence of a linear relation between features and the output, because there can be nonlinearity in the dataset.

1.2. Overfitting or Underfitting: As there are many features in the selected dataset which may results in overfitting. Regularization methods (Ridge or Lasso) will be tried to be implemented to prevent the overfitting or underfitting.

1.3. Imbalanced Data: In this project, the number of customers who will be labeled due to the credit card default can be considered as a rare event generally. This results in imbalance in data.

2. Shallow Neural Network

2.1. Overfitting: Shallow neural networks are inclined for overfitting so using this method may require some regularization techniques.

2.2. Training Time: Neural networks will work slower compared to other methods of machine learning. Using a single-layered shallow neural network compute faster compared to multi layered networks, but it still can be slower compared to other methods, especially when working with large datasets.

2.3. Hyperparameter Tuning: Finding an optimal spots for hyperparameters such as learning rate, batch size, number of neurons can be challenging and may require a series of experiments.

3. Support Vector Machine (SVM)

3.1. Kernel Selection: To determine the decision boundaries between classes, suitable kernel function parameters are tried to be selected and tuned.

3.2. Scalability: SVMs can draw a lot of resources when dealing with large datasets.

3.3. Parameter Tuning: SVMs have hyperparameters such as the regularization parameter and kernel parameters which may be tuned to achieve optimum performance.