Junwen Fang, Lyric Li, Rosy Xu

1 Introduction and Data Processing

This report explores what machine learning methods can do for us. We will apply various supervised machine learning models, including Linear Regression, Ridge regression, Lasso regression and Logistic Regression to predict the movie ratings. Also, we utilized metrics such as R^2, RMSE, and AUC values to assess the quality of our models.

We first process null values in movie ratings by imputation. We imputed movie ratings by taking the average of the user ratings mean and the movie ratings mean. If the user hasn't rated any movies, we instead impute their ratings using movies ratings mean. For all of our questions, we utilized this imputed value when building our model. In regards to null values in users' self-reported questions, we will discuss how we process the null values in the specific question.

2 Baseline Model

2.1 Simple Linear Regression Models

To predict the ratings of each movie, we used the ratings of the other 399 movies in the dataset. We built 399 models for each of the 400 movies by writing nested for loops and updated the **average COD**, **best predictor movie** for each movie. We then drew the histogram of 400 **average COD** values using plt.hist() as shown in Figure 2.1.1. From the graph, we can see that the average COD, which explains on average only about 20% of variance in movie ratings, can be explained by other movie ratings, which is relatively low explainability.

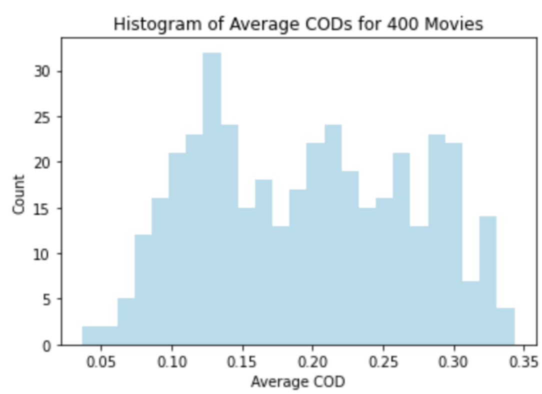

Figure 2.1.1 Histogram of Average CODs for 400 Movies

We assume the movie that predicts ratings the best as the movie that has the **highest COD** with respect to the predicted movie. We also assume the movie that is easiest predicted from the ratings of a single other movie as movie that has the **highest average COD** calculated by averaging the 399 models' COD. Based on these assumptions, we generated the table with 10 movies for highest average COD and lowest average COD respectively using sort and concat, as shown in Table 2.1.2. The highest average COD is about 34%, representing on average 34% of variance in ratings of "Escape from LA (1996)" can be explained by other movie. The lowest average COD is about 3%, representing 3% of variance in ratings of "Avatar (2009)" can be explained by other movie.

2.2 Multiple Regression Models

In this question, we used the best and worst 10 movies generated from Section 2.1. Additionally, we added gender, sibship, and social viewing preference as additional predictors and built multiple regression models. There are only 24 rows that have null values in gender, sibship status, or social viewing preferences, so we simply **dropped** these rows with null values. In this question, we assume we are comparing the old CODs versus the new CODs for 20 movies mentioned in Section 2.1. To find the new average CODs, similar to Section 2.1, we first loop in 20 movies, and for each movie, we used their best predictor in Section 2.1 to predict by running the multiple linear regression model between one of the 20 movies and predictor movies with gender, sibship, and social viewing preferences. Then, we evaluated the model by R^2 and drew a scatter plot by **plt.scatter** function in Figure 2.2.1. Overall, the graph showed little improvements in CODs after adding three features. We found that for 12 out of 20 of predicted movies, adding three more features in linear regression improved their CODs. For the rest of 8 predicted movies, the models' COD decreased only a little bit.

