

Project Workflow

DATA 301

Generic workflow

1. What do you want to do? **predict? Cluster? Something else?**

Generic workflow

1. What do you want to do? predict? Cluster? Something else?
2. Get data proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?

Generic workflow

1. What do you want to do? **predict? Cluster? Something else?**
2. Get data **proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?**
3. Preprocess the Data **you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc...**

Generic workflow

1. What do you want to do? **predict? Cluster? Something else?**
2. Get data **proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?**
3. Preprocess the Data **you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc...**
4. Explore the Data **plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?**

Generic workflow

1. What do you want to do? **predict? Cluster? Something else?**
2. Get data **proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?**
3. Preprocess the Data **you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc...**
4. Explore the Data **plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?**
5. Model Data **build, fit and validate a model**

Generic workflow

1. What do you want to do? **predict? Cluster? Something else?**
2. Get data **proprietary Data? Web Scraping? Public Domain? What about confidentiality? How to join datasets? What if data is too big? Or too small?**
3. Preprocess the Data **you will spend a lot of time here, clean NaN's, normalize fields, ensure strings are consistent (VA,Va, Virginia etc..), encode ordinal data, etc...**
4. Explore the Data **plot it, are there anomalies? can you see patterns? How to plot if you have >3 features?**
5. Model Data **build, fit and validate a model**
6. Evaluate Model **not good enough? Go to step 2, consider ensembling multiple models, consider simple default model to compare against**

Critical Bits

1. Domain Expertise is essential. For instance:
 - When evaluating medical images for tumors, it helps if you can recognize a tumor.
 - When exploring clustered botanical data its useful to be able to verify that groupings make sense.
 - When checking engine sensor data, it helps if you have some idea of what typical running parameters are and how anomalous behavior presents.
2. Communication and presentation skills are the most important of all.
 - If you cannot convince stakeholders to follow you, then you have wasted your time

What we covered

General project workflow
Data preprocessing, cleaning, EDA, data leakage, Pandas
Clustering
Splitting a dataset
Handling Dataset imbalance
Regressors verses classifiers
Linear regression
Decision Trees
Random Forest
Gradient Boosted Trees
What tree based algorithms cannot do that regressions and Neural Networks can
Explainability
Hyperparameter tuning
Cross validation
Algorithm evaluation metrics(accuracy, Precision, Recall F1, R squared)
Time series analysis

See course website for complete list

Where to go after this course

- SQL – there is a lot of data in databases
 - Do you need a Database class?
 - Not for this course.
 - Professionally, it's a little trickier. You can learn what you need in a week or so and get pretty good after a month or 2.
- Do you need to scale your compute?
 - Not for this class
 - Professionally, yes. You don't use a laptop. At a minimum use a local desktop with a GPU (or GPUs)
 - What if your data is huge and will not fit in memory? Next step is something cloud based (pyspark, Dask, terality ..etc this area is changing fast)
- Dashboards
- You have to learn to use Linux
- Start creating your own projects
- Participate in Data Science competitions (Kaggle etc.)