User movie preferences based on tag relevance, standard movie attributes and controversies

Submission by: ***Team 22*** | Colin Flaherty, Jonathan Yuan, Zijian Li, Abhinav Singh

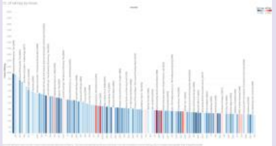
Highlights

- Linear regression model to understand the relationship between movie-tag relevance scores and average rating.
- Further extended our study by eliminating collinearities and interactions among independent variables and observing the coefficient values.
- Derived a controversy score to determine how does various movie attributes effect the controversiality and vice versa.
- Provided extra models to predict user ratings using k means clustering and community detection of user graphs.

Background

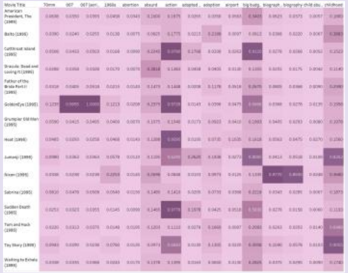
Customer-generated "tags" provide a convenient way of categorizing movies. By examining what tags correlate strongly with customer preferences, we are able to better understand what kinds of movies are preferred by customers.

Building on this, we explore what kinds of movies tend to be "controversial". Controversy within these ratings is an interesting abstraction because it alludes to the polarization in opinion for each movie and points at a different aspect of the distribution of consumer preference.



Data

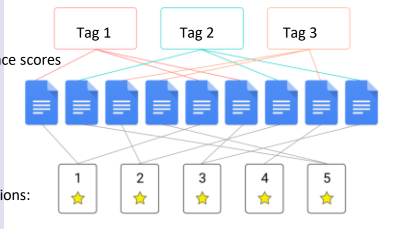
We performed our exploratory analysis on the MovieLens data set along with supporting data for oscar awards and movie attributes. The dataset provides us user ratings for about 50K movies along with relevance scores of some specific customer generated tags.



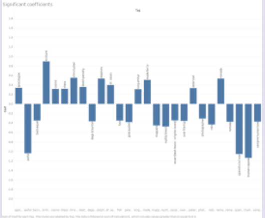
Movie Title	70mm	007	007 (seri..	1960s	abortion	absurd	action	adapted ..	adaption	airport	big budg..	biograph..	biography	child abu..	childhood
American President, The (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Balto (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Cutthroat Island (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Dracula: Dead and Loving It (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Father of the Bride Part II (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GoldenEye (1995)	[illegible]	[illegible]	1.0000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Grumpier Old Men (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Heat (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Jumanji (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Nixon (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Sabrina (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Sudden Death (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Tom and Huck (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Toy Story (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Waiting to Exhale (1995)	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Model

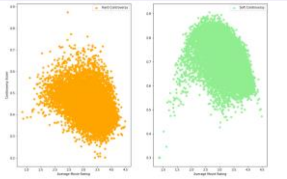
We began with a naive linear model to investigate a relationship between the relevance score between a movie and a tag and the average rating of a movie. This was further extended to removing multicollinearity from the model by removing less significant terms. Finally we correlated this data with a controversy score that we generated based on awards and other attributes.



Results

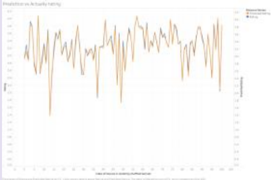


The linear regression model generated had low multicollinearity providing a better understanding of a relationship between movie tags and movie ratings



We were also able to infer that as the average rating of a movie increased, its corresponding hard and soft controversy score decreased. This is evident from the above scatter plot.

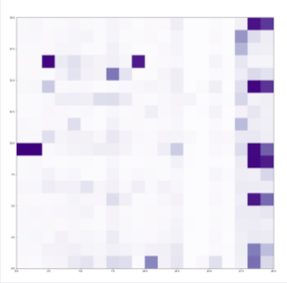
Visualizations



High accuracy prediction of the rating.

MSE_test = 0.0042

Low mean squared error on unseen data.



Correlation heatmap between tags.
Some tags have very correlated relevances.

To conclude our findings, we can say that the we can identify clear relationships between movie attributes and the ratings that users provide. We introduced a concept of controversy which was identified as a very key metric establish relationships with the final user rating.