IMDb Data Exploration & Machine Learning

Katie Winkle, Katja Durrani & Vivienne Kuh

Women in Tech: Machine Learning Workshop Series

Introductions

What could we predict?

- IMDb data set included a huge range of features covering film details, actors & directors, financials etc.
- Each explored something different – Facebook likes, IMDb score & country of origin
- Presenting back on trying to predict the number of Facebook likes a film might receive
  - Social media massively relevant to modern day marketing & advertising
  - Some measure of pervasiveness/pervasiveness and therefore success

Where to start?
Pre-processing the Data

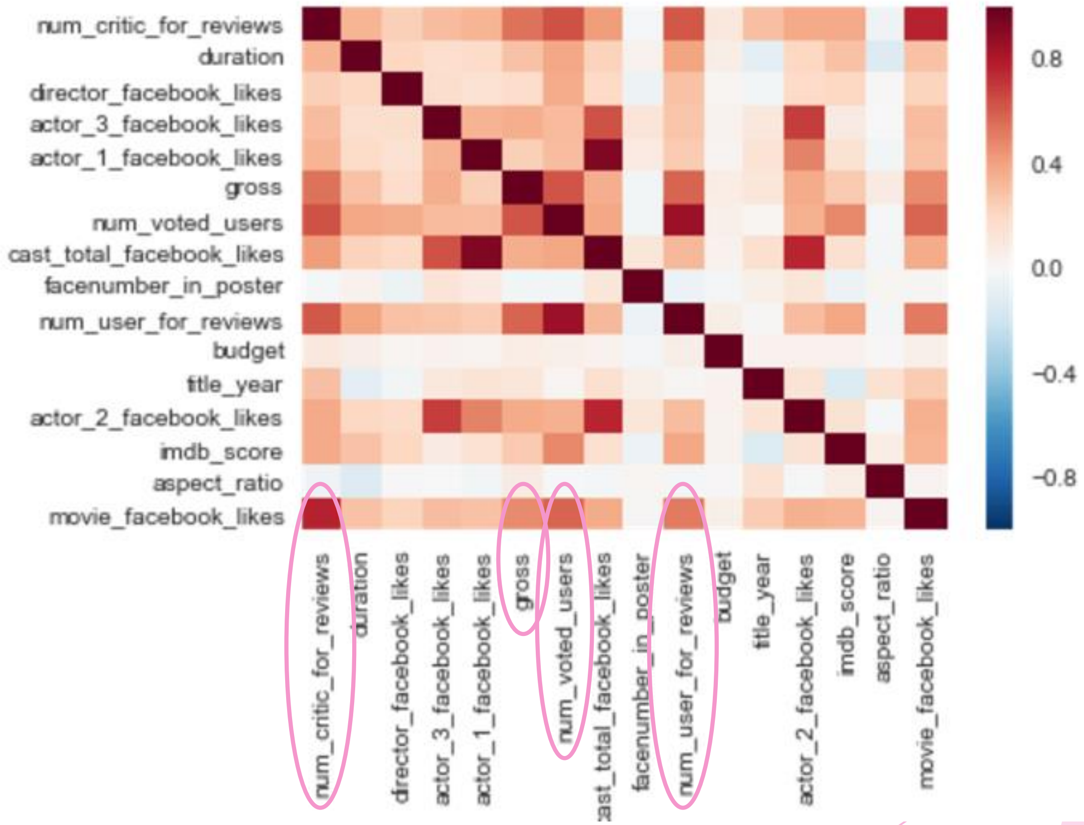
- Many 0 like entries, some of which definitely incorrect (e.g. Pirates of the Caribbean?!) so deleted all; ~2k data entries!
- Facebook likes is a continuous scalar – so need to discretise under pre-defined labels
- Wanted to assign ‘Low – Medium – High’ labels based on relative number of likes (bottom 25% = low, top 25% = high)
- Applying directly to Facebook likes didn’t work – ended up with too many films in the ‘medium’ category
- Instead ranked films based on Facebook likes and applied labels based on those rankings

Ready for a first run
Quantitative Only Full Feature Set KNN

- Minimum data manipulation for building a KNN classifier:
  - NaN’s replaced with zeros
  - 0 Facebook like data entries removed
  - Facebook like Low/Medium/High labels created
  - All qualitative data dismissed – KNN requires numbers!
- Used sklearn to scale/split data and train KNN just as in class
- Achieved a prediction accuracy of **54.8%** on the training data