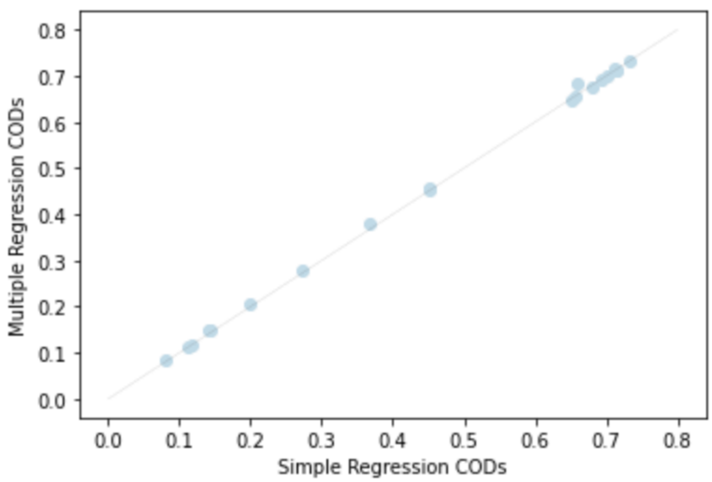

Figure 2.2.1 Scatterplot of Single Regression CODs v.s. Multiple Regression CODs
for Best and Least Well Predicted Movie

3 Linear Regression Model with Regularization

3.1 Ridge Regularization

From the sorted COD obtained in Section 2.1, we selected the middle 30 movies and then randomly chose 10 movies from the rest. For each of the 30 movies, we implemented the following procedure: First, we split the data into an 80/20 train/test ratio. Next, we built a Ridge Regression model using the training data from the 10 movies to predict movie ratings. We employed **RandomizedSearchCV** for hyperparameter tuning. Finally, we extracted and reported the model's alpha, beta, intercept, and RMSE as shown in Table 3.1.1. For different movies, the beta distributions are different. Compared to alpha used in Lasso, the alphas in Ridge Regression models are much larger, ranging from 7 to 108. During parameter search, we first tried a larger range of alpha and at last narrowed it down to [7, 8, … , 107, 108].

3.2 Lasso Regularization

Using the same data as in Section 3.1, we applied Lasso Regression in place of Ridge Regression. The process followed similar steps: initially, we divided the data into an 80/20 train/test ratio. Subsequently, we constructed a Lasso Regression model. For hyperparameter tuning, **RandomizedSearchCV** was utilized. At the conclusion of this process, we extracted and reported the model's alpha, beta, intercept, and RMSE as shown in Table 3.2.1. Compared to alpha using the Ridge Regression model, the alphas for Lasso Regression models are much smaller, even 0. Indeed, Lasso Regression selected specific features by setting other features' beta as 0. Similar as Section 3.1, we narrowed down the beta ranges to [0.0001, 0.0006, …, 0.9995, 1].

4 Logistic Regression

We first sorted the average ratings of all movies and selected four with intermediate ratings. For each user, we calculated the average ratings, and if the user hasn't rated any movies, the average for that user will be zero. For each movie, we encoded their ratings: 1 for ratings **higher than** the median and 0 for ratings **lower than or equal to** the median. We then employed **k-fold cross-validation** to mitigate overfitting and reported the resulting accuracies as shown in Table 4.1.

Predicted Movie	**Accuracy**
Fahrenheit 9/11 (2004)	0.94
Happy Gilmore (1996)	0.88
Diamonds are Forever (1971)	0.95
Scream (1996)	0.85

Table 4.1 Accuracy by K-Fold Cross-Validation

Next, we used train-test split (80:20 for training:testing data) to prevent overfitting. We then build logistic regression models to predict these encoded ratings based on the average score a user gives to all movies. Finally, we evaluated the model's quality by constructing a confusion matrix on predicted testing data and plotting the ROC curve. Our model has high quality due to the large area under the ROC curve.

