Allison Redfern - DS1001 Data Analysis Project 2

1: Finding Movie Rating Predictors Using Simple Linear Regression

The data was cleansed by removing any rows that were missing movie ratings for all movies. Then, I replaced the remaining missing data with a blend (50/50) of the user's and movie's average rating. This imputed data was used to perform 399 simple linear regressions for each of the 400 movies. The predictor movie that produced the highest Coefficient of Determination (COD) was considered the best predictor. By performing a simple linear regression, we are assuming there is a linear relationship between how users rate two different movies, and that the residuals are normally distributed, independent of the predicted movie, and have similar variances.

A simple linear regression was chosen to determine the best predictor because we are only using one variable for the prediction. We use the highest COD to determine the best predictor because we are looking to maximize the proportion of the variance that is accounted for by the model.

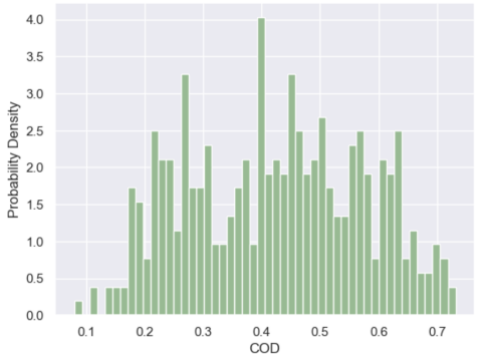

Figure 1: Histogram of COD for 400 movies predicted by their top predictor

A histogram showing the COD for each movie and its top predictor is shown in Figure 1. The average COD across all movies and all top predictors was 0.424. The 10 movies with the highest COD and the bottom 10 movies with the lowest COD, their best predictor movie, and their associated COD value are shown in Table 1 and 2.

Table 1: Top predictor movie and COD for top predicted movies

Movie	Top Predictor Movie	COD
I.Q. (1994)	Erik the Viking (1989)	0.732
Erik the Viking (1989)	I.Q. (1994)	0.732
Patton (1970)	The Lookout (2007)	0.714
The Lookout (2007)	Patton (1970)	0.714
The Bandit (1996)	Best Laid Plans (1999)	0.711
Best Laid Plans (1999)	The Bandit (1996)	0.711
Congo (1995)	The Straight Story (1999)	0.701
The Straight Story (1999)	Congo (1995)	0.701
The Final Conflict (1981)	The Lookout (2007)	0.700
Heavy Traffic (1973)	Ran (1985)	0.693

Table 2: Top predictor movie and COD for bottom predicted movies

Movie	Top Predictor Movie	COD
Avatar (2009)	Bad Boys (1995)	0.079
Interstellar (2014)	Torque (2004)	0.111
Black Swan (2010)	Sorority Boys (2002)	0.117
Clueless (1995)	Escape from LA (1996)	0.141
The Cabin in the Woods (2012)	The Evil Dead (1981)	0.144
La La Land (2016)	The Lookout (2007)	0.149
Titanic (1997)	Cocktail (1988)	0.154
13 Going on 30 (2004)	Can't Hardly Wait (1998)	0.160
The Fast and the Furious (2001)	Terminator 3: Rise of the Machines (2003)	0.169
Grown Ups 2 (2013)	The Core (2003)	0.171

Given the COD, even the best model only accounts for 73% of the variance. Simple linear regression with just a single predictor column is limited in providing an accurate estimate. Another limitation is the fact that so much of this data was imputed. Using a 50/50 blend of user and movie average ratings can be problematic for movies with very little data, since we don't have an accurate representation of users' ratings on those movies to predict others.

2: Predicting Movie Ratings Using Multiple Linear Regression

The data set for this analysis was cleansed by removing any rows in the imputed data that had missing data or "-1" (Did Not Respond) responses in the gender identity, sibling status, and social viewing preferences column. Since the gender identity column has three options (Male, Female, and Self Described), two "dummy" columns were created its place. A female response is represented by 1 and 0 in the first and second column, respectively. A male response is represented by 0 and 1 in the first and second column, respectively. A self-described response is represented by 0 and 0 in the first and second column, respectively. This "dummy" column approach did not need to be taken for the sibling status and social viewing preferences column since those already contain binary entries (after the "Did Not Responds" were removed).

