An End-to-End Data Science Project

Deena Gergis
Lead Data Scientist @ Bayer

#ACCELERATE

Workshop overview:

Session 1
Prep & Analytics
12.09.2021

Start with the business problem, set the foundation up, find data source, preprocess Start the descriptive analytics pipeline)

Session 2
Machine learning
19.09.2021

Implemented analytics pipeline, Build and evaluate prediction model(s), use MIflow to keep track of the various experiments

Session 3
Deployment in Prod
26.09.2021

Create prediction functions and production class, develop an API, create a dashboard that the user will access and call the API

What you will do:

- **During the sessions:** You will get tasks to be done
- **After the sessions:**
  - You will complete the whole covered phases
  - Dig deeper into the various technologies discussed

i.e.: No Spoon-feeding :-)

#ACCELERATE

& let's get started and pick-up where we've left off

Part 1
Prep & Analytics follow-up

Part 2
Machine learning hands-on

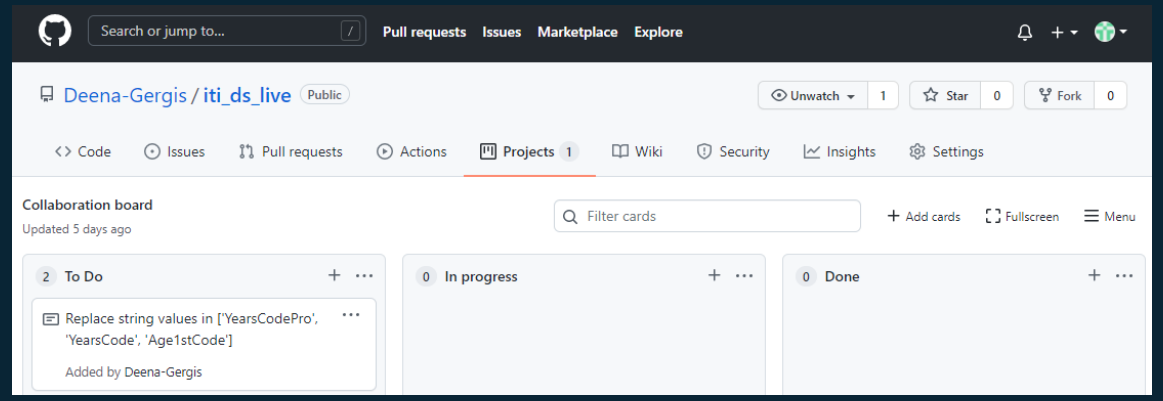
Part 1
Prep & Analytics follow-up

#ACCELERATE

Open points from session 1

Github board:

https://docs.github.com/en/issues/organizing-your-work-with-project-boards/managing-project-boards/about-project-boards



Legal responsibility:

https://www.nytimes.com/2017/08/25/business/volkswagen-engineer-prison-diesel-cheating.html

The New York Times

Volkswagen Engineer Gets Prison in Diesel Cheating Case

Mr. Liang, who helped develop the software that concealed high levels of pollutants generated by Volkswagen's diesel engines, reached a plea deal with prosecutors last year after agreeing to assist in the government's investigation of the company.

But even after that pledge, Mr. Liang received a harsher sentence than the government recommended for pleading guilty to conspiracy to defraud the United States and violating the Clean Air Act.

Part 1: Preprocessing

1. Let's see a sample notebook

2. Your turn:

Refactor your preprocessing notebook to have:

1. **Reusable Functions**
2. **Proper docstrings**
3. **Readable code**

Part 2: Descriptive analytics

1. Framework recap:

Asking the right question is half of the answer

- **Categories:** *General, Jobs, Skills, Relation*
- **Output:** *Numeric, Visualization, Unsupervised learning*

2. Your turn:

- **Show us one your descriptive analytics results**
- **Tell us how those results will help you with the modelling**

3. Let's see a sample analysis

Part 2: Descriptive analytics

1. Framework recap:

Asking the right question is half of the answer

- **Categories:** *General, Jobs, Skills, Relation*
- **Output:** *Numeric, Visualization, Unsupervised learning*

2. Your turn:

- **Show us one your descriptive analytics results**
- **Tell us how those results will help you with the modelling**

3. Let's see a sample analysis

Part 3. Unsupervised to Supervised

Unsupervised to Supervised

T-SNE

Stands for t-distributed stochastic neighbor embedding.
Nonlinear dimensionality reduction technique

Agglomerative Clustering

Recursively merges the pair of clusters that minimally increases a given linkage distance

Silhouette metric

The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

Part 4.
Mlflow Tracking

Tracking GUI

mlflow

Github Docs

Listing Price Prediction

Experiment ID: 0 Artifact Location: /Users/matei/mlflow/demo/mlruns/0

Search Runs: metrics.R2 > 0.24 Search

Filter Params: alpha, lr Filter Metrics: rmse, r2 Clear

4 matching runs Compare Selected Download CSV

					Parameters		Metrics		
	Time	User	Source	Version	alpha	l1_ratio	MAE	R2	RMSE
☐	17:37	matei	linear.py	3a1995	0.5	0.2	84.27	0.277	158.1
☐	17:37	matei	linear.py	3a1995	0.2	0.5	84.08	0.264	159.6
☐	17:37	matei	linear.py	3a1995	0.5	0.5	84.12	0.272	158.6
☐	17:37	matei	linear.py	3a1995	0	0	84.49	0.249	161.2

Tutorial

Assignment:

Train and track your predictive models

#ACCELERATE

Questions?

#ACCELERATE