
A burrito
Pusheen.com


IMDb

- Intro to team Pyfun
- Aim
- Data of interest
- Pre-processing:
- Results
- What we would do next

Our aim:

Knowing only the things that we know about a film before it is made, predict the return on investment it will generate

Our aim:

Knowing only the things that we know about a film before it is made, predict the return on investment it will generate

Budget	Director CV	Genre	Content rating	Duration
$, continuous	*?*	*List of tags*	*One tag*	*Minutes, continuous*

ROI
$\frac{Gross\ earnings}{Budget}$

Pre-processing

- Remove null and duplicates
- Remove all films not in USA – so budget and earnings are in same currency(!)
- Correct for inflation
- Calculate director track record
- Normalise the feature data
- Standardise certification feature

```
In [7]: grouped = ddata.groupby('director_name')
```

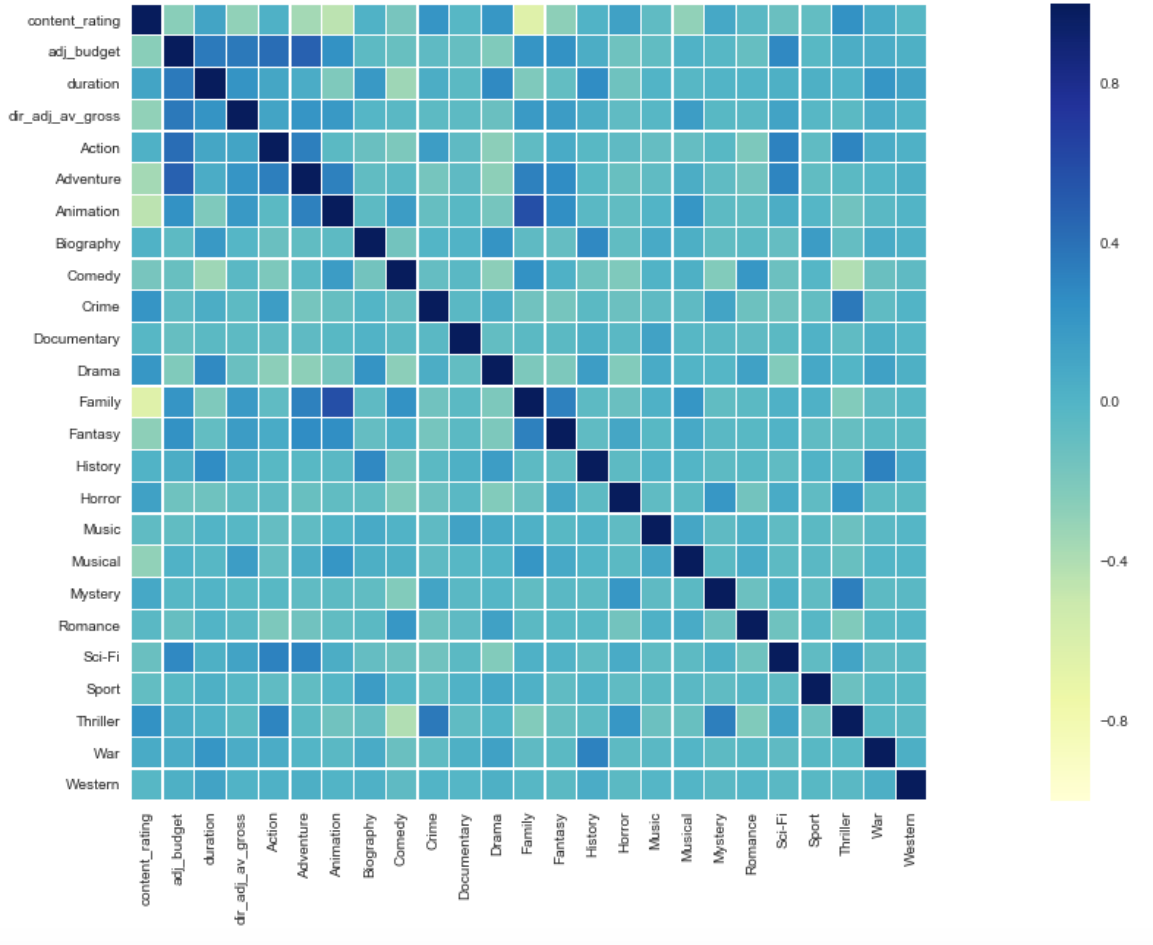
```
In [8]: av = grouped.mean()
```

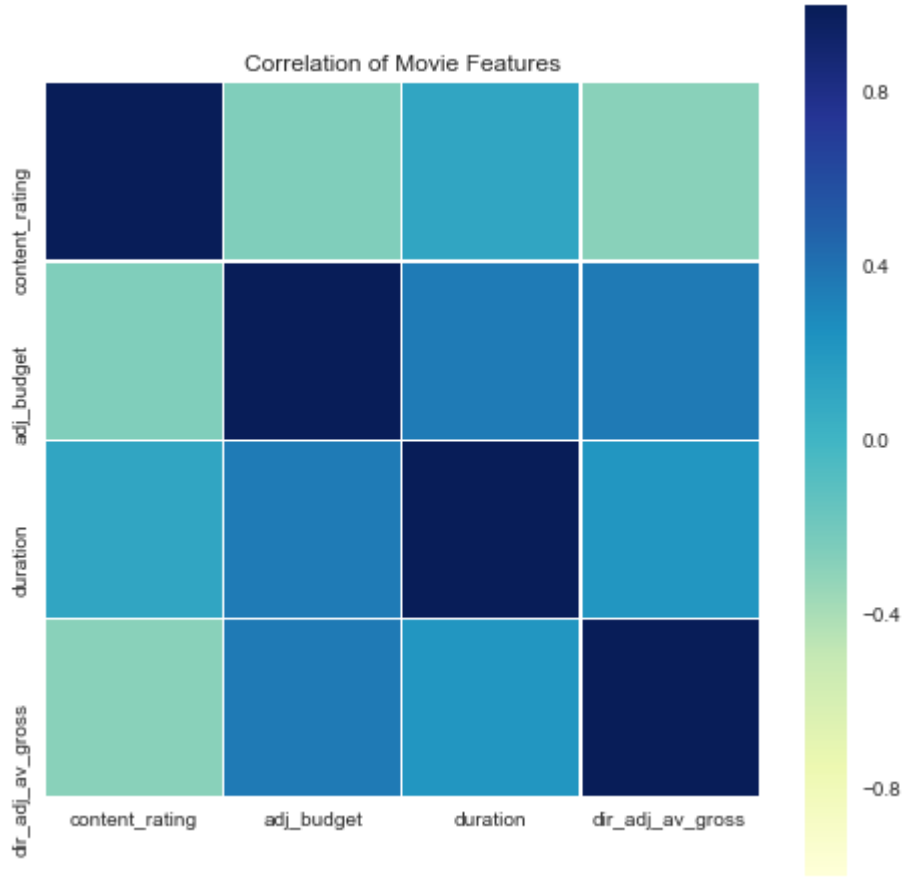
```
In [126]: budg_norm = (budg-budg.min())/(budg.max()-budg.min())
          gross_norm = (gross-gross.min())/(gross.max()-gross.min())
          dur_norm = (duration-duration.min())/(duration.max()-duration.min())
```

G – General Audiences *All ages admitted*. Nothing that would offend parents for viewing by children.	**G**: General Audiences – all ages admitted	**G**: General Audiences – All ages admitted	**G** : General Audiences