After cleansing the data, multiple linear regression was performed for each of the top and bottom movies from Question 1. The predictors for each movie consisted of each of their top predictor movie's ratings, their gender identity, sibling status, and social viewing preferences. The coefficients for each predictor and the COD for the model were then determined. Like the simple linear regression, by performing a multiple linear regression we are assuming a linear relationship between movie ratings and each of the predictors that and that the residuals are normally distributed, independent of the predicted movie, and have similar variances. We also assume that the predictors (predictor movie, gender identity, sibling status, and social viewing preferences) are independent of each other. Multiple linear regression was chosen because we are trying to predict a movies rating given multiple predictors and to what extent each predictor contributes to the prediction (as shown by the coefficients). The coefficients for each model's predictors can be found in the Jupyter notebook.

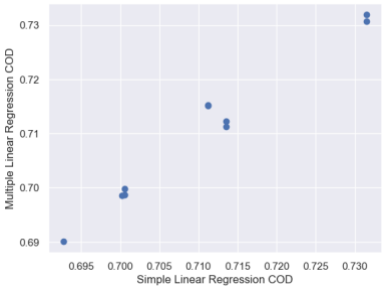

Figure 2: Old vs new COD for top predicted movies

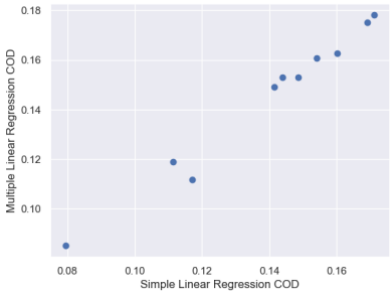

Figure 3: old vs new COD for bottom predicted movies

The COD across all 20 movies only changed by less than a percent as shown in Figures 2 and 3. The largest increase in COD was for the movie, "The Cabin in the Woods (2012)", by only 0.009 and the largest decrease in COD was for the movie, "I.Q. (1994)", by -0.008. In total 3/10 COD values increased from the top movies group and 9/10 COD values increased from the bottom movies group. So, we can conclude that using gender identity, sibling status, and social viewing preferences along with each movie's predictor movies does not provide much better of an estimate for top predicted movies and provides a slightly better estimate for bottom predicted movies. If the COD had increased a lot, one concern with multiple linear regression could be overfitting our particular data set. This would be solved by using cross-validation which will be used in the remaining parts.

3: Predicting Movie Ratings Using Ridge Regression

For the 30 movies to be predicted, I selected the exact middle 30 movies in the data set sorted by COD from Question 1. For the 10 predictor movies, I selected the following 10 movies in the sorted data set. This selection can be easily modified to replicate this analysis for other sets of 30 and 10 movies. The data was broken up using an 80/20 train/test split (using a random seed of 5) to avoid overfitting. For hyperparameter tuning, I used RidgeCV with built in cross-validation (to avoid overfitting) over a wide range of alphas, and selected the optimal alpha for each of the 30 movies. Using this alpha in each model, I found the RMSE and betas of each of the 30 models. (I also repeated the models using one optimal alpha across all movies. Please see Jupyter notebook for details.)

In ridge regression we are assuming that there is a linear relationship between movie ratings and each of the 10 predictor movies and that the residuals are independent of the predicted movie and have similar variances. Ridge regression was chosen as a regularization method to be able to penalize overfitting the data by taking away the impact of correlated predictor variables.

The optimal alphas after hyperparameter tuning and the RMSE are shown in Table 3 (see Jupyter Notebook for titles of movie predictors). Many of the models have very high alpha values, which lowers the contribution of the beta coefficients. Generally, the higher the beta value, the greater the effect that movie has on predicting the ratings of the target movie. In ridge regression, these betas can approach zero, but not become zero.

