

CSC 380 Summer 2023

Final Project

Due: August 7, Monday, 11:59 pm

Instructions

- Final is an extension of HW4. You will see HW4 questions and some new ones repeated here.
 - It will assess your overall understanding of the course, the fulfillment of course objectives, and the achievement of learning outcomes.
 - You will be working on an individual data science project from start to finish.
 - Remember, both the course instructor and TA Bennett are available to assist you with any challenges you encounter. Office hours are available by appointment too. If in doubt, ask.
 - The project does not have a predefined "solution"; you will be evaluated based on your approach and implementation of the data science project.
 - Please refer to the shared grading criteria and project structure below for evaluation details.
-

Grade Distribution

Submission	Total	% Contribution to Final Grade	Impact of 1 point of Submission in Final Grade
HW1	14	10.5	0.75
HW2	50	19.5	0.39
HW3	50	20	0.40
HW4	50	20	0.40
Final Project	50	20	0.40
Weekly Checkin	14	5	0.35
Participation Activities	5	5	0.20
Total	-	100	-
Bonus Questions	5	5	1

Table 1: Grade Distribution

1 Setup

Dataset Selection done in HW4. Github repo is the same as HW4.

2 [FINAL] Improvements [10 points]

Recommendations and feedback may have been received from the course staff for Homework 4 . Failure to implement the changes will result in a 10-point penalty apart from the regular penalty for the wrong/incorrect approach. If no feedback or recommendations are received, 10 points will be awarded automatically since you did it well the first time around.

If you have received recommendations and feedback, in a numbered list below, mention the feedback received and the actions taken. Make the actual changes in the appropriate sections of the notebook.

3 [HW4] Data Collection [5 points]

3.1 Dataset Description [5 points]

Write a short description of the dataset (200-word minimum). Below are some questions that can serve as a starting point. You may not have answers to all of these. These questions are meant to serve only as a guide.

- Where was the data collected from?
- How did they collect the data?
- Whose data is this?
- What are the main columns/features of the dataset? What does each of the column represent? You can add a table here if you want to.
- How old is it?
- Does this dataset get updated often? What is the update frequency?

3.2 Dataset Download [Incomplete Penalty -2 points]

- Download the dataset to the folder data.
- Add data to the git ignore file. This way, git will not add your dataset to your pushed repository. Since it is not our data to upload and distribute, we should not upload it to our repo.

4 [HW4] Data Preprocessing [20 points]

4.1 Data cleaning [5 points]

1. Perform some initial data exploration to understand the dataset better.
2. Address missing data, if any, by appropriately identifying and quantifying missing values in the dataset.
3. Apply suitable techniques to handle missing values, such as imputation or exclusion, based on a logical approach.
4. Justify the chosen method for dealing with missing data in at least 50 words.
5. If you can find no missing values, this section will still contain the code where you are checking for missing values.

4.2 Outlier Detection [5 points]

1. Check for outliers. You may need data visualization here. Go for a method you see fit.
2. If you do find outliers, apply effective outlier treatment techniques, such as removal, transformation, or capping, depending on the data and context.
3. Explained the rationale behind the chosen outlier handling approach.
4. If you can find no outliers, this section will still contain the code where you are checking for outliers.

4.3 Data Quality and Consistency [5 points]

1. Ensured data quality by verifying data consistency and accuracy.
2. Check for any duplicate records or potential errors and correct them as needed.
3. If you did perform some transformation on the dataset, in at least 25 words, explain what and why?

4.4 Data Preparation for Model training [5 points]

1. Convert the dataset into a form that a model can ingest.
2. This may include shuffling the dataset and splitting it into train, test, and validation.
3. Label encoding may be needed.
4. You may also have to normalize data.
5. Whatever transformation you performed to the data, explain why you did so in at least 25 words.

5 [HW4] Data Visualisation [8 points]

Choose any two charts/plots etc to visualize the dataset and its properties. For each viz,

- Ensure your plot is clear and has all necessary labels marked.
- Write a short desc of why you choose the viz and what you understood about the dataset from it.

6 [FINAL] Evaluation Metrics [5 points]

1. What metric(s) are you using to evaluate the model?
2. Write a short description of each metric, including the definition and formula used. Pros and Cons of picking the metric.

7 [HW4] Model (One of Two) Training [15 points]

7.1 Picking a model [5 points]

1. Pick any machine learning algorithm that you think works for your problem and data. It needn't necessarily be an algorithm taught in class.
2. Write a short paragraph explaining how the algorithm/model works. I am not looking for a full stats-oriented explanation (even though you are welcome to do so). Your explanation should help your classmates understand how the model basically works.

7.2 Model Training [10 points]

1. Train your model. Use any of the cross-validation methods available to pick your best model.
2. Choose appropriate tasks that go along with your model. Examples below.
 - If you are using a neural network model, show your training and validation loss plotted as a graph.
 - If you are using K-means clustering, plot the elbow curve. Choose an appropriate K.
3. List the hyperparameters of the model.

8 [FINALS] Model (Two of Two) Training [15 points]

8.1 Picking another model [5 points]

1. Pick another machine learning algorithm that you think works for your problem and data. It needn't necessarily be an algorithm taught in class.

2. Write a short paragraph explaining how the algorithm/model works. I am not looking for a full stats-oriented explanation (even though you are welcome to do so). Your explanation should help your classmates understand how the model basically works.

8.2 Model Training [10 points]

1. Train your model. Use any of the cross-validation methods available to pick your best model.
2. Choose appropriate tasks that go along with your model. Examples below.
 - If you are using a neural network model, show your training and validation loss plotted as a graph.
 - If you are using K-means clustering, plot the elbow curve. Choose an appropriate K.
3. List the hyperparameters of the model.

9 [FINAL] Report Results [4 points]

1. Test both of your models and report the performance. Which is a better model?
2. Write a summary to your project aka summarise what you did in the project.

Ex: In this project, I aimed to create a model that can predict a song's popularity, given its properties. For this purpose, I choose to work with The 1M Spotify Songs Dataset. The features available were While exploring the dataset, I observed the following... I choose A and B algorithms to train the models. The best model in each was picked using k-cross validation with k=3. When tested, model A gave better results at 89% acc. This is 5% more than Model B. In conclusion, we were able to create an effective predictor for songs popularity.

10 [HW4] Notebook Presentation [2 points]

While a clean with all outputs displayed notebook gets a credit of 2 points, unorganized and difficult-to-read notebooks get a penalty of 5 points.

11 [FINAL] README to the repo [8 points]

Update your project repo's README file. Here is the template. The bullets within each heading are meant to be guide.

12 [FINAL] Ethics Question [8 points]

We had a reading and/or activity around an ethics topic every week. You were given a question prompt to answer. Choose any week and answer the question from that week in a more detailed manner here, about 300 words or more.