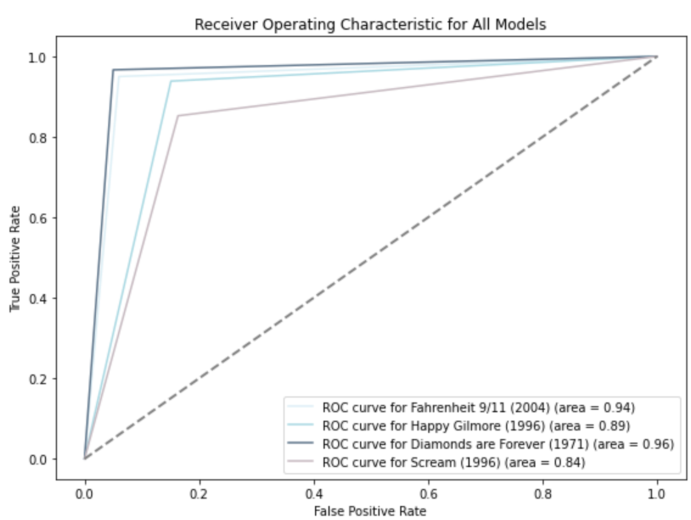

Figure 4.2 ROC Curve for Middle Four Average Movie Ratings

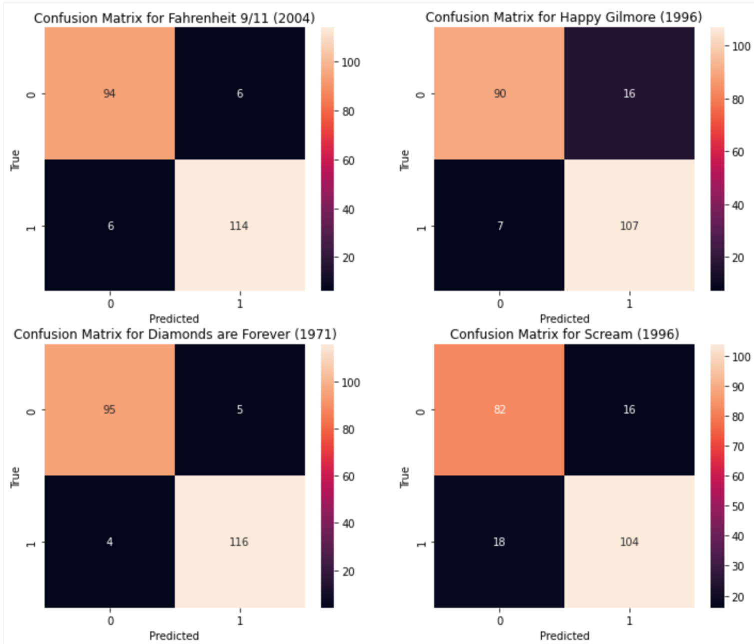

Figure 4.3 Confusion Matrix for Middle Four Average Movie Ratings

5 Extra Credit

In this part, we investigated the features selection for movies that belong to several franchises by building a simple linear regression. Specifically, we will build **linear regression** for movies in franchises ([‘Star Wars’, ‘Harry Potter’, ‘The Matrix’, ‘Indiana Jones’, ‘Jurassic Park’, ‘Pirates of the Caribbean’, ‘Toy Story’, ‘Batman’]). We then built simple linear regression models to predict each movie in the franchise, and compared the average CODs for movie when rating features are from the other movie in the franchise, with features that are ratings from the movie not in the franchise. When there is null rating value, we still impute the rating as the average of the mean of the movie and the mean of the user’s rating. The predicted movie, average COD by movies from franchise, and average COD by movies not from franchise is listed in Table 5.1. As shown in Table 5.1, we found that the average CODs predicted by movies from franchise are much higher than the CODs predicted by those movies not from franchise. Majority of average CODs predicted by franchise are approximating to 0.34, which is the highest average CODs gained from 2.1. As a result, for movies that have a franchise, we conclude that using **movie ratings in franchises** can yield a **higher model explainability** compared to using other movie ratings.

Also, by calculating average COD, it also to some extent shows that the franchise “Harry Potter” has **consistent quality**. Recall in project 1, we concluded that all franchises except “Harry Potter” have inconsistent quality by Kruskal-Wallis H-test. Among all these franchises, “Harry Potter” franchise has the highest average COD if we use the movie in the franchise to build linear regression, meaning that the ratings from other “Harry Potter” franchises are predictable. If the franchise has inconsistent quality, for example, the Batman, building a simple linear regression model by movie in “Batman” franchise will have almost no difference as other movies that are not in “Batman” franchise.

