UNIT 2 ASSIGNMENT

Managing Your Data in ML

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought responses.

Except as indicated, use this document to record all your project work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator as part of your project completion. You may also have additional supporting documents that you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

Complete each project part as you progress through the course. Wait to submit the project until all parts are complete. Begin your course project by completing Part One below. A submit button can be found on the final Course Project assignment page. Information about the grading rubric is available on any of the course project assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the project.

Week 2 Written Portion

Building a Modeling Dataset

Answer the questions below about building a model dataset and understanding your data through analysis and visualization.

## Questions:

1. What does it mean to have a “modeling dataset”?

|  |
| --- |
| Having a “modeling dataset” means having a dataset that is cleaned (i.e. inputs are sanitized and null values are dealt with). A modeling dataset is ready to be run by machine learning algorithms. |

1. What steps would you take with a raw dataset to end up with a modeling dataset?

|  |
| --- |
| I would first analyze the raw dataset. Then I would sanitize the inputs by, for example, making sure they’re all of the same datatype or making sure they don’t contain illegal or extraneous characters. Next, I would deal with null values, if any. After that, I would visualize the data to see if there are any general trends and see if there are any other transformations I can apply to obtain trends in data (e.g. apply logarithmic function). Finally, I will change all features to have numerical values (e.g. one-hot-encode, convert Booleans to 0/1). |

1. What is the difference between nominal data and ordinal data? Explain with an example.

|  |
| --- |
| Ordinal data has an order whereas nominal data does not. For example, the rank of software engineers can be considered ordinal data, ordered as follows: software engineer, senior software engineer, staff software engineer, principal engineer, fellow. An example of nominal data could be different colors. |

1. Why is data visualization an important part of the data preparation process?

|  |
| --- |
| Data visualization is an important part of the data preparation process as it helps us understand the data better and identify general trends. For example, plotting the data may help us identify that a certain set of datapoints follows a general quadratic curve. |

1. What is an outlier?

|  |
| --- |
| An outlier is a datapoint that is very different from all the other ones (could be caused by human error, machine error, or situational anomalies). |

1. Name a few libraries used for data analysis and visualization and explain when you would use each library.

|  |
| --- |
| Pandas for reading in data and applying transformations to the Dataframes.  Numpy for applying functions to Dataframes.  Scipy.stats for winsorizing Dataframes.  Matplotlib and seaborn for visualizing data. Seaborn makes it look pretty. |

*To submit this assignment, please refer to the instructions in the course*.