Given the RMSE, the ridge regression models seem to predict some movies very well and others fairly well. The model may be improved by strategically choosing predictor movies rather than randomly choosing them (through a strategy like lasso regression or an elastic net). Another potential concern again is the fact that a lot of the data is imputed based on averages, which could cause overfitting on this particular data set and its noise for movies with less overall ratings.

Table 3: Ridge regression hyperparameters and RMSE

Movie	Alpha	RMSE
Gone in Sixty Seconds (2000)	39.816	0.059
Crossroads (2002)	36.408	0.156
Austin Powers: The Spy Who Shagged Me (1999)	73.898	0.318
Austin Powers in Goldmember (2002)	60.265	0.240
Goodfellas (1990)	50.041	0.122
The Big Lebowski (1998)	67.082	0.069
Twister (1996)	16.000	0.162
Blues Brothers 2000 (1998)	107.980	0.098
Dances with Wolves (1990)	70.490	0.149
28 Days Later (2002)	67.082	0.104
Knight and Day (2010)	53.449	0.136
The Evil Dead (1981)	67.082	0.092
The Machinist (2004)	87.531	0.079
The Blue Lagoon (1980)	46.633	0.119
Uptown Girls (2003)	43.224	0.113
Men in Black II (2002)	33.000	0.199
Men in Black (1997)	94.347	0.239
The Green Mile (1999)	53.449	0.115
The Rock (1996)	60.265	0.130
You're Next (2011)"	50.041	0.141
The Poseidon Adventure (1972)	94.347	0.082
The Good the Bad and the Ugly (1966)	67.082	0.206
Let the Right One In (2008)	43.224	0.072
Equilibrium (2002)	56.857	0.147
Just Married (2003)	56.857	0.170
The Mummy Returns (2001)	67.082	0.173
The Mummy (1999)	63.673	0.175
Reservoir Dogs (1992)	46.633	0.097
Man on Fire (2004)	36.408	0.144
The Prestige (2006)	84.122	0.145

Table 4: Lasso regression hyperparameters and RMSE

Movie	Alpha	RMSE
Gone in Sixty Seconds (2000)	0.0054	0.061
Crossroads (2002)	0.0039	0.157
Austin Powers: The Spy Who Shagged Me (1999)	0.0045	0.315
Austin Powers in Goldmember (2002)	0.0070	0.242
Goodfellas (1990)	0.0009	0.122
The Big Lebowski (1998)	0.0041	0.068
Twister (1996)	0.0029	0.165
Blues Brothers 2000 (1998)	0.0052	0.099
Dances with Wolves (1990)	0.0052	0.148
28 Days Later (2002)	0.0035	0.103
Knight and Day (2010)	0.0047	0.138
The Evil Dead (1981)	0.0025	0.093
The Machinist (2004)	0.0054	0.085
The Blue Lagoon (1980)	0.0043	0.121
Uptown Girls (2003)	0.0019	0.113
Men in Black II (2002)	0.0039	0.201
Men in Black (1997)	0.0031	0.241
The Green Mile (1999)	0.0060	0.115
The Rock (1996)	0.0019	0.132
You're Next (2011)"	0.0043	0.142
The Poseidon Adventure (1972)	0.0003	0.083
The Good the Bad and the Ugly (1966)	0.0066	0.210
Let the Right One In (2008)	0.0023	0.075
Equilibrium (2002)	0.0025	0.147
Just Married (2003)	0.0052	0.169
The Mummy Returns (2001)	0.0062	0.171
The Mummy (1999)	0.0064	0.173
Reservoir Dogs (1992)	0.0078	0.097
Man on Fire (2004)	0.0060	0.146
The Prestige (2006)	0.0015	0.143

4: Predicting Movie Ratings Using Lasso Regression

For the lasso regression, the same target movies, predictor movies, and 80/20 train/test split as Question 3 were used. For hyperparameter tuning, I used LassoCV with built in cross-validation (to avoid overfitting) over a wide range of alphas, and selected the optimal alpha for each of the 30 movies. Using this alpha in each model, I found the RMSE and betas of each of the 30 models. (I also repeated the models using one optimal alpha across all movies. Please see Jupyter notebook for details.)

