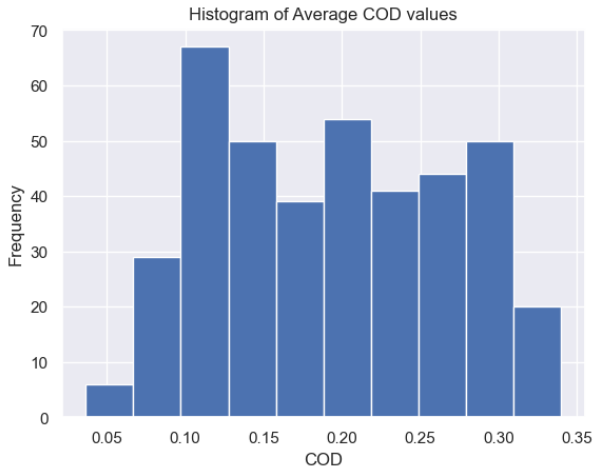
Data Analysis Project 2
Applying machine learning methods to movie ratings data
Written by Mrugank Dake

The dataset is movie rating data for 400 movies. In addition, the responders have also responded to other questions pertaining to sensation-seeking behaviors, personality questions, self-reported movie experience ratings, gender identity, whether they are the only child or had siblings, and whether they tend to enjoy movies alone.

First, the movie ratings were screened for NaN entries. One row had all NaN entries and hence was removed. For other NaN entries, the data was imputed using a 50/50 blend of column and row mean entries, while imputing with the blend, the imputed value was rounded to first decimal place to match the scale of the other ratings. This is a very crude approach but has been used nevertheless as a decent approximation.

Q1. There are a total of 400 movies. Using each movie as a target movie, one of the other 399 movies was used as a predictor and a linear regression model was fit to the data. The model predictions were used to compute coefficient of determination (COD) for the model. This resulted in 399 linear regression models for each movie i.e. a total of 400 * 399 models. The model with the highest COD and the corresponding predictor was the movie that best predicted the rating of the target movie. This was used to determine top 10 that were best predicted using a single predictor and least 10 movies that were most difficult to predict using ratings of another movie. Additionally, an average COD was computed for each movie across all 399 models for the movie.

From the histogram of average COD, we can observe that the models do not perform well at predicting a single movie from another target movie, with on average models explaining 5-35% variance in the data.



Looking at the top 10 models based on COD, we can see that the best COD comes in pairs, wherein the target and predictor are interchangeable. And the highest predicted movie Erik the Viking (1989) or I.Q. (1994) have a very high COD of ~0.73.

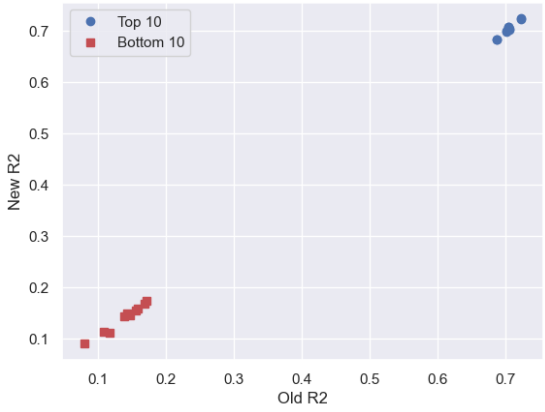
Target movie	best COD	Best Predictor
Erik the Viking (1989)	0.723343	I.Q. (1994)
I.Q. (1994)	0.723343	Erik the Viking (1989)
The Lookout (2007)	0.704884	Patton (1970)
Patton (1970)	0.704884	The Lookout (2007)
Best Laid Plans (1999)	0.703987	The Bandit (1996)
The Bandit (1996)	0.703987	Best Laid Plans (1999)
Congo (1995)	0.701036	The Straight Story (1999)
The Straight Story (1999)	0.701036	Congo (1995)
Heavy Traffic (1973)	0.687072	Ran (1985)
Ran (1985)	0.687072	Heavy Traffic (1973)

Looking at the bottom 10 models based on COD, we can see that the best COD can go quite low up to 0.08. Additionally, the movies now don't seem to occur in pairs. Therefore, the low prediction is only one way but not the other way. This is because the target movies in this case are in general the worst predicted movie. However, their predictors when used as targets might have other movies that can predict them well.

