Distinction TASK

About this task

Step-1

This task is designed to assess the Distinction level expectations. There are two sets of evidence requested in this task: one is for SIT307 students and the other one for SIT720 students. Please select the set based on your unit code. DO NOT SUBMIT BOTH SETS OF EVIDENCE.

Step-2

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: Friday 12 May (No submission before this date means no feedback!)
Submission deadline: Before creating and submitting portfolio.

Background

The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories.

In this assignment, the task is to build models to predict employee productivity given some other factors in a factory. In order to maintain real-time capability, model sizes should be as small as possible. Note that the employee productivity estimation in production will be deployed on best-cost hardware of traction drives in an automotive environment, where lean computation and lightweight implementation is key.

Specifically, the problem you are going to solve is: Can you

- Accurately predict the actual productivity given the collected data?
- Well explain your prediction and the associated findings? For example, identify the key factors which are strongly associated with the response variable, i.e., actual productivity.

Data set

The data set contains 1197 instances, each of which have 15 columns: the first 14 columns corresponding to the attributes, the 15th column ``actual_productivity'' is the variable that we will predict. The details of the data set can be found and downloaded in the original UCI Repository.

Evidence of Learning – SIT307

Execute your code into a jupyter notebook (.ipynb file) and keep the output, write a report (.pdf file) to answer the following questions, and submit your code and report to OnTrack.

1. Load and explore dataset, do necessary pre-processing and split the dataset into training set and test set with an appropriate ratio. Explain the steps that you have taken (e.g. show dataset size, dealing with missing values, feature exploration and representation, label distribution, split dataset etc).

2. Based on the training data, create three supervised machine learning (ML) models for predicting actual_productivity.

    a. Report performance score using a suitable metric on the test data. Is it possible that the presented result is an underfitted or overfitted one? Justify.
    b. Justify different design decisions for each ML model used to answer this question.
    c. Have you optimised any hyper-parameters for each ML model? What are they? Why have you done that? Explain.
    d. Finally, make a recommendation based on the reported results and justify it.

3. Analyse the importance of the features for predicting actual_productivity using two different approaches. Give statistical reasons of your findings.

Evidence of Learning – SIT720

Execute your code into a jupyter notebook (.ipynb file) and keep the output, write a report (.pdf file) to answer the following questions, and submit your code and report to OnTrack.

1. Load and explore dataset, do necessary pre-processing and split the dataset into training set and test set with an appropriate ratio. Explain the steps that you have taken (e.g. show

dataset size, dealing with missing values, feature exploration and representation, label distribution, split dataset etc).

2. Based on the training data, create three supervised machine learning (ML) models except ensemble methods for predicting actual_productivity.

    a. Report performance score using a suitable metric on the test data. Is it possible that the presented result is an underfitted or overfitted one? Justify.
    b. Justify different design decisions for each ML model used to answer this question.
    c. Have you optimised any hyper-parameters for each ML model? What are they? Why have you done that? Explain.
    d. Finally, make a recommendation based on the reported results and justify it.

3. Analyse the importance of the features for predicting actual_productivity using two different approaches. Give statistical reasons of your findings.
4. Try an ensemble method to predict actual_productivity, do you get better model performance? Explain your findings.