Individual Reflection Report

Tong Wu
22205061

1. Introduction

This homework focuses on training and evaluating prediction models for particular problems and datasets. The goal of this homework is to work with the data to build and evaluate prediction models that capture the relationship between the descriptive features and the target feature *death_yn*. The original datasets of this homework are based on the previous homework of all team members. Another member of the team got a better grade in the previous, we decide to use her approach which is used in the first homework to process my original dataset and take them as the input datasets in this homework.

2. Personal Contributions

In our team, I am responsible for the 2, 3, and 4 parts, implementing the three models including Linear Regression, Logistic Regression, and Random Forest. Another one is responsible for 1 and 5 parts, understanding and preparing data, and improving predictive models.

For my contributions, I split the datasets processed in the first part into two portions, 70% for training and 30% for testing. Firstly, I trained the data using Linear Regression. I used the trained models to predict the target feature according to the training data. Then using *metrics.classification_report* to show the detail of the performance of this model. To further validate the accuracy of the model, I again used the trained model to predict the target feature according to the testing data. For both tests, I obtained very similar results. It suggests that the trained model is generalizing well and is not overfitting. To avoid this trained model is likely to perform well on the same data it was trained on, but it may not perform well on new data. I also performed cross-validation on the trained model to double-check the performance of this model. After doing that, the validation results show that the model does not suffer from problems such as overfitting and shows similar performance and accuracy when predicting different datasets. I then used the same procedure as above to get the trained model using Logistic regression and did the cross-validation. No problem was found during the processing.

The process of training a model using the random forest model differs from the above two models. We trained the dataset using Randon Forest model and then drew two decision trees, with depths of 4 and 10 respectively. We can get the importance of different features from a decision tree. The accuracy of the trained model is higher than Linear and logistic Regression. We also performed cross-validation tests on this model and obtained the desired results. Follow-up on improving predictive models will be done by my teammates.

3. What did I learn from this project

I have learned a lot of knowledge and skills from this project. The first thing is the importance of teamwork and assigning tasks according to each person's strengths. For example, my teammate is more conscientious and meticulous, then she is better suited to the part of handling data and improving models. I was responsible for the machine learning part in the teamwork of the course Software Engineering I am suited to implement the three models. Although the process of training the model is important, but data preprocessing, including understanding and preparing the data is also important. As the quality of the data directly decides the accuracy and performance of the trained models. After training models, we should make a test for the model to ensure that it is accurate enough for different data. Besides, I know how to select features that will be used in the training process such as importance scores, feature interactions, etc. In addition, I believe that the evaluation of a model's suitability for a project should not focus solely on the accuracy of the model. For example, some projects need to be deployed in a short time, so training the model will not be allowed to take too much time, and then the accuracy of this model may be not too high. But I have to admit that a good model needs to be accurate enough.

It would be best to document each step of the project and keep track of the decisions made during the project. This documentation could help you avoid any errors or issues that might occur later on. On one occasion when I closed the file, I didn't pay attention to whether the file was saved successfully or not. The result was that all the code I had written that day was not saved because the file was not trusted in Jupyter Notebook and was not saved automatically. Hence, before closing the file, you should double-check whether the file is saved successfully. It is a good habit for my career. Another key point is essential to have good communication within the team to ensure that everyone is on the same page and that there are no misunderstandings. This will greatly improve efficiency.

4. Other

There is a question. In part 5, we select the model which was trained by Random Forest Regression. But when we tested the accuracy of this model using the varying number of features as input, we found that using all features has higher accuracy than using the features which were selected according to our analytics and their correlationships. Does this mean that all the analytical work we do is meaningless? It is confusing for me. The answer may be found later in the learning process. All in all, this project was very rewarding for me.