COMP4434 Big Data Analytics

Individual Project

Introduction

AlphaMoney is a platform for issuing loans. For people who want to borrow money from the AlphaMoney platform, the platform requires them to provide relevant personal information to decide the Loan Amount. The platform has collected personal information from users, including Gender, Age, Credit Score, Property Age, Profession, Loan Amount Request, Property Price and Income Stability, etc., and determined the Loan Amount based on them. Now, some new users apply for loan amounts and provide their personal information. In this project, you are required to help AlphaMoney to design a Loan Amount prediction model for the new users based on the historical data and the knowledge you learned in Big Data Analytics.

Dataset

This dataset is composed of two parts:

1. *train.csv* contains historical personal information of 27,000 users and their specific information **including** Loan Amount.
2. *test.csv* contains personal information of 3,000 users and their specific information **without** Loan Amount.

p.s. '-999' vague values are handled as outliers.

Task

Design prediction models (at least **Linear Regression** and **DNN**) to predict the **loan amount**. Other models are also encouraged.

Submission Format

1. Predict the Loan Amount of the 3000 users in test.csv. Save one file with the following format and name it as *your_student_ID.csv*. (e.g., *190XXXXXXD.csv*)

Customer_ID	Loan Amount
1	
2	
…	

2. The source code (name it as *your_student_ID.py*), readme file (name it as *your_student_ID.md*), report (name it as *your_student_ID.pdf*), and the above prediction result need to be submitted to Blackboard. (Please upload them separately.)

Final Report

A final report should include following information of this project, e.g.

- Introduction
- Data preprocessing/analytics
- Model design and implementation
- Performance evaluation and discussions
- Summary and future work
- Reference

Project Grading

The project is 25% of the total subject assessment.

- Final report (10%)
- Code (15%)
    1. Work independently
    2. Language is not limited, however, Python is recommended.
    3. Only using the Third-party related packages for the Linear Regression model will lose the point for coding.

Grading Criteria for Project Report

We will grade your report based on the following 3 aspects:

1. Integration of course content:
    - You are encouraged to apply the knowledge you learned in the course as much as possible.
2. Diversity of your methodology:
    - You are encouraged to use more than two models. Creative comparisons, critical thinking, and valuable discussions will be very impressive in your project report.
3. Performance of your models:
    - Your prediction model will be compared with the true value and the grading is based on the Root Mean Squared Error (RMSE). Lower the RMSE as much as possible.

Timeline:

Submission deadline:

- Final report: **23:59 15 Apr 2022**