Target movie	best COD	Best Predictor
Avatar (2009)	0.0791126	Indiana Jones and the Kingdom of the Crystal Skull (2008)
Interstellar (2014)	0.108968	Torque (2004)
Black Swan (2010)	0.116818	Once Upon a Time in America (1984)
Clueless (1995)	0.138304	Love Story (1970)
The Cabin in the Woods (2012)	0.142478	The Evil Dead (1981)
La La Land (2016)	0.146454	The Lookout (2007)
Titanic (1997)	0.155495	Cocktail (1988)
13 Going on 30 (2004)	0.15811	Can't Hardly Wait (1998)
The Fast and the Furious (2001)	0.168981	Terminator 3: Rise of the Machines (2003)
Grown Ups 2 (2013)	0.170761	The Core (2003)

Given this data, we can say that some movies can be reliably predicted using a single other movie, but others cannot be. In general, it appears that models using single other movie as predictors are not great at predicting ratings of other movies.

Q2. Now using the 10 best predicted movies from Q1 and using their corresponding predictors in addition to other predictors like gender, sibship status, and social viewing preferences, multiple regression models for each target movie. In order to account for missing data and non-responded entries data in the last three predictors, they were eliminated and then one-hot encoded. The resulting models did not exhibit a difference in COD.



Given that the model fit did not change either for worst predicted or best predicted movies, we can say that gender, sibship status or social viewing preferences of people are not great predictors of how they would rate any single movie. Instead, their ratings for other movies are more reliable predictors.

Q3. Here, the middle 30 movies were selected based on their best COD from Q1 were used as targets, and top 10 predicted movies from Q1 were used as predictors. Ridge regression was used to determine weights for the predictor movies to determine which of the 10 predictor movies were the best predictors for the target movie. Ridge regression penalizes the predictors the weights for the predictors that are not good predictors, thereby providing a better estimate of the top predictors for each movie. First, the data was split into train and test sets. Grid Search cross validation was used on training data to determine the optimal penalization term (α) for each model. This was repeated twice, first with a broader range of α values and the obtained value for each model was then used to fine-tune α on a narrower range. The obtained α is only based on training data obtained via cross-validation and thus reduces the problem of overfitting. Ideally, the predictors should be scaled before performing ridge regression as predictors with high values would be penalized more. Since all the ratings are on Likert scale of 0-5, their scale is identical and hence no normalization was applied to the data. The best-fit α obtained using grid search was then used to fit a ridge regression model on the training data. The model was then evaluated on both training and testing data to obtain root mean square error (RMSE).

The table here reports the target movies as rows and the first 10 columns (0-9) are the beta weights of the top 10 movies used as predictors, followed by intercept, chose α, RMSE for train and RMSE for test. We can see that the models do not appear to overfit the training data, as RMSE for train and test are comparable. Therefore, the obtained α and model parameters are reliable. We can also see that, for most target movies, the chosen model has almost equal weights across the movies. Therefore, it appears that having more than one predictor can provide a better prediction of target movie.

