DSC550-T301 Data Mining (2235-1)   Week 8: Term Project Milestone 2

Week 8: Term Project Milestone 2

Introduction

Contents of the Week

- 📋 Introduction
- 📑 Readings
- 🔖 Supplemental Materials
- 💬 8.1 Discussion/Participation
- ⭐ 8.2 Term Project: Term Project Milestone 2: Data Preparation

Topics

- **A** Data preparation/cleaning
- **B** Data selection/extraction
- **C** Feature engineering

Readings

📑 Read the following:
  - There are no textbook readings this week.

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings 📖	Videos 🎞

🌐 [What Is Data Preparation in a Machine Learning Project](#)

🌐 [Six Steps to Master Machine Learning with Data Preparation](#)

🌐 [Discover Feature Engineering, How to Engineer Features and How to Get Good at It](#)

8.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. Why is it so important to clean your data before building a model?

2. What are some of the key data cleaning steps?

3. How do you know when your data is ready for modeling?

4. What is feature engineering?

8.2 Term Project: Term Project Milestone 2: Data Preparation 👆

Now that you have created your idea, located data, and have started your graphical analysis, you will move on to the data preparation process of your project. After completing Milestone 2, your data should be ready for the

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

model building/evaluation phase.

Here is a list of steps to consider performing in Milestone 2:

👉 Drop any features that are not useful for your model building and explain why they are not useful.

👉 Perform any data extraction/selection steps.

👉 Transform features if necessary.

👉 Engineer new useful features.

👉 Deal with missing data (do not just drop rows or columns without justifying this).

👉 Create dummy variables if necessary.

Explain your process at each step. You can use any methods/tools you think are most appropriate. Do what makes the most sense for your data/problem. This will vary greatly among different projects. Be careful to avoid data snooping in these steps.

It is important to note that these milestones are meant to keep you on track for the final project submission. At any point, you can pivot or modify your project as needed based on what you discover. These milestones are not final versions; they are drafts of the many steps you need to complete along the way.

As a reminder, Teams is a great place to discuss your project with your peers. Feel free to solicit feedback/input (without creating a group project!) and collaborate on your projects with your peers.

Each milestone will build on top of each other, so make sure you do not fall behind. Submit Milestones 1 & 2 together. I recommend building your project milestones in a Jupyter Notebook, building upon one another. However, make sure it is clear where Milestone

Submit your code, output, and analysis at the link above. Comment all your code and address all items listed in the instructions. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Term Project Rubric

1 ends and Milestone 2 begins.