BIRMINGHAM CITY
University

CMP6230 Draft Pipeline

DRAFT VERSION

Lewis Higgins - Student ID 22133848

Word count: XXXX

Contents

Candidate Data Sources

For the first stage of the pipeline, data ingestion, three data sources will be identified in order to find the one that would be most optimal for the production and deployment of a machine learning model to complete a supervised learning task.

1.1 Notes - DELETE BEFORE SUBMISSION

Lucidchart can generate ERDs from CSVs. For each, show the pandas head, column data types and ERD. Finding one with multiple CSVs can be good to make the ERD look more complex. **So far, all of your data is from Kaggle. Consider another source like data.gov.uk, especially considering that it can give you raw data for preprocessing.** On Kaggle, the "Provenance" section will have the source if the description doesn't. If neither have a source, it's probably fake data.

- Loan data
  - Fictional, so hard to give a good problem statement because this isn't real.
  - Classification - Should they be given a loan?
  - Unlikely to be of any use, should find another.
- Smoke detection
  - Real data
  - Lots to explain (how the alarms work etc)
  - Preprocessed, but you could still do more (remove timestamp etc)
  - Classification - Should the smoke/fire alarm sound?
- Employee data
  - Allegedly real data though it seems hard to believe.
  - Classification - Is the employee likely to find another job instead?
- Australian weather
  - Promising. Real data, but very large. Can do lots of preprocessing.
  - Classification - Will it rain tomorrow?
- Cardiovascular disease
  - Good data, enormous amount of it (laptop might not handle it), good source.
  - However, it's already been processed. Check if that's fine or not.
  - Classification - Do they have heart disease?
- Diabetes
  - In consideration for CMP6200, cannot use a dataset in both.

- It's actually real data according to the Kaggle page, from "multiple healthcare providers" and the electronic health records (EHRs) they keep.
- Preprocessed already, but duplicates exist in it.
- Only 9 features, is that a bad thing?
- Classification - Do they have diabetes?

The size of the datasets used for this module do not matter, unlike CMP6202. As such, OpenML and the smaller Kaggle sets are allowed. It might actually be *disadvantageous* for you to use large datasets because this will be done on a VM. Consider the classic Heart dataset with the 300 values. Simple and instant processing, even on your laptop. Even if you don't use it, it should be a candidate.

1.2 Candidate 1 - Smoke Detection Dataset