	0	1	2	3	4	5	6	7	8	9	Intercept	alpha	RMSE train	RMSE test
Twister (1996)	0.06	0.08	0.13	0.1	0.13	0.09	0.08	0.05	0.09	0.1	0.32	81.53	0.32	0.39
Aliens (1986)	0.1	0.11	0.13	0.18	0.07	0.09	0.06	0.11	0.08	0.1	0.20	77.01	0.35	0.47
Austin Powers in Goldmember (2002)	0.1	0.05	0.19	0.26	0.06	0.1	-0.01	0.04	0.19	0.16	-0.28	27.26	0.51	0.60
Austin Powers: The Spy Who Shagged Me (1999)	0.12	0.23	0.1	0.2	0.17	0.11	0.07	0.12	0.11	-0.03	-0.35	45.35	0.56	0.47
Gone in Sixty Seconds (2000)	0.04	0.02	0.12	0.1	0.1	0.13	0.06	0.05	0.06	0.05	0.69	141.46	0.33	0.37
28 Days Later (2002)	0.09	0.0	0.09	0.18	0.07	0.14	0.05	0.08	0.12	0.1	0.41	65.70	0.35	0.37
The Big Lebowski (1998)	0.07	0.02	0.17	0.2	0.2	0.01	-0.05	0.04	0.15	0.2	0.37	29.52	0.36	0.29
Blues Brothers 2000 (1998)	0.08	0.06	0.09	0.05	0.06	0.12	0.07	0.09	0.07	0.09	0.67	150.50	0.36	0.37
Goodfellas (1990)	0.12	0.17	0.04	0.18	0.04	0.08	0.1	0.04	0.15	0.09	0.44	51.01	0.34	0.38
Dances with Wolves (1990)	0.08	0.1	0.07	0.08	0.09	0.1	0.04	0.07	0.12	0.08	0.59	103.02	0.34	0.42
The Green Mile (1999)	0.09	0.01	0.11	0.15	0.04	0.09	0.03	0.13	0.13	0.12	0.68	54.40	0.31	0.44
The Blue Lagoon (1980)	0.1	0.07	0.27	0.12	0.01	0.11	0.05	0.17	0.02	0.03	0.28	30.65	0.34	0.33
Uptown Girls (2003)	0.16	0.15	0.07	0.15	0.14	0.12	0.04	0.17	0.05	0.06	0.03	41.96	0.40	0.34
The Machinist (2004)	0.07	0.09	0.06	0.1	0.12	0.01	0.02	0.13	0.13	0.15	0.45	65.70	0.34	0.31
Knight and Day (2010)	0.14	0.14	0.07	0.07	0.19	0.23	-0.06	0.14	-0.01	0.03	0.30	37.44	0.35	0.41
The Evil Dead (1981)	0.16	0.22	0.1	0.14	0.15	0.03	0.04	0.03	-0.05	0.15	0.23	25.00	0.33	0.34
Men in Black (1997)	0.17	0.07	0.19	0.16	0.1	0.1	0.03	0.11	0.1	0.03	0.28	55.53	0.51	0.54
Men in Black II (2002)	0.13	0.06	0.12	0.13	0.12	0.07	0.1	0.12	0.12	0.14	0.05	95.10	0.50	0.56
Equilibrium (2002)	0.08	0.09	0.03	0.02	0.08	0.03	0.07	0.09	0.11	0.1	0.86	83.79	0.32	0.29
The Good the Bad and the Ugly (1966)	0.07	0.11	0.09	0.07	0.11	0.12	0.04	0.11	0.07	0.09	0.67	121.11	0.36	0.41
The Rock (1996)	0.15	0.06	0.14	0.07	0.18	0.12	-0.01	0.01	0.09	0.02	0.66	51.01	0.35	0.29
Let the Right One In (2008)	0.04	0.05	0.08	0.12	0.06	0.09	0.1	0.06	0.11	0.1	0.60	84.92	0.32	0.25
You're Next (2011)	0.07	0.07	0.16	0.16	0.09	0.16	0.0	0.11	0.04	0.09	0.13	47.61	0.31	0.35
Reservoir Dogs (1992)	0.07	0.01	0.05	0.22	0.08	0.06	0.07	0.21	0.09	0.18	0.27	35.18	0.31	0.40
The Poseidon Adventure (1972)	0.06	0.07	0.04	0.05	0.07	0.07	0.08	0.06	0.07	0.06	0.90	233.04	0.34	0.32
The Prestige (2006)	0.16	0.16	0.04	-0.01	0.12	0.08	0.11	0.14	0.12	-0.01	0.73	54.40	0.34	0.34
There's Something About Mary (1998)	0.08	0.16	0.06	0.19	0.09	0.07	0.16	0.11	0.07	0.05	0.07	61.18	0.35	0.38
The Mummy Returns (2001)	0.06	0.08	0.09	0.1	0.08	0.09	0.09	0.07	0.1	0.08	0.56	179.90	0.47	0.52
The Mummy (1999)	0.13	0.1	0.12	0.05	0.08	0.03	0.1	0.11	0.13	0.11	0.37	83.79	0.51	0.62
Just Married (2003)	0.17	0.03	0.08	0.03	0.08	0.1	-0.06	0.27	0.14	0.15	0.10	28.39	0.35	0.44

