BT5152

AY2018/2019 Semester 1 Week **12**

W12: Association Rules and Imbalanced Dataset

Things we covering today

1. Association Rules [Demo]
2. Sampling Techniques [10 mins]
3. Assignment 4 Walkthrough
4. Assignment 5 Question 2

Associative Rules

What is/are the likely movies that customers will want to watch given she/he has watched LOTR1, LOTR2?

- This is also more known as collaborative filtering
- Other algorithms
  - Alternating Least Square (Matrix Factorisation technique)
  - Neural Collaborative Filtering (work by NUS!)
- However, there is a limitation of these algorithms

How to do Associative Rules in R?

Using **arules**

```
apriori(data = transactions, parameter = list(support = <support>, confidence = <minimum confidence>, minlen = <minimum basket size>))
```

The more important part is preparing the **transactions**.

Binary Classification is boring

Often not, in real problems, classification is always > 2 classes and perhaps, sometimes, multi-label as well.

Sampling techniques are available to overcome imbalance.

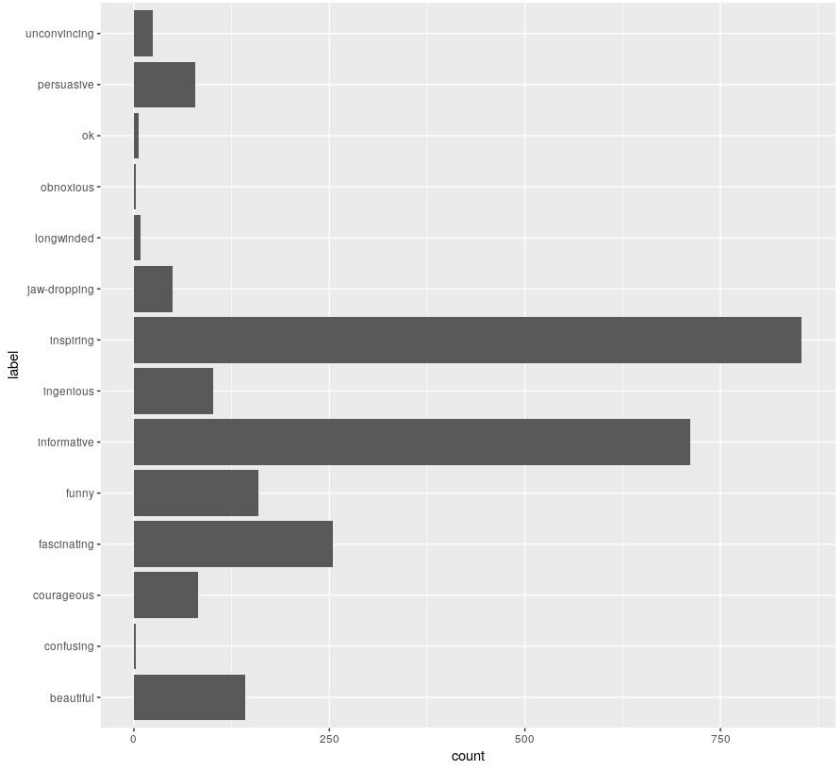
More often not, talking with your stakeholders to “balance” the dataset as the first step.

Pre-sampling

In Assignment 4, we have a few classes of ratings with very low frequency.

Reduce the number of classes to predict

- Group low frequencies ratings as one: name it as “others”?
  - `unconvincing + ok + obnoxious + longwinded + confusing`

