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”?

|  |
| --- |
| Modeling dataset means to analyze and transform data to a model/graph for visualization, better understanding of the data, and to tell a story. |

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

|  |
| --- |
| The steps I would take with a raw dataset to end up with a modeling dataset is to first ask ourselves what are we aiming to model. So, to define my unit analysis in my problem statement. Then I want to understand my data. Perform EDA, ask experts or those around me for information, and to build visualizations all to gain any insights. Then do data preparation. I would have sample population to aim towards the goal of generalization. Inspect my data, and determine if there’s any data cleaning I would need to perform. This involves looking for outliers, missing data, etc. Specify my label, have a data matrix, and analyze the data for any more data cleaning. |

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

|  |
| --- |
| The difference between nominal data and ordinal data is that nominal data is often represented in strings. Because it’s in the form of strings and most machine learning algorithms cannot operate on strings, we have to convert it to numeric, and we can use one hot coding for that. Another difference is that there’s no ordering for nominal data. An example of nominal data could be a survey where they asked “How was your experience?”, and participants answered. Meanwhile, ordinal data typically uses numbers as datatypes; and there is a natural order to them. An example of ordinal data is a survey conducted where participants answer a series of question using a rating scale of 1-5. |

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

|  |
| --- |
| Data visualization is an important part of the data preparation process because it helps us better understand our data. It also helps us to determine the skew of the data, shape of its distribution, and if there’s any outliers—which are difficult to determine if you’re just analyzing data on its data table. |

1. What is an outlier?

|  |
| --- |
| An outlier is defined as data that is far apart from other datapoints. They can be both univariant and multivariant in nature, and determining the outlier thresholds can be subjective. They’re typically problematic because it can skew the mean values towards the outlier which increases the variance which as a result increases the errors. |

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

|  |
| --- |
| For data analysis I would use NumPy and Pandas for data analysis. Pandas allows us to use several methods that allows us to create DataFrames, analyze the data, and manipulate the DataFrame, etc. (.shape, .head). As for visualization, I would also use Pandas, Matplotlib, and Seaborn. Matplotlib includes several plotting packages, Seaborn helps elevate the power of Matplotlib and it supports the visualization of our data and their statistical properties. Pandas also helps with plotting. |

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