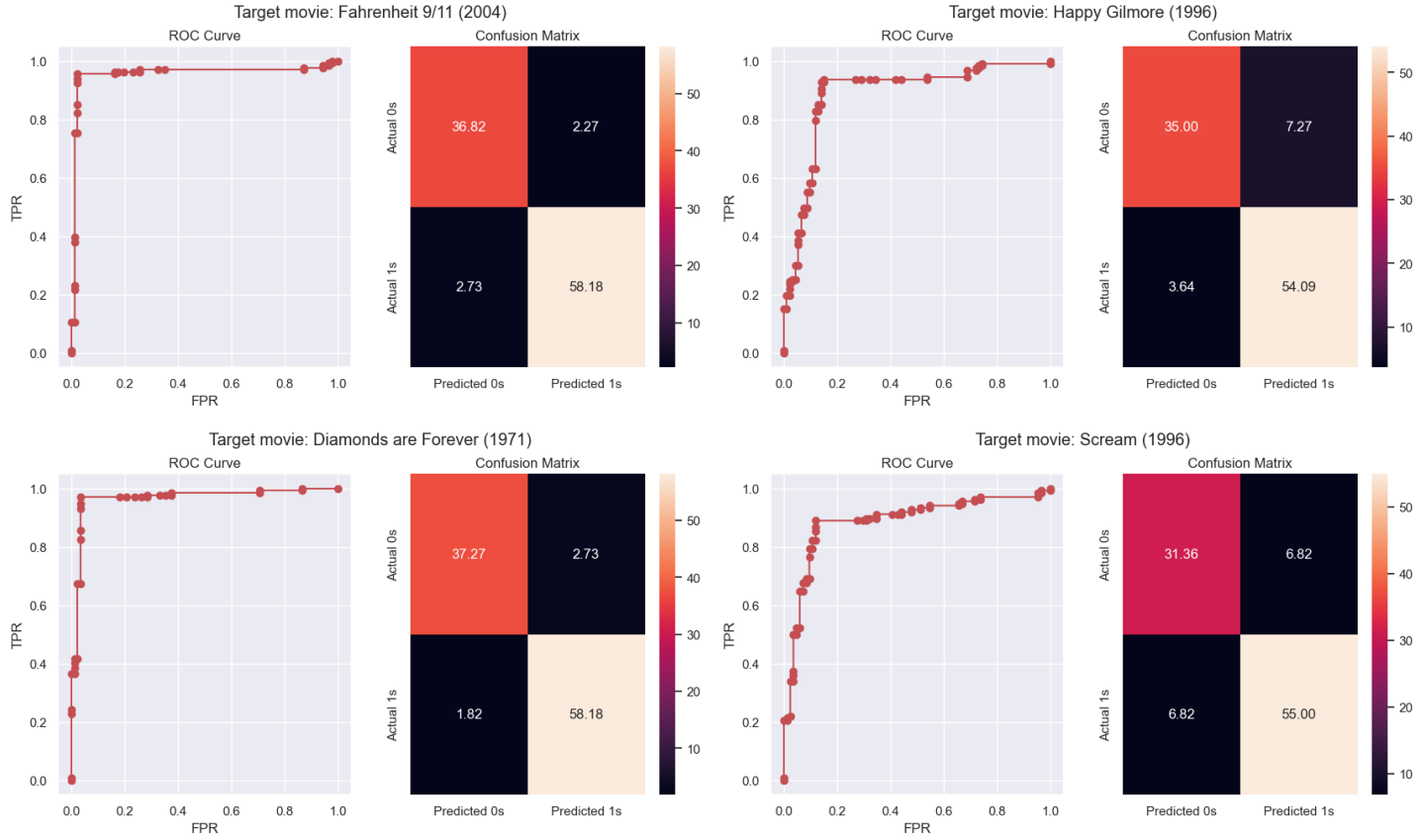
Q4. This is similar to Q3, however, here the regression models are penalized using L1 instead of L2 norm. Lasso regression was used to determine weights for the predictor movies to determine which of the 10 predictor movies were the best predictors for the target movie. Lasso regression penalizes the predictors the weights for the predictors that are not good predictors, similar to ridge regression, however, in the case of Lasso, the predictor weights can end up being 0. First, the data was split into train and test sets. Grid Search cross validation was used on training data to determine the optimal penalization term (α) for each model. This was repeated twice, first with a broader range of α values and the obtained value for each model was then used to fine-tune α on a narrower range. The obtained α is only based on training data obtained via cross-validation and thus reduces the problem of overfitting. Ideally, the predictors should be scaled before performing lasso regression as predictors with high values would be penalized more. Since all the ratings are on Likert scale of 0-5, their scale is identical and hence no normalization was applied to the data. The best-fit α obtained using grid search was then used to fit a lasso regression model on the training data. The model was then evaluated on both training and testing data to obtain root mean square error (RMSE).