PG – Parental Guidance Suggested *Some material may not be suitable for children*. Parents urged to give "parental guidance". May contain some material parents might not like for their young children.	**PG**: Parental Guidance Suggested – some material may not be suitable for childre	**GP**: All Ages Admitted – Parental Guidance Suggested (Changed to PG in 1972)	**M** : Suggested for Mature Audiences – parental discretion advised
PG-13 – Parents Strongly Cautioned *Some material may be inappropriate for children under 13*. Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.	**PG-13 –** added Parents Strongly Cautioned – some material may be inappropriate for children under 13		
R – Restricted *Under 17 requires accompanying parent or adult guardian*. Contains some adult material. Parents are urged to learn more about the film before taking their young children with them.	**R**: Restricted – Under 17 requires accompanying parent or adult guardian	**R**: Restricted – Under 17 requires accompanying parent or adult guardian	**R**: Restricted – persons under 16 not admitted, unless accompanied by parent or adult guardian.
NC-17 – Adults Only *No One 17 and Under Admitted*. Clearly adult. Children are not admitted. (Changed from X in 1990)	**X**: No One Under 17 Admitted	**X**: No One Under 17 Admitted	**X**: Persons Under 16 Not Admitted
late 1990s	**1984 to 1990**	**1970 (1972) to 1984**	**1968 to 1970**

