

Question: Delivery Time Estimation

Problem Statement

A product supplier company is facing a problem. Whenever a customer places an order for a product, the company has to estimate the number of days it will take for the product to be delivered to the customer, so that it can communicate the expected delivery date to the customer. However, thus far, many customers have complained that the expected delivery date communicated to them is not accurate. To resolve this issue, the company has decided to engage your Data Science expertise to predict the delivery time based on historical data on product orders.

Datasets

- data_train.xlsx: this dataset is used to train your model, which contains the order information. Specially, column “Delivery Time (days)” is the target that the company wants to predict.
- data_test.xlsx: this dataset is used to test your model performance.

Task

The company has tasked you to create an algorithm that uses Data Science modelling to accurately predict the delivery time. However, the company has also mentioned that besides accuracy, the model should ideally have the following features:

1. Try to avoid a prediction time which is less than the actual delivery time, since this case will necessitate a change in delivery date (i.e. a delayed delivery date) and result in customer dissatisfaction.
2. Once you have built a model, the algorithm should have the functionality to update this existing model with new datasets, since the delivery time pattern may change overtime. In particular, the existing model should be able to be updated with new datasets under two scenarios:
 - I. The training dataset or new dataset size is small, and your model training time is short.
 - II. The training dataset or new dataset size is quite large, and your model training time is quite long. Try to make it efficient for the updating process.

Taking the above requirements into consideration, build your model with the training dataset and test on the testing dataset. It is suggested that you do not spend all your efforts to get the best accuracy since accuracy is just one of the requirements.

Files to Submit

1. A coding file that clearly outlines your whole modeling and testing process, including intermediate output or charts as appropriate. We are flexible on the choice of software

you use to generate this coding file. For instance, you can choose to use an IDE such as Jupyter Notebook/Spyder/PyCharm/other IDEs. You may also choose to use a Word Processor program such as Microsoft Word.

2. An executable python file (.py) which can be used to run on the testing data file, the expected output is a file with predicted and actual delivery time for each order, and an evaluation accuracy result based on your chosen model evaluation metrics. The python file must be in Python version 3 and above.
3. An executable python file (.py) which can be used to update the existing model with new datasets under the two scenarios highlighted earlier. The python file must be in Python version 3 and above.
4. Other files you deem necessary for us to run your file.
5. (If any) Any other suggestions / further improvements to the algorithm / model on top of the requirements already specified.