CS 463 - Data Challenge Fall, 2023

Background

The goal of this data challenge is to demonstrate your understanding of general machine learning by predicting an identified outcome on a dataset. In answering the question:

1. You are allowed (and expected to) use external packages and libraries (eg. scikit-learn, matplotlib, pandas).
2. You are allowed to consult external sources (notes, etc.).
3. You are allowed to discuss your model's performance and general ideas.

… but you are not allowed to use external data sources, nor are you allowed to discuss specific solutions with others inside or outside the course.

You are required to document your findings and explain all choices you make. You may use a Jupyter notebook to capture the output at every step. Any plots (charts, graphs) you create in terms of documenting or explaining must be constructed using matplotlib or must have approval from the instructor. Work hard, and have fun!

Cargo Volume

This task requires that you create a model to predict cargo volume through San Francisco International Airport. The file contains a header row with column names. An example row of data — the first row — is as follows:

```
200507,ABX Air,GB,ABX Air,GB,Domestic,US,Deplaned,Cargo,Freighter,20.6038728
```

There are 11 columns in each part of this dataset. The columns of the dataset are described as shown in Table 1.

The data comes in two parts:

- A train set — a CSV file containing all the columns in Table 1.
- A test set — a CSV file containing all the columns in Table 1 EXCEPT for the target column (Cargo Metric Tons).

There are two tasks in this data challenge:

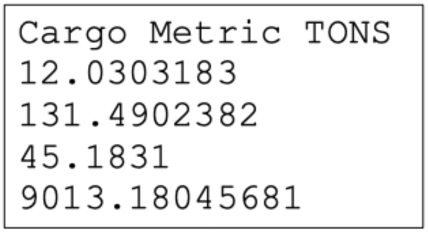
- Task 1: Choose and defend the best metric for quantifying performance of a model.
- Task 2: Produce the best model you can for predicting the target (Cargo Metric Tons) using the train set data.
- Task 3: Using the best model, generate predictions for the target on the test set.

Your solution will be judged based on several criteria, as described in the “Grading” section, including: the appropriateness of the metric you select, the degree to which you follow a good process and your model’s performance on the test set. You may change the existing data in any way you see fit as long as the data retrieval is part of your submission.

Order	**Column Name**	**Description**
1	Activity Period	Year and month described by this row (eg. 202012).
2	Operating Airline	Airline operating aircraft (eg. United Airlines).
3	Operating Airline IATA Code	IATA code for airline operating aircraft (eg. UA).
4	Published Airline	Company for which airline is being operated (eg. DHL Express). This may be the same or different to Operating Airline.
5	Published Airline IATA Code	IATA code for published airline.
6	GEO Summary	“International” or “Domestic” (with respect to the USA).
7	GEO Region	Continent or (North American) country to which the airline is bound or from which it comes.
8	Activity Type Code	“Deplaned” if arriving at SFO; “Enplaned” if leaving SFO.
9	Cargo Type Code	“Cargo”, “Express”, “Mail”, etc.
10	Cargo Aircraft Type	“Combi”, “Freighter”, “Passenger”, etc.
11	Cargo Metric Tons	Volume of cargo.

Table 1: Columns in the Cargo Volume dataset

Submission



```
Cargo Metric TONS
12.0303183
131.4902382
45.1831
9013.18045681
```

Figure 1: Example CSV output file

Submit the following items:

1. Your choice of best metric and the reason why you chose it.
2. Your implementation — eg. `dc1.ipynb` = Jupyter notebook you used to answer this data challenge question. Make it clear — either through code comments or an additional accompanying document — what procedure or procedures you follow and the reasons why. For

example, you may decide to use the Logistic Regression classifier based on the relationships between features and response. Your process for deciding on this classifier should be clear and the evidence for it should be part of your submission.

3. Your predictions for the target on the test set — eg. `dc1.csv` = a CSV file with your model's predictions for the target (Cargo Metric Tons) given the features in the test set. The file should have a header row ("Cargo Metric TONS") and one hypothesis for each row in the test set. An example of a 4-row test set is shown in Figure 1. (Your model will generate different data and more rows than appears in this figure.)

Grading

Each submission will be graded as follows:

40%	Performance	The competitive[1] performance of your model as executed on a neutral system.
20%	Execution Time	The competitive wall clock execution time of your model as executed on a neutral, CPU-based system
20%	Process / Documentation	The degree to which your solution follows a reasonable process and have documented this process: 20% = Completely 14% = Partially: missing process details / module documentation 06% = Poorly: missing several major details / most documentation 00% = Does not follow process / does not document
10%	Best Metric Choice	The appropriateness of your choice of metric and the reason for choosing it.
10%	Code Quality	The degree to which your solution is modular and easy to read. 10% = Completely 06% = Partially 02% = Poorly

[1] For competitive grading, the submissions with the top performance get a full-credit score (eg. 40/40 on Performance). Other submissions which do not yield top performance are ranked and graded accordingly. Your model will be executed once to ensure compliance with your output, so be wary of models with varying / random performance.