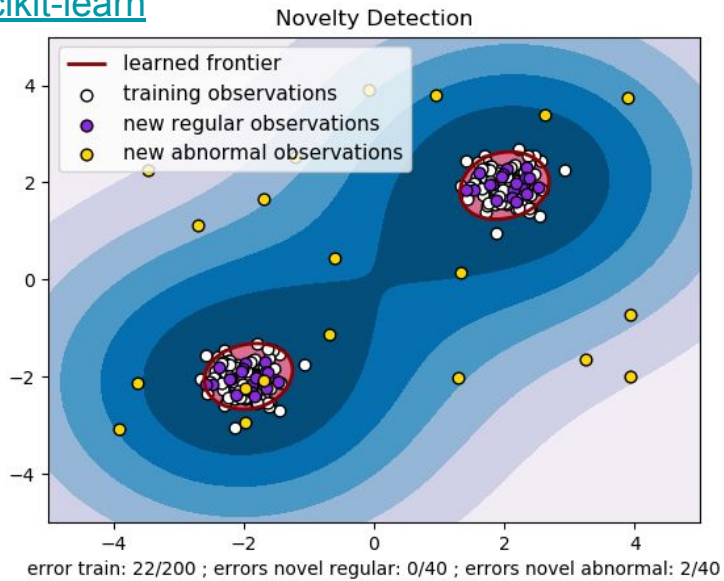


Alternatives

What if your stakeholders want to predict the low frequency classes as well?

- Train another model to just work on the low frequency classes; but even your “low” frequency classes need to have a decent sample size
- Detecting these low frequency classes is like anomaly detection?
  - Train multiple 1-class classifiers to detect these “anomaly” classes?
    - Novelty detection

Scikit-learn



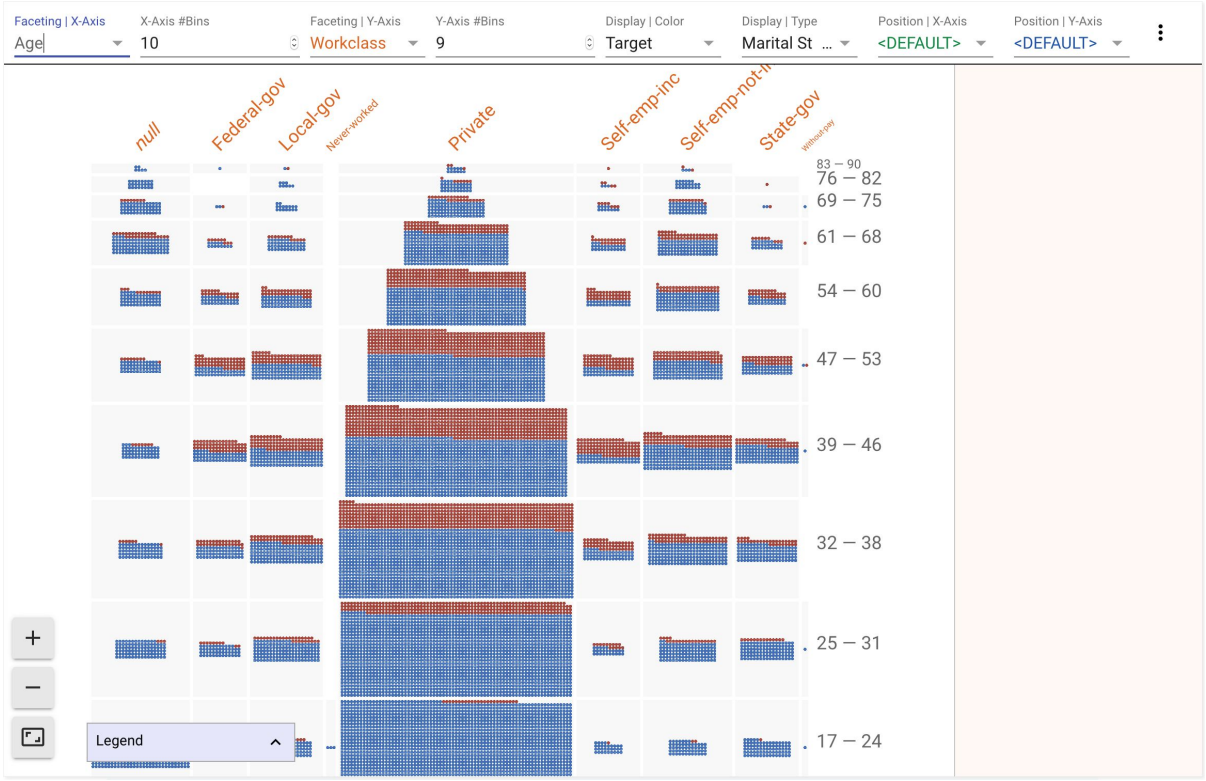
Understanding the data is key

More complex algorithms and models won't help.

