Data Scientist Case Study - Directions

Our data analysts have found that a portion of patients who schedule an initial appointment don't actually show up to that appointment. These patients either reschedule their appointment, cancel their appointment, or are unable to show up to their appointment without providing prior notice.

It is important to the organization to predict – at the time of a patient schedules an appointment – which patients will show up to their appointments and which patients will not.

As a data scientist, you have been tasked with developing a machine learning model that **predicts whether or not a patient will attend an appointment**. Your model will run right after the patient schedules their appointment (usually online). Your model's predictions will be used to determine if more appointment slots need to be opened up on the organization's schedule. This will help keep the organization's schedule full, which ultimately optimizes business performance and increases the number of patient's lives improved.

Our data engineers have put together a dataset for you to use in training your machine learning model. The data engineers have pulled raw data, so any data cleaning or feature engineering is up to you. This data can be found in the "data" tab of the accompanying workbook.

Each data field is described in the "data_dictionary" tab of the accompanying workbook. You should use your knowledge of the business' scheduling process (found in the "Patient Phases" tab of the accompanying workbook) to determine which features can be used in the model.

From there, it is up to you to use your Python abilities and your data science training to develop a machine learning model. You have freedom to perform this case study in the software environment of your choosing (any Python IDE or a Jupyter Notebook, for example) but your deliverables should include the following:

- Your Python code (exploratory data analysis, data preparation, model training, model evaluation, etc.). This can be a .py file (or files) or a Jupyter Notebook file (or files).
- The final dataset you used to train your model on. This should be in .csv format.
- A basic summary of your work that contains:
    - A list of the features that your model uses. Include descriptions for the features you engineered, if any.
    - The relevant performance metrics that you used in model evaluation. Include metric name and values for your model.
- Your concise answers to the following questions:
    - What features, if any, did you decide to exclude from the original dataset? How did you decide?
    - What are the strengths and weaknesses of your model in classifying whether a patient will arrive at their appointment or not?
    - Broadly speaking, what steps do you think would be necessary to implement this model in production?