Data Science Workshop

Hosted at Park University

Micah Melling

Micah Melling

Current: Chief Data Scientist at Americo Financial

Former: Senior Director of Data Science at Spring Venture Group

Bachelors in Econ from UCM

Professional Certificate in Data Science from Georgetown

Masters in Data Science from Rockhurst

UMKC Analytics Advisory Board

Park University Adjunct Faculty

Why We Are Here

Build a production-ready machine learning application, deploy it on AWS, and learn data science tips and tricks along the way.

Agenda

1. Get set up.
2. Walk through the data science pipeline with a real project.
3. Deploy!

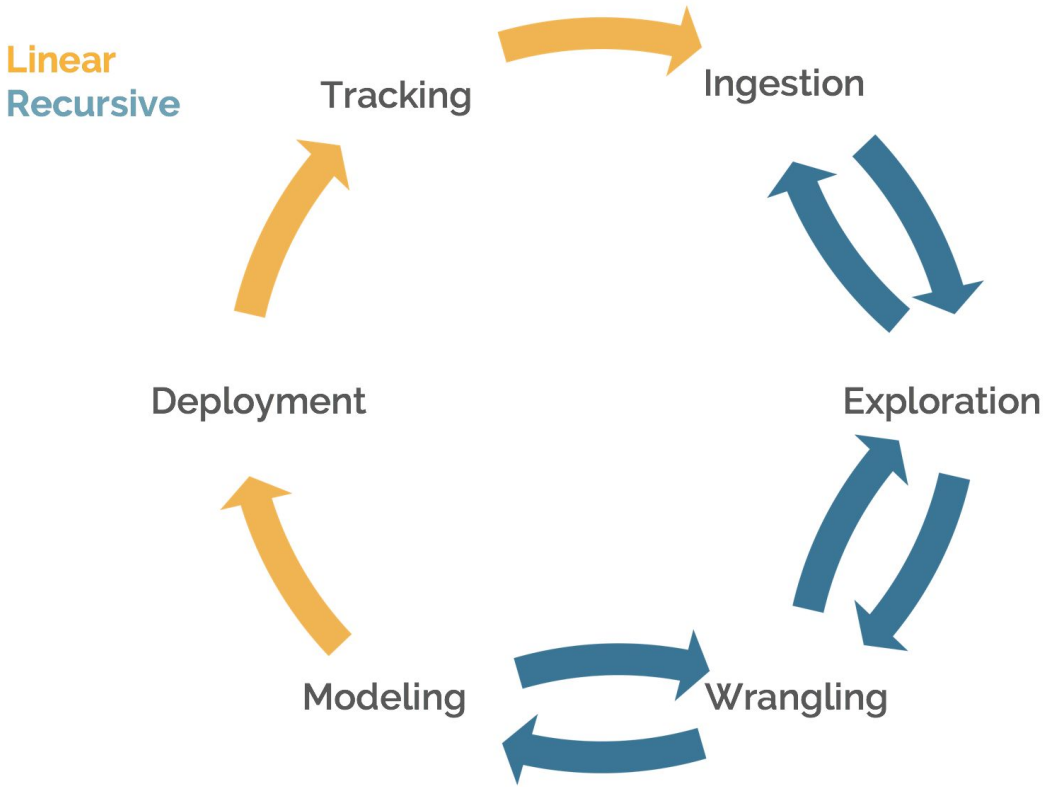
Learning Outcomes

1. Understand the components of a production ML application.
2. Get experience with putting a model in the cloud and making it accessible.
3. Learn data science tips and tricks along the way.