This is a data issue. Using visualisation tools to help you understand the data.

Example of such software and tools:

- https://pair-code.github.io/facets/
- https://seaborn.pydata.org/generated/seaborn.pairplot.html
- Or any of your off-the-shelves BI visualisation software like Tableau or Power BI

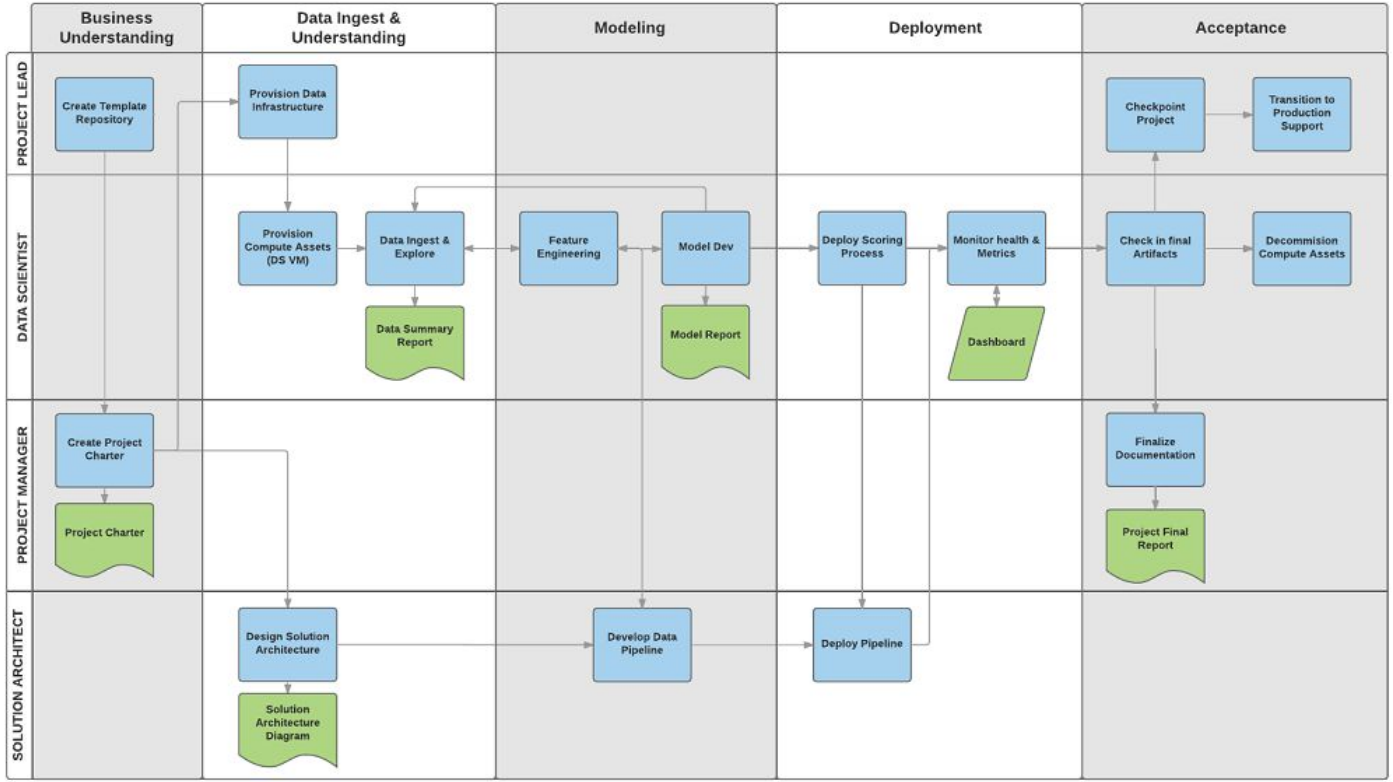


Assignment 4 Walkthrough

- Don’t over-complicate the solution 🧢
- Work within the constraints you have
- Make sure your naive end-to-end solution can run first, then you can start changing things like feature engineering, tuning, etc
    - Similar to software development lifecycle
- There is no “right” answer in machine learning, or modeling in generally. ***“All models are wrong, only some are useful.”***