The table here reports the target movies as rows and the first 10 columns (0-9) are the beta weights of the top 10 movies used as predictors, followed by intercept, chose α, RMSE for train and RMSE for test. We can see that the models do not appear to overfit the training data, as RMSE for train and test are comparable. Therefore, the obtained α and model parameters are reliable. Unlike ridge regression, we can see that some of the movies have very low weights, while some movies have very high weights. Therefore, lasso regression seems to be more suitable at choosing best predictors in this case.

	0	1	2	3	4	5	6	7	8	9	Intercept	alpha	RMSE train	RMSE test
Twister (1996)	0.0	0.09	0.28	0.06	0.26	0.02	0.09	0.0	0.06	0.1	0.22	5.477387e-03	0.32	0.39
Aliens (1986)	0.09	0.13	0.12	0.4	0.02	0.04	0.0	0.16	0.01	0.06	0.12	6.969849e-03	0.35	0.49
Austin Powers in Goldmember (2002)	0.1	0.0	0.18	0.37	0.0	0.09	0.0	0.0	0.29	0.12	-0.23	6.969849e-03	0.51	0.61
Austin Powers: The Spy Who Shagged Me (1999)	0.0	0.42	0.0	0.35	0.27	0.03	0.0	0.12	0.0	0.0	-0.29	9.457286e-03	0.56	0.46
Gone in Sixty Seconds (2000)	0.0	0.0	0.17	0.14	0.04	0.39	0.0	0.0	0.0	0.0	0.66	1.542714e-02	0.32	0.39
28 Days Later (2002)	0.15	-0.29	-0.19	0.52	0.02	0.36	-0.09	0.11	0.38	0.0	0.20	1.000000e-14	0.34	0.43
The Big Lebowski (1998)	0.0	0.0	0.12	0.27	0.27	0.0	-0.0	0.0	0.1	0.24	0.39	5.477387e-03	0.36	0.29
Blues Brothers 2000 (1998)	0.11	0.0	0.08	0.0	0.0	0.38	0.0	0.08	0.0	0.18	0.52	9.954774e-03	0.36	0.38
Goodfellas (1990)	0.09	0.26	0.0	0.32	0.0	0.03	0.05	0.0	0.27	0.01	0.40	5.477387e-03	0.34	0.37
Dances with Wolves (1990)	0.02	0.11	0.0	0.07	0.05	0.21	0.0	0.0	0.34	0.03	0.57	9.457286e-03	0.33	0.44
The Green Mile (1999)	0.03	-0.0	0.06	0.25	0.0	0.06	0.0	0.2	0.22	0.1	0.63	5.477387e-03	0.30	0.45
The Blue Lagoon (1980)	0.1	0.05	0.53	0.0	0.0	0.03	0.0	0.26	0.0	0.0	0.23	4.482412e-03	0.34	0.34
Uptown Girls (2003)	0.18	0.14	0.0	0.27	0.19	0.08	-0.0	0.29	0.0	0.0	-0.05	2.492462e-03	0.40	0.34
The Machinist (2004)	0.0	0.05	0.0	0.1	0.17	-0.0	-0.0	0.21	0.2	0.18	0.39	6.472362e-03	0.34	0.33
Knight and Day (2010)	0.13	0.18	0.0	0.06	0.23	0.44	-0.25	0.21	-0.06	0.03	0.20	1.497487e-03	0.34	0.41
The Evil Dead (1981)	0.17	0.4	0.14	0.17	0.26	-0.06	0.02	-0.04	-0.38	0.3	0.18	1.000000e-14	0.32	0.33
Men in Black (1997)	0.34	0.0	0.37	0.15	0.04	0.0	0.0	0.09	0.01	0.0	0.39	1.343719e-02	0.51	0.54