In lasso regression, we are again assuming that a linear relationship between movie ratings and each of the 10 predictor movies and that the residuals are independent of the predicted movie and have similar variances.

Lasso regression was used in this question for regularization to be able to penalize overfitting the data by taking away the impact of correlated predictor variables. The main difference between lasso regression and ridge regression is that in lasso regression we can fully remove predictors effect (their beta can be 0) rather than only approaching 0, to reduce the overall number of predictors.

The optimal alphas after hyperparameter tuning and the RMSE are shown in Table 4 (see Jupyter Notebook for titles of movie predictors). Many of the models have very low alpha values, which increases the contribution of the beta coefficients. Generally, the higher the beta value, the greater the effect that movie has on predicting the ratings of the target movie. There are many betas in this model that become 0, a unique feature of lasso regression.

Compared to the ridge regression, the lasso regression worked comparably well. Overall, the RMSE decreased for 10 of the 30 movies using lasso regression, so with these given predictor movies, both models were almost identical in performance with the ridge regression being slightly more accurate for this data set. Once again, a potential concern with these models is the imputed data which may cause potential overfitting.

5: Predicting Movie Enjoyment Using Logistic Regression

The average movie enjoyment per user (X) was found using the original data set. After sorting all movies by their average ratings, I found the 4 movies exactly in the middle of the set. I categorized each rating within the movie in the imputed data as 1 if it was above the movie's median rating and 0 if it was below to represent whether users enjoyed the movie or not (Y). I split this new data set using an 80/20 train/test split (using a random seed of 5) to avoid overfitting. Then on the test data, I fit logistic regression models for the 4 movies' enjoyment using the average ratings per user as the predictor and found the betas. Then, I found the area under the ROC curve (AUC values) on the test data.

Logistic regression was chosen because we are categorizing the data into binary outcomes using continuous data. When performing logistic regression we are assuming our binary outcomes of enjoyment are legitimate, each user's average movie rating is independent of other users' ratings, and their enjoyment of these 4 movies is independent of other users' enjoyment. Finally logistic regression assumes linearity between the enjoyment and the log odds and that there is a sufficient sample size.

Table 5: Enjoyment model confusion matrices, beta, and AUC Values

Fahrenheit 9/11 (2004)

Beta = 7.214

AUC Value = 0.969

True Label	0	1
0	102	2
1	5	111

Predicted Label

Happy Gilmore (1996)

Beta = 5.097

AUC Value = 0.881

True Label	0	1
0	91	13
1	13	103

Predicted Label

Diamonds are Forever (1971)

Beta = 6.748

AUC Value = 0.955

True Label	0	1
0	101	9
1	1	109

Predicted Label

Scream (1996)

Beta = 4.245

AUC Value = 0.872

True Label	0	1
0	90	14
1	114	102

Predicted Label

The outcomes of the model are shown in the confusion matrices in Table 5. Each ROC curve is shown in Figures 4-7.