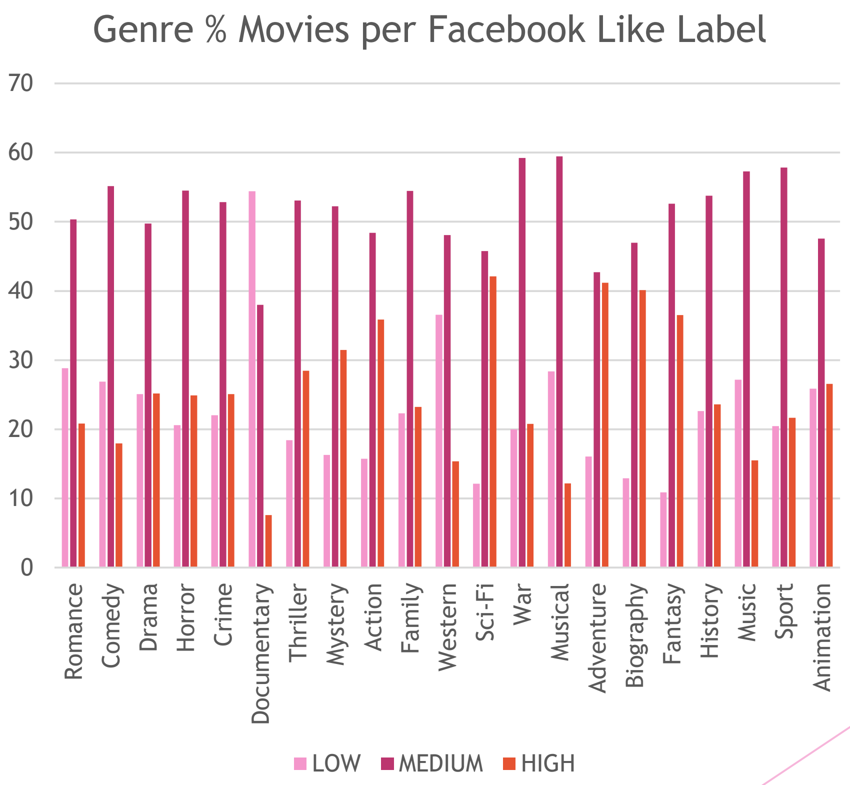
What's really relevant? Data Exploration & Visualisation

- Heatmap based on correlation matrix using pandas df.corr()
- Picked out variables with correlation >= 0.5 (0.499 for gross)
- Satisfying reflections:
  - Number of critic reviews strongest correlation
  - IMDb users like movies on Facebook?
  - Some correlation with gross
  - *Not* linked to actor/director Facebook likes
- Improved KNN classifier with **60.4%** accuracy



What about the text data we ignored? Data Exploration & Visualisation

- Example: is genre relevant?
- Pre-processing:
  - Multiple genres per film – treated as ‘tags’ rather than primary genre
  - Python scripting for splitting genre list in CSV, counting etc.
- Satisfying reflections:
  - Documentaries rarely labelled as high, most Sci-Fi films medium/high – Facebook user demographics?



How to deal with text?
Coding Qualitative Data

- Example workaround for genres: binary classification
  - Does it have the action 'tag'?

	num_critic_for_reviews	gross	num_voted_users	num_user_for_reviews	Action	Adventure	Fantasy	Sci-Fi	Thriller	Documentary	...	Drama	History	Sport	Crime	Horror	War	Biography	Music	facebook_label
0	1.0	0.0	57.0	1.0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	LOW
1	0.0	0.0	128.0	3.0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	LOW
2	0.0	0.0	33.0	0.0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	LOW
3	0.0	0.0	114.0	6.0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0	LOW
4	6.0	0.0	117.0	6.0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	LOW

- Not full solution (depends on number of possibilities e.g. language vs directors)
- Addition of all genre data reduced classifier accuracy to **59.8%**
- Genres with one label >50% pushed accuracy up to **61.2%**
  - Higher than what we achieved with the quantitative data alone

Did we miss anything?
Further Data Pre-Processing

- Took another look at selected features data set
- Removed some missing entries for gross
- Huge difference in scale of e.g. gross versus number of critic reviews, so normalised all data with respect to their maximum
  - Increased accuracy again to **65.7%**
- Realised gross in different currencies – removed
  - Increased accuracy again to final value of **67.9%**

