CS150 Database and Data Mining
Course Project

Student 1
ID: xxxxxxxx
email1@shanghaitech.edu.cn

Student 2
ID: xxxxxxxx
email2@shanghaitech.edu.cn

Guideline

Compared with developing a novel machine learning algorihtm, building a machine learning system is less theoretical but more engineering, so it is important to get your hands dirty. To build an entire machine learning system, you have to go through some essential steps. We have listed 5 steps which we hope you to go through. Read the instructions of each section before you fill in. You are free to add more sections.
If you use PySpark to implement the algorithms and want to earn some additional points, you should also report your implementation briefly in the last section.

1 Explore the dataset

Instruction:
Explore the given dataset, report your findings about the dataset. You should not repeat the information provided in the 'Data Format' section of project.pdf. Instead, you can report the data type of each feature, the distribution of different values of some important features(maybe with visualization), is there any missing value, etc

Your work below:

2 Data cleaning

Instruction:
Some people treat data cleaning as a part of feature engineering, however we separate them here to make your work clearer. In this section, you should mainly deal with the missing values and the outliers. You can also do some data normalization here.

Your work below:

3 Feature engineering

Instruction:
In this section, you should select a subset of features and transform them into a data matrix which can be feed into the learning model you choose. Report your work with reasons.

Your work below:

4 Learning algorithm

Instruction:
In this section, you should describe the learning algorithm you choose and state the reasons why you choose it.

Your work below:

5 Hyperparameter selection and model performance

Instruction:
In this section, you should describe the way you choose the hyperparameters of your model, also compare the performance of the model with your chosen hyperparamters with models with sub-optimal hyperparameters

Your work below:

6 PySpark implementation (optional)