Men in Black II (2002)	0.17	0.0	0.05	0.22	0.16	0.0	0.01	0.19	0.1	0.19	0.05	1.343719e-02	0.50	0.57
Equilibrium (2002)	0.06	0.1	0.0	-0.0	0.09	0.0	0.0	0.14	0.25	0.1	0.76	4.482412e-03	0.31	0.30
The Good the Bad and the Ugly (1966)	0.0	0.18	0.08	0.0	0.16	0.23	0.0	0.15	0.0	0.12	0.54	8.462312e-03	0.35	0.41
The Rock (1996)	0.25	0.0	0.24	0.0	0.27	0.06	-0.0	0.0	0.0	0.0	0.65	5.974874e-03	0.34	0.29
Let the Right One In (2008)	0.0	0.0	0.0	0.22	0.0	0.17	0.14	0.0	0.25	0.05	0.56	6.969849e-03	0.32	0.27
You're Next (2011)	0.03	0.06	0.21	0.25	0.05	0.27	-0.08	0.16	-0.0	0.07	-0.03	1.497487e-03	0.31	0.34
Reservoir Dogs (1992)	0.0	0.0	0.0	0.37	0.09	0.0	0.0	0.37	0.0	0.23	0.23	4.979899e-03	0.31	0.42
The Poseidon Adventure (1972)	0.0	0.1	0.0	0.0	0.13	0.03	0.16	0.02	0.12	0.02	1.03	2.139698e-02	0.34	0.34
The Prestige (2006)	0.22	0.2	0.0	-0.0	0.15	0.0	0.06	0.2	0.1	-0.0	0.70	5.477387e-03	0.34	0.34
There's Something About Mary (1998)	0.0	0.27	0.0	0.39	0.08	0.0	0.24	0.09	0.0	0.0	0.00	4.979899e-03	0.35	0.38
The Mummy Returns (2001)	0.0	0.05	0.07	0.19	0.03	0.16	0.09	0.0	0.27	0.01	0.49	1.592462e-02	0.47	0.53
The Mummy (1999)	0.24	0.04	0.21	-0.0	0.0	-0.0	0.06	0.13	0.28	0.05	0.22	4.979899e-03	0.51	0.63
Just Married (2003)	0.2	-0.0	0.05	0.0	0.05	0.07	-0.19	0.48	0.19	0.15	0.02	1.994975e-03	0.34	0.45

Q5. First, the average movie enjoyment for each user on non-imputed data by averaging the rating for all the movies by each user. Next, the movies were sorted in the increasing order of rating and the middle 4 movies were selected as target movies. For each movie, a logistic regression model with 10-fold cross validation was

fitted on the training data once the data was split into train and test set. The resulting model obtained was then evaluated on the test data set.



	AUC test	beta	intercept
Fahrenheit 9/11 (2004)	0.956612	10.514312	-30.458468
Happy Gilmore (1996)	0.887393	5.865999	-16.899860
Diamonds are Forever (1971)	0.966598	8.070634	-23.324315
Scream (1996)	0.887824	5.047525	-14.479993

We can see from the model fit and the confusion matrix obtained on the test dataset, that the model does fit well. The AUC score also confirms the same.

Hence, we can conclude that people's enjoyment of movies is a good predictor of whether they would like any given movie.