	Raw Data	Pre-processed Data
lines	*5044*	*2941*
features	*28*	*25 including genre tags* *4 not including genre tags*

Looking at the data before we do anything





What we know about profit

- What score do we have to beat to know our model is any good?

- In the whole dataset – 60.3% of the films returned a profit

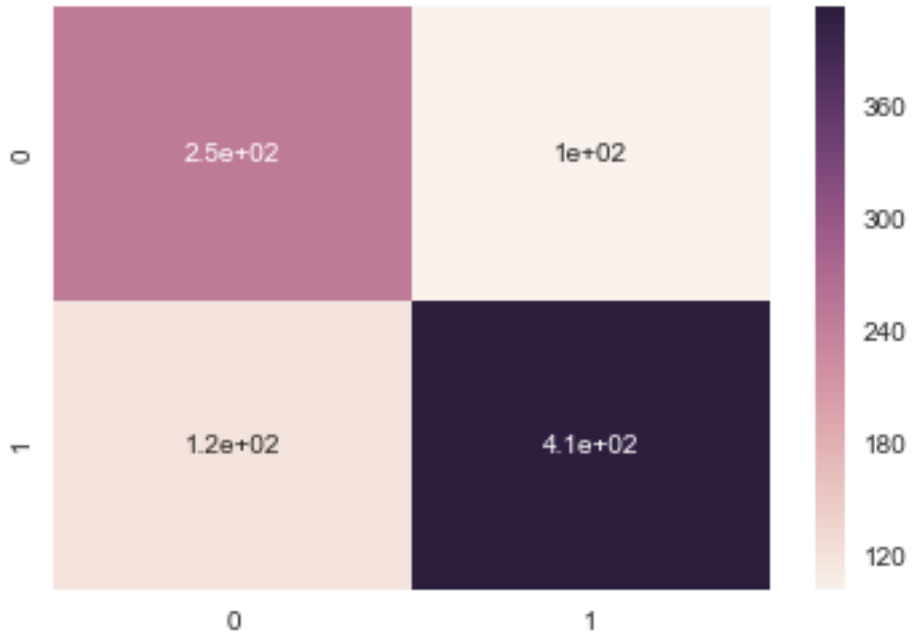
KNN binary classifier

- N neighbours:
  - 30 or 300? 30 is ~2% better
- Accuracy: 63%

Making a profit	Making a loss
Precision: 69%	Precision: 53%
Recall: 70%	Recall: 51%

Random Forest binary classifier

- Feature significance:
  - Content rating – 0.04
  - Budget – 0.35
  - Duration – 0.19
  - Director – 0.42

