CS 663 - Data Challenge Spring, 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 and others).
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!

Cash or E-ZPass?

Accept a repo through Github Classroom through the link: https://classroom.github.com/a/VYf_zT9Q.

This task requires that you create a model to predict how a motorist will pay for tolls on the New York State (NYS) Thruway. It might be useful for you to read details about the NYS Thruway, which can be found on Wikipedia. The file contains a header row with column names. An example of the data — the first two rows — is as follows:

```
02/20/2018,46,49,1100,2L,2,CASH
01/09/2018,47,48A,1030,2L,1,E-ZPass
```

There are 7 columns in each part of this dataset. The columns of the dataset are described as shown in Table 1. The first row (above) indicates that between 11:00AM (and 11:15AM) on the 20th of February, 2018, two (2) class "2L" vehicles (cars, vans, SUVs and small trucks) entered the NYS Thruway at Exit 46, exited at Exit 49 and paid their tolls in cash.

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.

There are three 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 (Cash or E-ZPass) 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 degree to which you follow a good process and the 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	Date	The date on which the vehicle entered the NYS Thruway.
2	Entrance	The exit number where the vehicle entered the NYS Thruway.
3	Exit	The exit number where the vehicle departed the NYS Thruway and paid the toll.
4	Interval Beginning Time	The time at which the motorist entered the NYS Thruway.
5	Vehicle Class	The class (type, with respect to toll charge) of the vehicle. Vehicle classes are available at https://www.thruway.ny.gov/travelers/tolls/classes.html.
6	Vehicle Count	The number of vehicles reported corresponding to the Entrance, Exit, Interval Beginning Time and Vehicle Class.
7	Payment Type (Cash or E-ZPass)	The manner in which the toll was paid. “E-ZPass” is an electronic toll collection system (similar to FastPass, SunPass, etc.) used in New York state and other locations. E-ZPass information may be found at https://en.wikipedia.org/wiki/E-ZPass.

Table 1: Columns in the Cash or E-ZPass? dataset

Submission

CASH
E-ZPass
CASH
CASH

Figure 1: Example CSV output file

Submit the following items:

1. Your implementation — eg. `dc2.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.
2. Your predictions for the target on the test set — eg. `dc2.csv` = a CSV file with your model's predictions for the target (Cash or E-ZPass) given the features in the test set. The file should have a header row ("Cash or E-ZPass") 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.
30%	Process / Documentation	The degree to which your solution follows a reasonable process and have documented this process: 30% = Completely 21% = Partially: missing process details / module documentation 14% = Poorly: missing several major details / most documentation 00% = Does not follow process / does not document
20%	Execution Time	The competitive wall clock execution time of your model as executed on a neutral, CPU-based system
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 may be executed once to ensure compliance with your output, so be wary of models with varying / random performance.