LIC LaGuardia

Data Analytics Academy

NYC HEALTH+ HOSPITALS

The Data Analytics Capstone Project Guidelines

All enrolled students are required to complete unit and final capstone projects. Students can use work data, subjected to approval from department/ organization, and use it in their class unit and final capstone projects. A student can deliver project as an individual or can team up with classmates also with max team size 3.

The below guidelines and checklist can be used as a guide during the process of any unit capstone data analytics project.

1. Answering the Question (Track 1 and Track 2)

a) Did you specify the type of data analytic question before touching the data?
b) Did you define the metric for success before beginning?
c) Did you understand the context for the question and the business application?
d) Did you consider whether the question could be answered with the available data?

2. Checking the Data (Track 1 and Track 2)

a) Did you identify any missing values in the data such as null, blank etc.?
b) Each variable is one column?
c) Each observation is one row/ record?
d) Do different data types appear in each column or table?
e) Is there any relationship between the data columns or data tables?
f) Did you check the units of all columns and rows/ records to make sure they are in the right range?

g) Did you try to identify any errors or miscoding of variables in your data table?

3. Exploratory Data Analysis (EDA) (Track 1 and Track 2)

a) Did you identify missing values?
b) Did you make univariate plots (histograms, density plots, boxplots)?
c) Did you consider correlations between variables (scatterplots)?
d) Did you check the units of all data points to make sure they are in the right range?
e) Did you try to identify any errors or miscoding of variables?
f) Did you consider plotting on a log scale?
g) Would a scatterplot be more informative?

4. Inference (Track 2)

a) Did you identify what large population you are trying to describe?
b) Did you clearly identify the quantities of interest in your model?
c) Did you consider potential confounders?
d) Did you identify and model potential sources of correlation such as measurements over time or space?
e) Did you calculate a measure of uncertainty for each estimate on the scientific scale?

5. Prediction (Track 2)

a) Did you identify in advance your error measure?
b) Did you immediately split your data into training and validation?
c) Did you use cross validation, resampling, or bootstrapping only on the training data?
d) Did you create features using only the training data?
e) Did you estimate parameters only on the training data?
f) Did you fix all features, parameters, and models before applying to the validation data?
g) Did you apply only one final model to the validation data and report the error rate?

6. **Reproducibility (Track 1 and Track 2)**
   a) Did you ensure that you are not performing any calculations manually?
   b) Did you create a script that reproduces all your analyses?
   c) Did you save the raw and processed versions of your data?
   d) Did you record all versions of your code that you used to process the data?
   e) Did you try to have someone else run your analysis code to confirm they got the same answers?

7. **Code Packages (Track 1 and Track 2)**
   a) Did you make your project package? For example: project folder, SQL or script files.
   b) Did you create any ER (entity relationship) diagram?
   c) Did you create your jupyter notebook files?
   d) Did you create your Tableau files and dashboard artifcats?
   e) Did you create any architectural artifacts?
   f) Have you eliminated all errors and warnings?

8. **Presentation (Track 1 and Track 2)**
   a) Did you provide background of data?
   b) Did you mention how you acquired the data?
   c) Did you specify the type of data analytic question you are answering?
   d) Did you try to anonymize the data to ensure that records are not personally identifiable if that was necessary?