Distinction Task (Supervised learning)

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 15 September (No submission before this date means no feedback!)

Submission deadline: Before creating and submitting portfolio.

Required documents

1. **Submit a report (pdf format) in Ontrack (https://ontrack.deakin.edu.au)**
2. **Complete the problem credit task and submit your code file (.ipynb) separately in the OnTrack (https://ontrack.deakin.edu.au).**

Background

The dataset provided consists of data collected from heavy Scania trucks. Each instance of the training set is classified as positive or negative. The positive class represents failures for the specific component considered, while negative instances represent trucks with failures for components not related to the APS. Based on this scenario, the goal is to develop a prediction model that minimises a total misclassification cost where there is a much greater cost for predicting false negatives (FNs) than false positives (FPs), namely 500 and 10 units, respectively.

Datasets Description

The attributes are as follows: class, then anonymised operational data. The operational data have an identifier and a bin id, like "Identifier_Bin". In total there are 171 attributes, of which 7 are histogram variabels. Missing values are denoted by "na".

Datasets: original UCI Repository.

Evidence of Learning- SIT307

1. What are the differences between hyperparameter and parameter of a machine learning (ML) model. Explain your answer using at least two machine learning models that you have learned in this unit.
2. Prove that Elastic net can be used as either LASSO or Ridge regulariser.
3. Analyse the importance of the features for predicting “class'' using two different approaches. Explain the similarity/difference between outcomes.
4. Create three supervised machine learning (ML) models except any ensemble approach for predicting “class''.
    a. Report performance score using a suitable metric. Is it possible that the presented result is an 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.
5. Build three ensemble models for predicting “class”.
    a. When do you want to use ensemble models over other ML models? Explain based on the models that you have used in Q4 and Q5.
    b. What are the similarities or differences among used ensemble models used in Q5?
    c. Is there any preferable scenario for using any specific model among the set of ensemble models?
    d. Write a report comparing performances of models built in question 4 and 5. Report the best method based on model complexity and performance.
    e. Is it possible to build ensemble model using ML classifiers other than decision tree? If yes, then explain with an example.

Evidence of Learning- SIT720

1. What is the mechanism used by the Voting Classifier to aggregate predictions from multiple base models and their differences? Explain with suitable example. If predictions conflict among the base models, what strategies can be employed to resolve the conflicts and make more reliable final predictions in the Voting Classifier?

2. Can Support Vector Machine (SVM) effectively perform multiclass classification? If so, explain the approach and techniques for successful implementation. If not, clarify the limitations or challenges that restrict SVM's suitability for multiclass classification tasks.

3. Analyse the importance of the features for predicting “class” using two different approaches. Explain the similarity/difference between outcomes.

4. Create three supervised machine learning (ML) models except any ensemble approach for predicting “class''.
    a. Report performance score using a suitable metric. Is it possible that the presented result is an 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.

5. Build three ensemble models **including voting classifier** for predicting “class”.
    a. When do you want to use ensemble models over other ML models? Explain based on the models that you have used in Q4 and Q5.
    b. What are the similarities or differences between among used ensemble models in Q5?
    c. Is there any preferable scenario for using any specific model among the set of ensemble models? Explain based on obtained outcomes.
    d. Write a report comparing performances of models built in question 4 and 5. Report the best method based on model complexity and performance.
    e. Is it possible to build ensemble model using ML classifiers other than decision tree? If yes, then explain with an example.