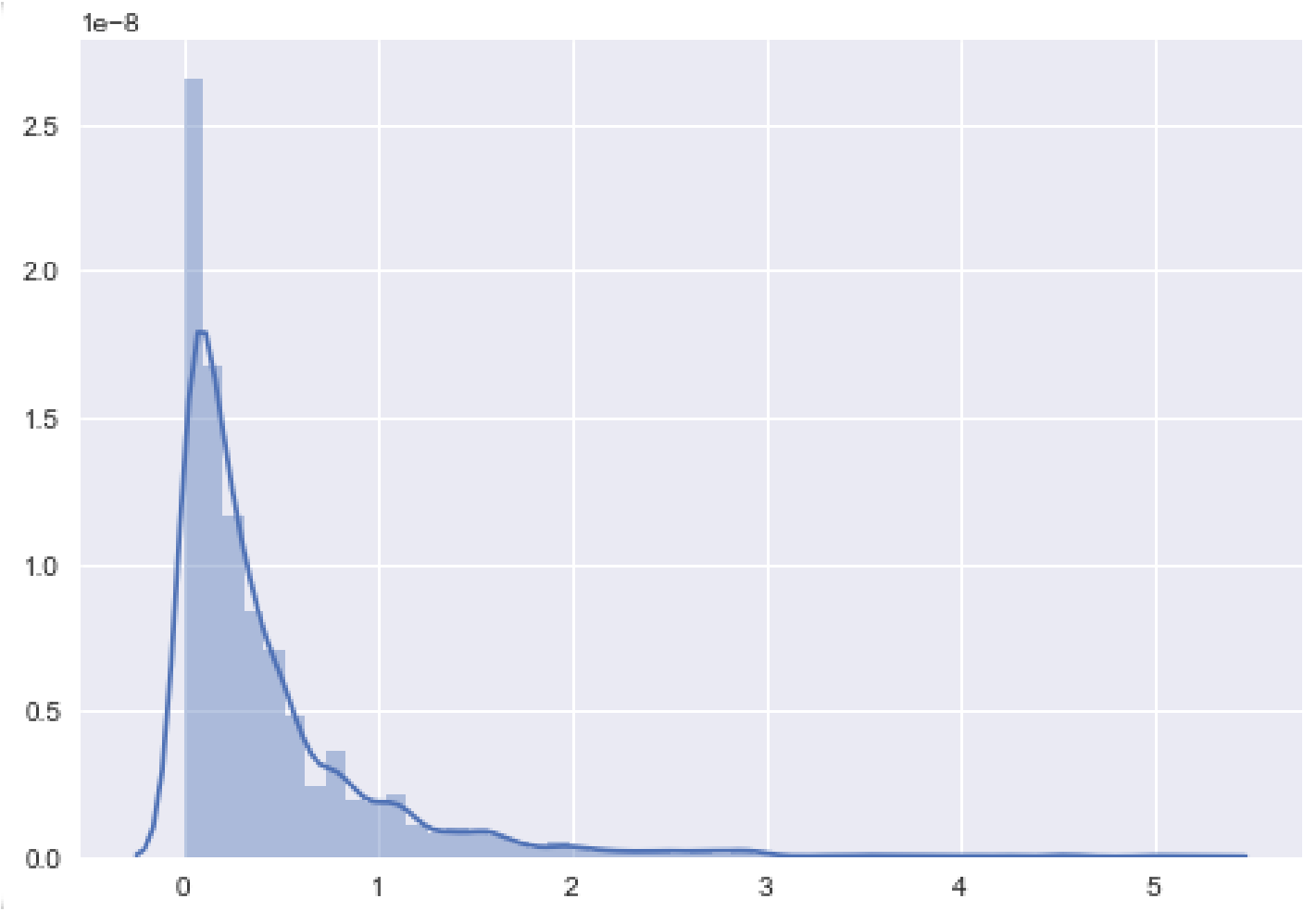


- Accuracy: 74%

Making a profit	Making a loss
Precision: 77%	Precision: 71%
Recall: 77%	Recall: 68%

