How to organize and showcase your Data Science portfolio project

Sadat Shahriar,
EMK center

Analysis stage

A data science portfolio project is really a continuous process. You have to work in a back and forth way to get the best possible model. But in general, the following steps are important:

1. Problem Definition
2. Clean the data
3. Answer critical business question. If not specified, try forming them yourself
4. Visualize the inter-feature relationship, and/or intra-feature distributions
5. Perform necessary feature engineering
6. If there are too much of features, try dimension reduction
7. Formulate the problem for machine-learning algorithm

Algorithm-Development stage

When you have to make prediction, you have to start with the simplest rational model possible. Often simplicity is the best. But in many situations, you have to dive deep. Follow the steps for that:

1. Define clearly the train, validation and test data
2. Define and explain what should be your evaluation metrics and why
3. Which learning model should you use. Try avoiding any random model. Be logical in your views and have answers for your baseline models.
4. What other models you are using and why?
5. Which model worked the best? What did you do different or why it’s performing better than any other model?
6. Visualize the model performance.
7. What else can be done to improve the performance even more? What are you gonna do in the fututre?

Project Showcasing stage

Unfortunately, developing a great model is not enough. You have to showcase your project for both technical and non-technical audience. Follow the tips:

1. MUST have a github repository for your project. This is not tough, and you DONOT need coding skill for that. Search on google (A nice link: https://www.youtube.com/watch?v=WfhRyz3Wf4o)
2. Organize your codes if and when showcasing a project in github. Give them meaningful names.
3. Organize your codes in-general. Try following pep-8 guidelines for Python code.
4. MUST have a self-explained readme file, where you clearly explain what you are doing.
5. Optional but recommended-- have a powerpoint or pdf presentation file

Deployment Stage

This is really an optional stage and NOT FOR NOW kind of a thing. Once you excel in previous stages, you have to concentrate on this very crucial and sophisticated stage.

1. Simple: Heruku deployment
2. Cloud deployment: AWS/GCP
3. Docker container
4. Other deployment tools (Jenkins, tf etc)

For this course project

You will have to have a --

1. Github repo, where the project code is uploaded
2. The first three stages.

I am not expecting a complete project but I am interested to see atleast 50% work has been done and organized in a github repo. With that, a clear plan of how to move forward.

Check my github repo for Data Science Project and more:
https://github.com/sadat1971/Price-Prediction-in-E-Commerce