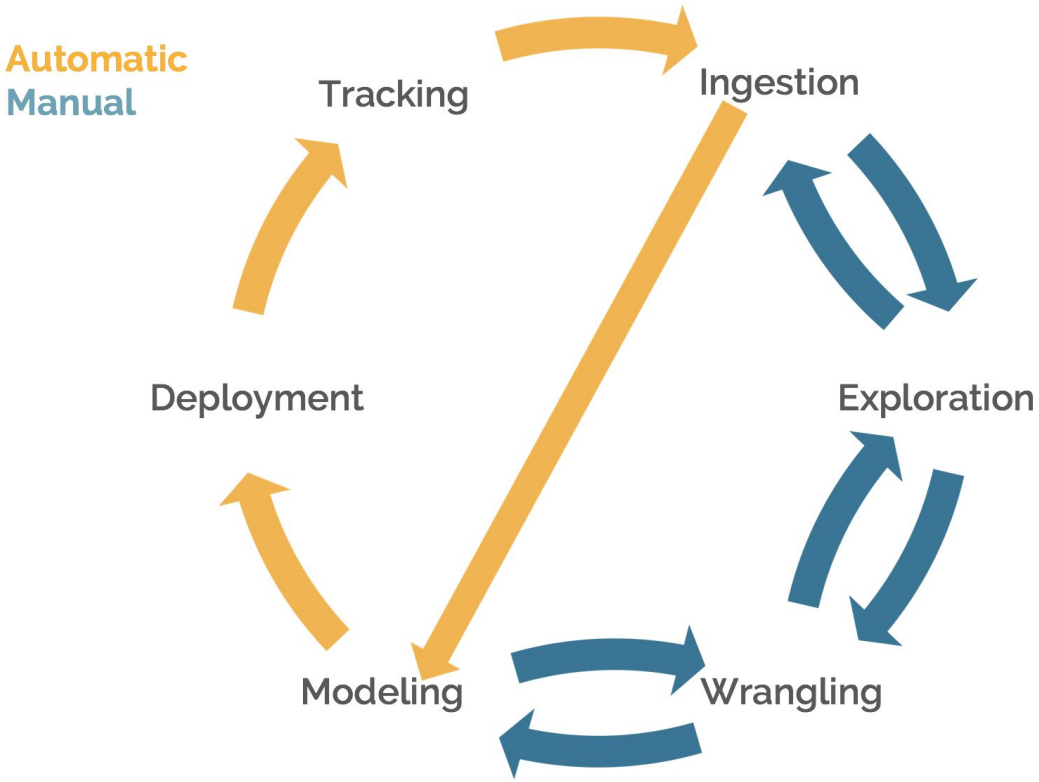
Data Science Pipeline

https://www.linkedin.com/pulse/understanding-data-science-pipeline-micah-melling/

Data Science Pipeline



Data Science Pipeline



Getting Started - GitHub

Clone the repo: git clone https://github.com/micahmelling/prod-app-workshop.git

An Aside: Using Copier for Project Templates

Templates are good - they give us a repeatable start place!

Copier is a useful solution.

https://copier.readthedocs.io/en/stable/

$ python3 -m copier path/to/project/template path/to/destination

Getting Started - AWS

AWS can incur charges! That said, our deployment will be serverless and should be de facto free (and can be taken down quickly and easily).

Create an AWS account

Create an Admin user with programmatic access

Set access keys as environment variables

Create an AWS Account

https://signin.aws.amazon.com/signup?request_type=register

Create a Programmatic Admin User and Set Env Variables

https://docs.aws.amazon.com/streams/latest/dev/setting-up.html

With more time, we would opt for a mechanism to grant temporary, one-time access keys.

```
$ export AWS_ACCESS_KEY_ID=...
$ export AWS_SECRET_ACCESS_KEY=...
$ export AWS_DEFAULT_REGION=us-west-2
```

The above will set the environment variables in your current terminal session. Once you exit the session, the keys will no longer be there. To set permanent keys - which is not always recommend (see above) - use your bash_profile.

Data Ingestion

In this case, it’s as simple as reading a csv!

Data Documentation:
https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset

(In our code repo, we make some slight adjustments to our data to make it more interesting).

Data Wrangling

The dataset is fairly clean, but we still need to perform some basic wrangling:

- Fill in missing values
- Handle outliers
- Drop unwanted columns

Data Wrangling Pipeline

Ideally, we want to wrap our wrangling code into a single pipeline that is coupled with our model.

Useful Videos (for later)

https://youtu.be/4dGv_6QT2Xw?si=gkDjqnBaOog0hvbI

https://youtu.be/frqcuPwgOl8?si=PcmNu33aVSzjxg1T

Optimizing Data Wrangling and Feature Engineering

Likewise, we can tune our wrangling / engineering in concert with our model's hyperparameters.

Useful Video (for later)

https://youtu.be/8rT4PM3w6ME?si=A4gfU3GN2vpk1jt6

Model Calibration and Prediction Intervals

In machine learning, we always want to quantify uncertainty.

In classification, we want a calibrated model. That is, we want our predicted probabilities to map to real-world probabilities.

In regression, we want a prediction interval. That is, we predict the value will be between X and Y 90% of the time, etc.

More on Model Calibration…

https://endtoenddatascience.com/chapter11-machine-learning-calibration

https://youtu.be/bbvZffubbIQ?si=57WKQstTpH6U14PJ

More on Prediction Intervals…

https://mapie.readthedocs.io/en/latest/

https://youtu.be/RTBmBZtBtuE?si=HISYe2Zywf5yUvzR

Model Optimization

Better options than grid search and randomized search exist.

https://endtoenddatascience.com/chapter10-machine-learning

https://youtu.be/_z8Ri_LwD5E?si=xJ7PyNoRUdKX0FxL

Model Evaluation

We want to evaluate our model on a suite of metrics. One metric does not rule them all.

Model Explanation

$ pip install auto-shap

https://youtu.be/1D_EaiyMwuI?si=Nz2VTTgom4AnbEPZ

Zappa

Zappa is an easy way to deploy Flask apps on AWS Lambda, a “serverless” architecture.

https://github.com/zappa/Zappa

Data Science Workshop

Hosted at Park University

Micah Melling