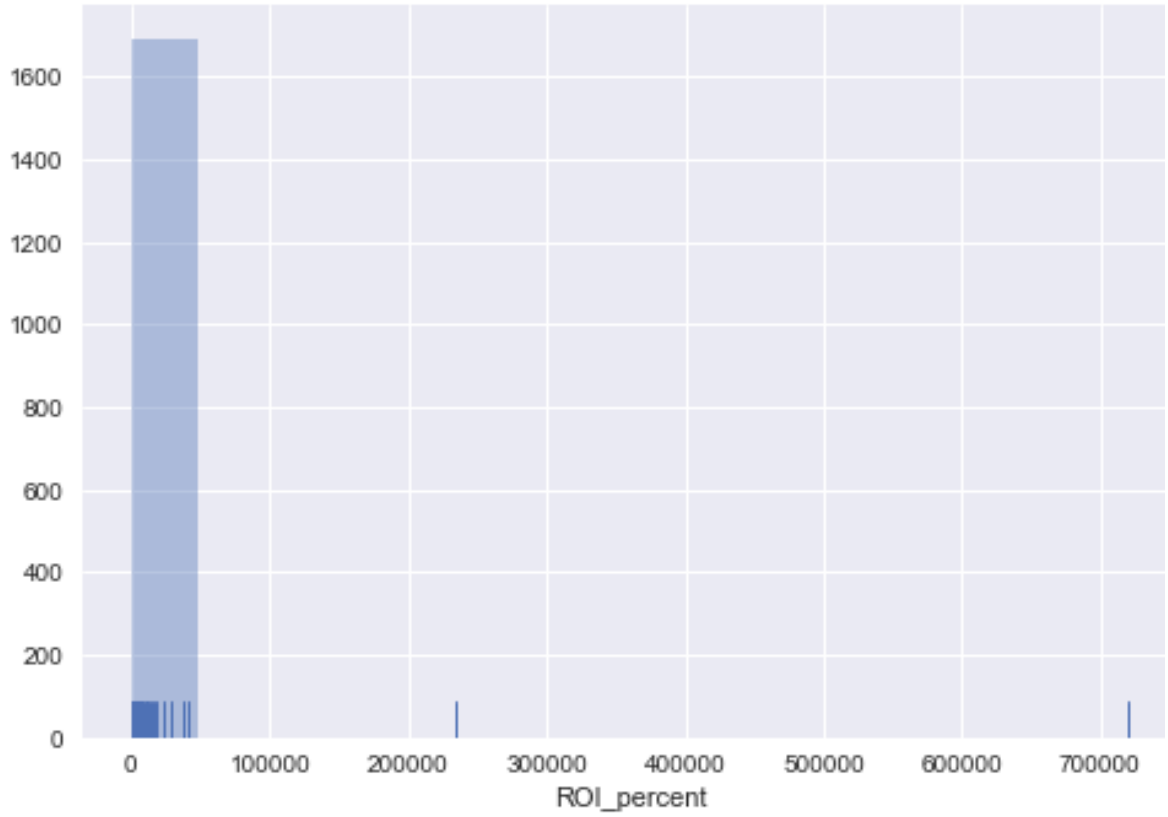
Binary to multi-class

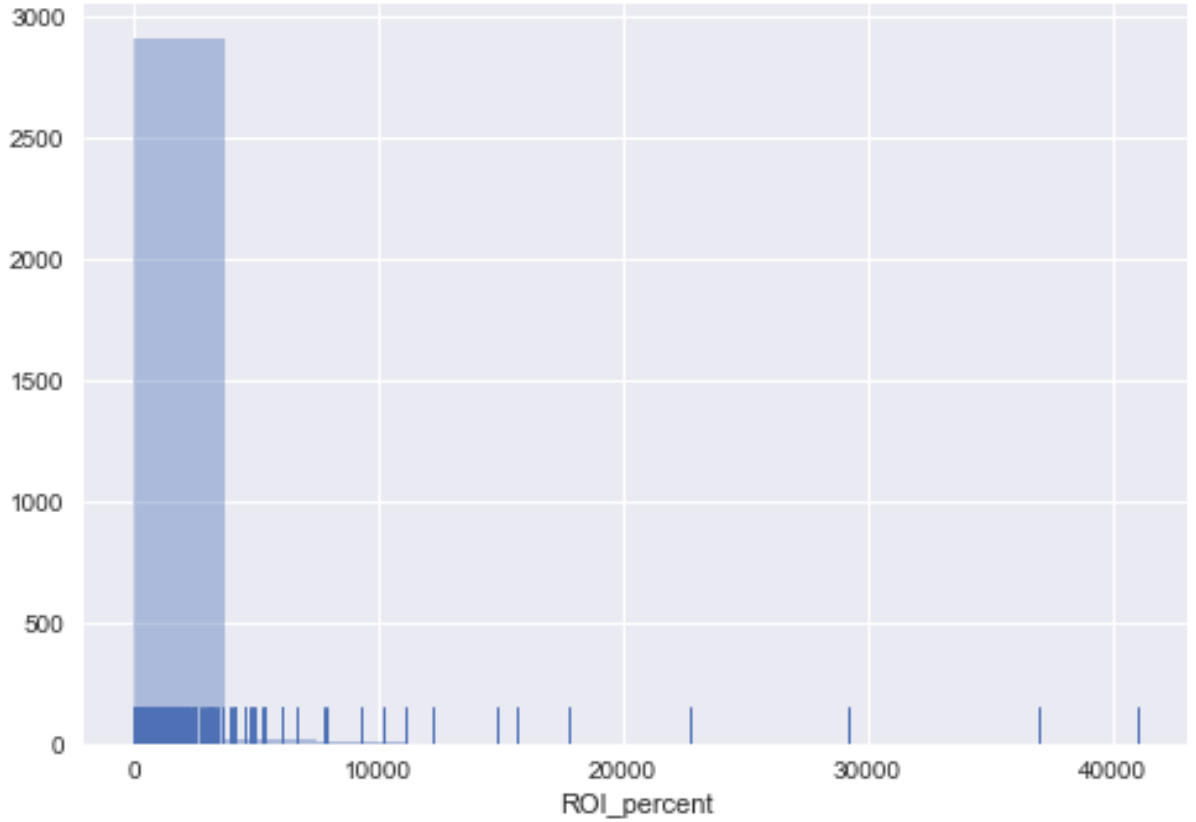
- Making buckets for the ROI



Binary to multi-class

- Weeding out the spectacularly successful indie horror films