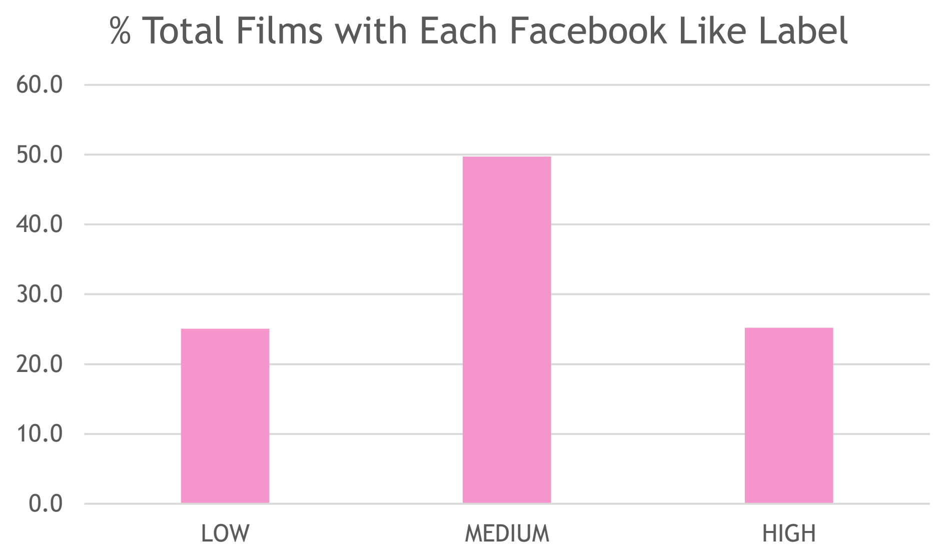
So can we predict Facebook likes? Final KNN & Reflections

- Final data features used by classifier:
  - Number of critic reviews
  - Number of users who’ve voted
  - Number of users who’ve left reviews
- Can predict Facebook like label (H-M-L) with accuracy of ~68%

- Suggests IMDb user behaviour somewhat reflective of Facebook user behaviour
  - Overlapping demographics in IMDb & Facebook users
  - IMDb users are a good test case for predicting film popularity

How good *really* is it?
Comparing to 'Chance'

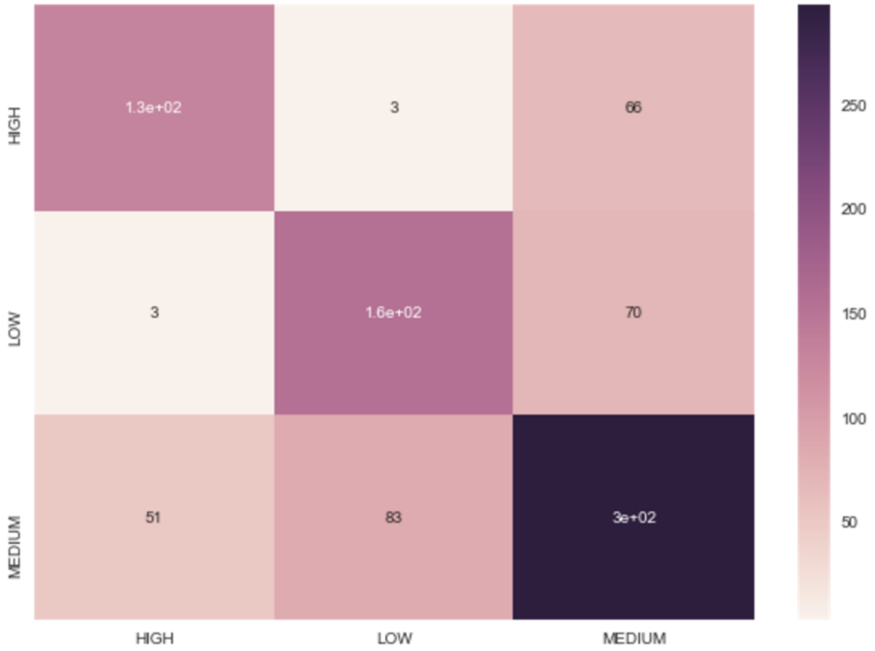
- Pure classifier accuracy isn't always enough
  - Breast cancer recurrence: always predicting no would be 70% accurate[1]
- In the Facebook like case, always guessing 'Medium' would be ~50% accurate
  - Initially caught us out, you would expect 'chance' to be 33% as 3 possible labels



[1]http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

How else can we check? Confusion Matrix

- Shows misclassifications by label
- Reflections:
  - High & low rarely confused
  - Most often correct on medium films -> reflection of data set as discussed prev?



So what did we learn?

- Importance of pre-processing of data (non-trivial)
  - Missing/incorrect entries
  - Qualitative data
  - Demonstrated accuracy improvement
- Data exploration, understanding & feature selection is vital
  - Demonstrated accuracy improvement
  - Also allows for meaningful reflections at this stage
  - Understanding data spread important for then assessing predictor performance
- Classification accuracy is not always the best performance indicator
  - Confusion matrix one example of an alternative

What would we ask next?

- How else might we have pre-processed the data?
  - Clearly this can have big impact on predictor performance!
- How else could we choose which features to train on?
  - Features analysis methods
- How else could we deal with qualitative data?
  - Other algorithms as well as coding methods
- What other learning algorithms might be better suited to this task?
- How might our reflections be useful in a business context?