- We can’t determine what is a “right” result, but we can identify non-ideal coding practices (DRY, magic number, etc) and wrong methodology.

Understanding language with math is hard 🤯

- Requires a lot of resources 💰 (RAM, CPU 🔥) because of its high dimensionality
- But with good coding practises, you can alleviate it
  - Create new variables judiciously; duplicating data complicates things
  - Use functionality of the libraries; don’t reinvent the wheel



Data Science Workflow (Microsoft, IBM)

One of my early runs

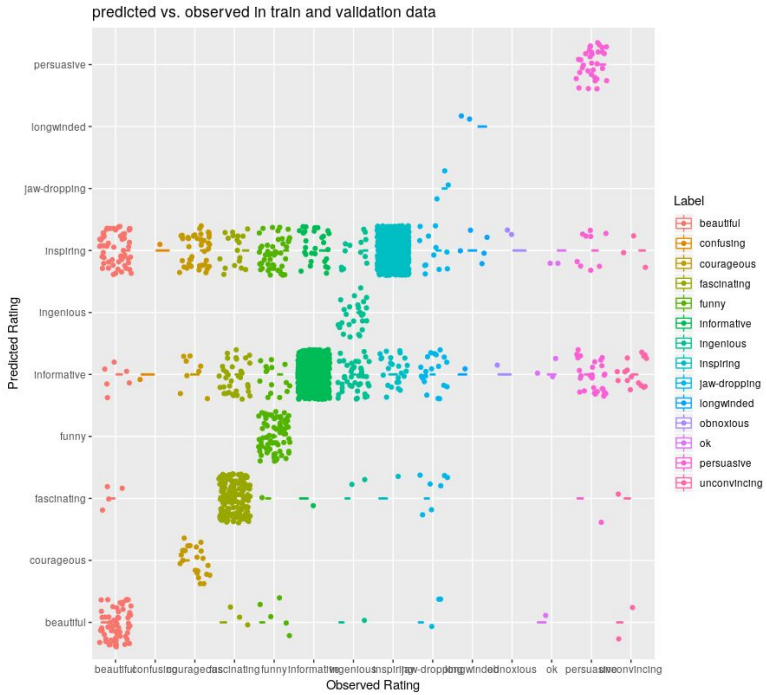
Training

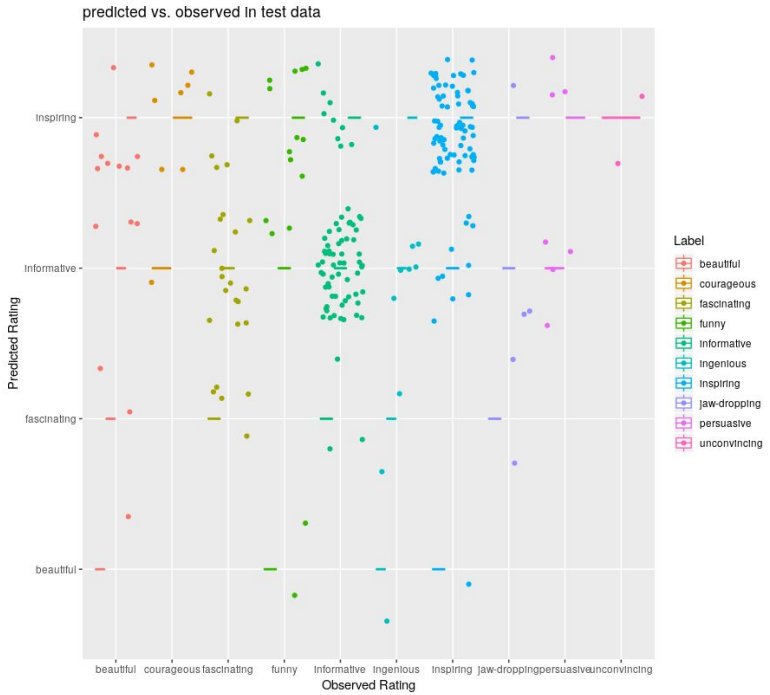
```
Overall Statistics

               Accuracy : 0.7838

               95% CI : (0.7661, 0.8007)

     No Information Rate : 0.3442

     P-Value [Acc > NIR] : < 2.2e-16

               Kappa : 0.7034
```