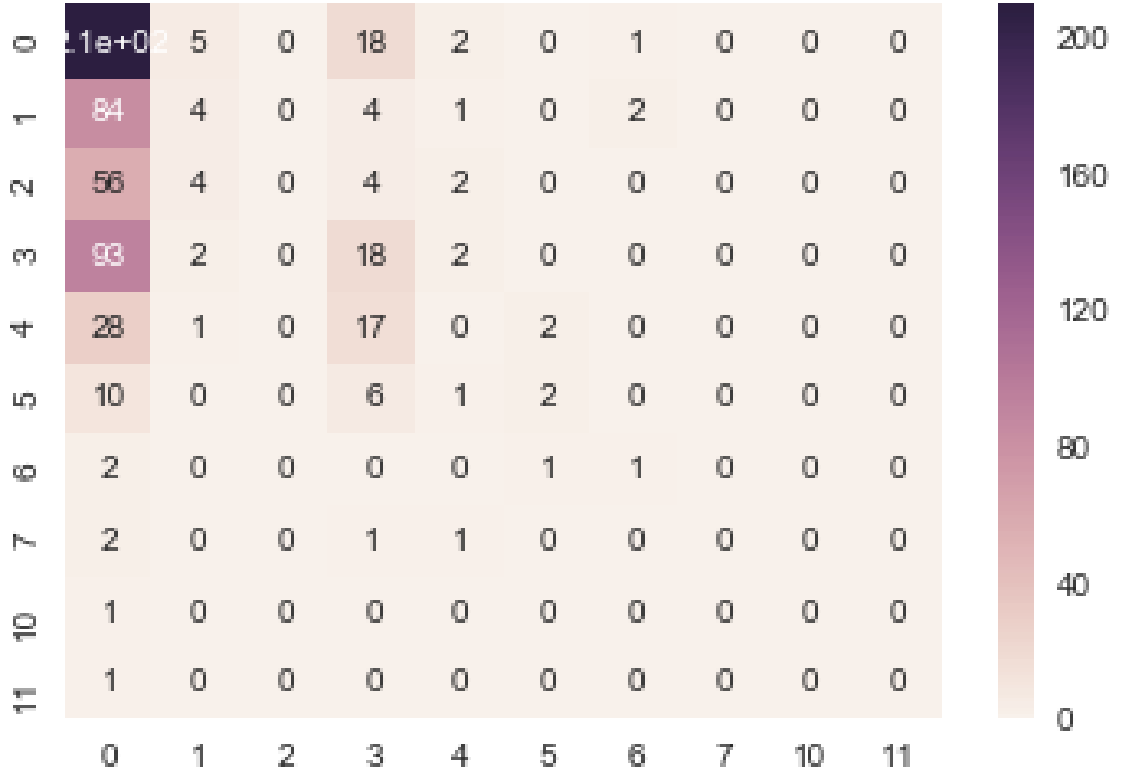




KNN multi-class classifier

- Accuracy: 40%

Making a profit	Making a loss
Precision: 74%	Precision: 43%
Recall: 21%	Recall: 89%