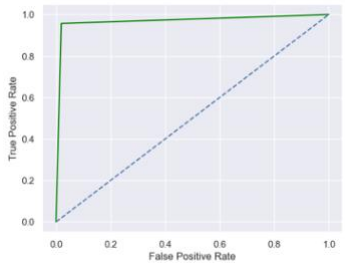

Figure 4: Fahrenheit 9/11 (2004) ROC curve

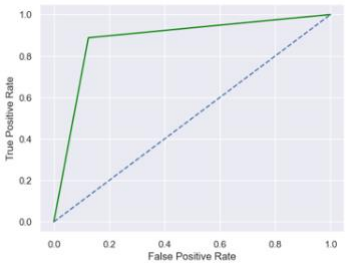

Figure 5: Happy Gilmore (1996) ROC curve

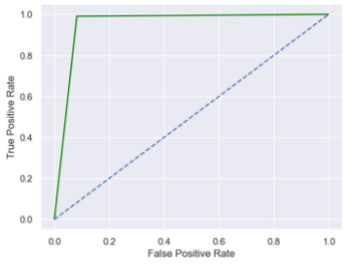

Figure 6: Diamonds are Forever (1971) ROC curve

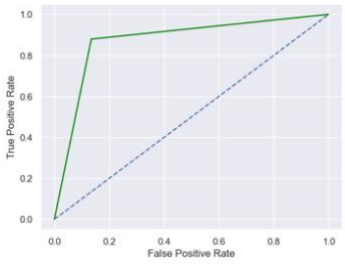

Figure 7: Scream (1996) ROC curve

Judging by the area under the ROC curve, these are highly accurate models for "Fahrenheit 9/11 (2004)" and "Diamonds are Forever (1971)" and fairly accurate models for "Happy Gilmore (1996)" and "Scream (1996)". The main concern for this model would once again be the imputed data. For a user that does not have many movies rated, it may be very easy to predict enjoyment since that would be solely based off the average ratings of only a few movies, and may not be truly representative.

Extra Credit: Predicting movie ratings with sensation seeking behaviors

For the bonus, I was curious on how the sensation seeking behaviors may be used to predict movie ratings. I began by imputing the missing lines of sensation seeking behavior questions using a 50/50 split of average user answer and average question answer. Then, I performed lasso regression using the LassoCV hyperparameter tuning and determined the alpha, betas, and RMSE.

I chose lasso regression because I wanted to allow the betas to become 0 since there were 21 variables for sensation seeking behavior and therefore a lot of potential multi-collinearity. The alphas and RMSE are shown in Table 6.

It turns out that sensation seeking behaviors are not good predictors of movie ratings because the RMSE increased for every movie from the lasso regression in Question 4. However, the lowest increases were for "The Rock (1996)" and "The Good the Bad and the Ugly (1966)" which are interestingly both action movies. Also, many of the betas (which can be viewed in the Jupyter notebook) became zero which makes sense given the multi-collinearity of the questions.

Table 6: Lasso regression using sensation seeking behavior

Movie	Alpha	RMSE
Gone in Sixty Seconds (2000)	0.0156	0.246
Crossroads (2002)	0.0195	0.257
Austin Powers: The Spy Who Shagged Me (1999)	0.0289	0.565
Austin Powers in Goldmember (2002)	0.0242	0.482
Goodfellas (1990)	0.0289	0.233
The Big Lebowski (1998)	200.00	0.267
Twister (1996)	0.0147	0.310
Blues Brothers 2000 (1998)	0.0242	0.283
Dances with Wolves (1990)	0.0039	0.238
28 Days Later (2002)	0.0147	0.292
Knight and Day (2010)	0.0078	0.233
The Evil Dead (1981)	2.8947	0.258
The Machinist (2004)	0.0289	0.225
The Blue Lagoon (1980)	0.0100	0.205
Uptown Girls (2003)	0.0100	0.334
Men in Black II (2002)	0.0156	0.429
Men in Black (1997)	0.0574	0.407
The Green Mile (1999)	0.0156	0.264
The Rock (1996)	200.00	0.170
You're Next (2011)"	0.0242	0.226
The Poseidon Adventure (1972)	0.0313	0.217
The Good the Bad and the Ugly (1966)	200.00	0.248
Let the Right One In (2008)	200.00	0.222
Equilibrium (2002)	0.0384	0.226
Just Married (2003)	0.0147	0.289
The Mummy Returns (2001)	0.0384	0.421
The Mummy (1999)	0.0147	0.393
Reservoir Dogs (1992)	200.000	0.328
Man on Fire (2004)	0.0010	0.401
The Prestige (2006)	0.0015	0.239