Test

```
Overall Statistics

               Accuracy : 0.5744

               95% CI : (0.5094, 0.6375)

     No Information Rate : 0.3512

     P-Value [Acc > NIR] : 1.252e-12

               Kappa : 0.3865
```

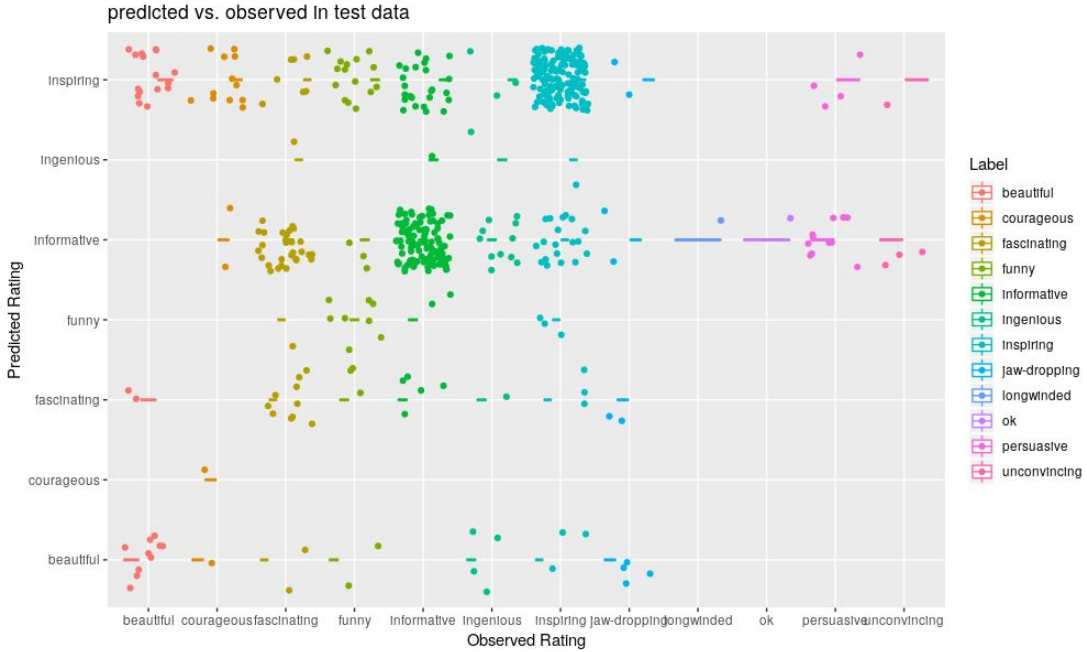
Confusion Matrix



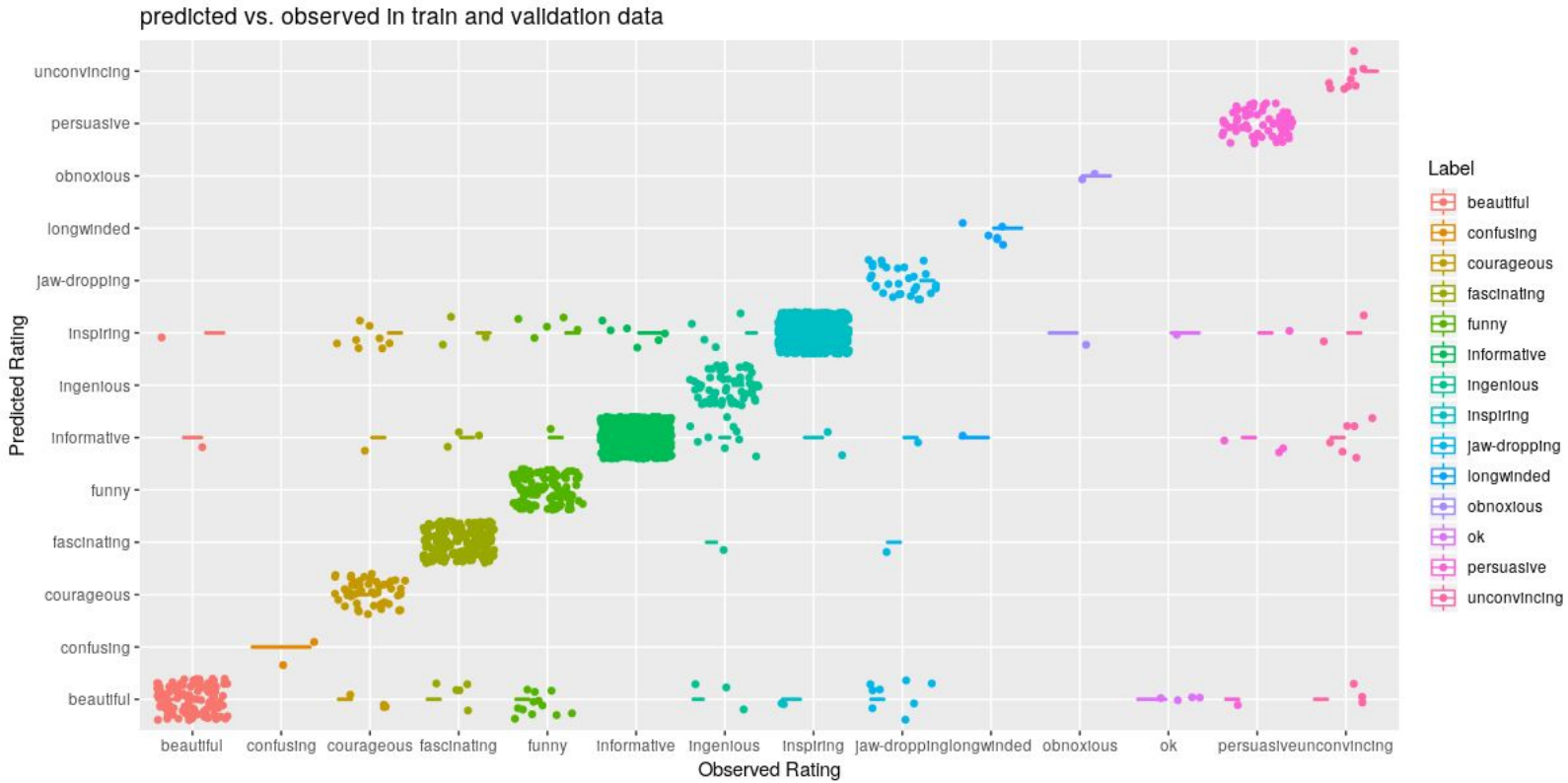


Best model with metrics

- 80% data used for training
- SMART stopwords
- Word Frequency > 350
- TF
- SVM Radial
  - Sigma = 0.00075
  - Cost = 2
- Macro F1: 0.3652629
- Micro F1: 0.5737705
- Kappa: 0.4031



Training Confusion Matrix



Assignment 5 Question 2

Home Credit Dataset (download from IVLE)

1. Using XGBoost to predict TARGET with AUC as the optimisation metric to establish a baseline
2. Sampling to improve performance
    - Under-sampling
    - Over-sampling
    - Under/over-sampling together
    - SMOTE