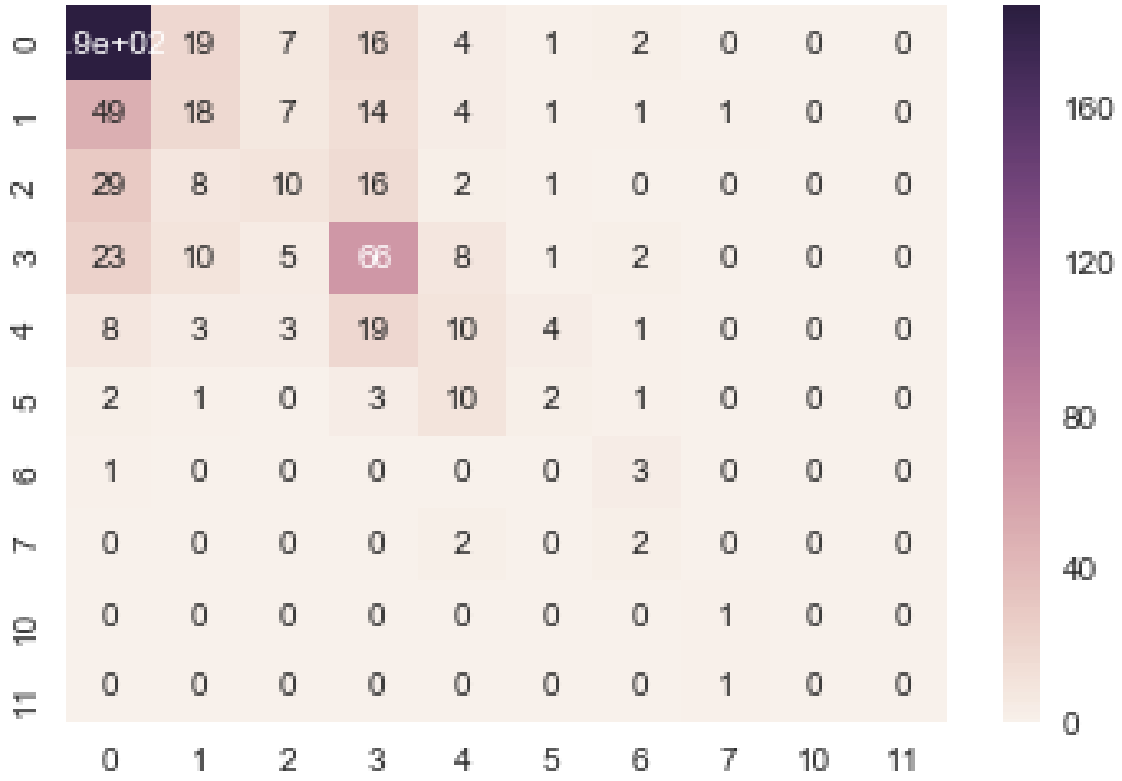


Random Forest multi-class classifier

- Feature significance:
  - Content rating – 0.06
  - Budget – 0.35
  - Duration – 0.23
  - Director – 0.37

- Accuracy: 49%

Making a profit	Making a loss
Precision: 83%	Precision: 62%
Recall: 68%	Recall: 79%



What we would do next

- Get more data – this is now about 3000 films
- Parameter sweeping on the models
- Split out content rating features
- Look in more detail at normalisation – some models got worse!
- 1-away results

But is our accuracy any good?

Binary: 74% vs 60% - i.e. 25% improvement

Multiclass: 49%

- Sharda, Ramesh, and Dursun Delen. "Predicting Box-office Success of Motion Pictures with Neural Networks." *Journal of expert systems and Applications v30 (2006) p243-254*

Neural network classification – average bin accuracy 37%

- Kaggle (logistic regression, SVM, Random Forest) – best average bin accuracy 33%