Appendix

Average COD	Predicted Movie	Best Predictor
0.036659	Avatar (2009)	Bad Boys (1995)
0.042864	The Conjuring (2013)	The Exorcist (1973)
0.052908	Interstellar (2014)	Torque (2004)
0.058093	Black Swan (2010)	Sorority Boys (2002)
0.065031	The Cabin in the Woods (2012)	The Evil Dead (1981)
0.065471	Shrek 2 (2004)	Shrek (2001)
0.069699	The Avengers (2012)	Captain America: Civil War (2016)
0.071286	Clueless (1995)	Escape from LA (1996)
0.071967	Pirates of the Caribbean: Dead Man's Chest (2006)	Pirates of the Caribbean: At World's End (2007)
0.075677	Shrek (2001)	Shrek 2 (2004)
0.325636	Heavy Traffic (1973)	Ran (1985)
0.326494	Miller's Crossing (1990)	The Lookout (2007)
0.326806	The Straight Story (1999)	Congo (1995)
0.327349	Patton (1970)	The Lookout (2007)
0.327877	The Bandit (1996)	Best Laid Plans (1999)
0.330112	Crimson Tide (1995)	The Straight Story (1999)
0.334636	Erik the Viking (1989)	I.Q. (1994)
0.339164	The Lookout (2007)	Patton (1970)
0.341764	Sexy Beast (2000)	The Silencers (1966)
0.343171	Escape from LA (1996)	Sexy Beast (2000)

Table 2.1.1 Easiest 10 and Hardest 10 Predicted Movie with Corresponding Average COD and Best Predictor

Predicted Movie	RMSE	Alpha	Beta	Intercept
The Thing (1982)	0.37	89	[0.03, 0.10, 0.12, 0.09, 0.15, 0.06, 0.07, 0.04, 0.11, 0.05]	0.34
12 Monkeys (1995)	0.38	84	[0.10, 0.11, 0.10, 0.04, 0.12, 0.04, 0.05, 0.00, 0.08, 0.05]	0.82
Bad Boys (1995)	0.39	85	[0.08, 0.05, 0.12, 0.03, 0.04, 0.06, 0.04, 0.03, 0.12, 0.16]	0.88
Armageddon (1998)	0.40	52	[0.03, 0.12, 0.13, 0.04, 0.07, 0.04, 0.17, 0.09, 0.15, 0.03]	0.38
Bad Boys 2 (2003)	0.42	59	[0.07, 0.10, 0.06, 0.03, 0.05, 0.06, 0.18, 0.05, 0.09, 0.15]	0.59
The Mist (2007)	0.37	30	[0.10, 0.03, 0.10, 0.01, 0.03, 0.03, 0.20, 0.00, 0.20, 0.14]	0.35
Braveheart (1995)	0.40	68	[0.11, 0.10, 0.07, 0.07, 0.06, 0.05, 0.14, 0.03, 0.10, 0.04]	0.74
The Others (2001)	0.37	42	[0.03, 0.03, 0.11, 0.05, 0.06, 0.04, 0.17, 0.07, 0.21, 0.01]	0.62
Baby Geniuses (1999)	0.41	75	[-0.01, 0.19, 0.11, 0.04, 0.12, 0.06, 0.05, 0.06, 0.13, 0.06]	0.27
One Flew Over the Cuckoo's Nest (1975)	0.42	72	[0.07, 0.08, 0.06, 0.07, 0.11, 0.04, 0.10, 0.01, 0.15, 0.09]	0.90
Honey (2003)	0.40	7	[0.03, 0.19, 0.21, 0.05, 0.08, -0.01, -0.07, 0.06, 0.34, 0.05]	0.11
Big Daddy (1999)	0.29	54	[0.14, 0.14, 0.11, -0.00, 0.04, 0.10, 0.17, 0.05, 0.08, 0.09]	0.44
Halloween (1978)	0.42	84	[0.09, 0.06, 0.10, 0.08, 0.06, 0.07, 0.09, 0.03, 0.13, 0.10]	0.56
The Poseidon Adventure (1972)	0.44	60	[0.06, 0.11, 0.14, 0.05, 0.06, 0.03, 0.11, 0.04, 0.13, 0.01]	0.49
Rocky (1976)	0.49	43	[0.12, 0.06, 0.08, 0.10, 0.14, 0.02, 0.12, 0.00, 0.05, 0.32]	0.15
The Intouchables (2011)	0.42	43	[0.14, 0.20, 0.08, 0.06, 0.06, 0.07, 0.02, 0.05, 0.09, 0.04]	0.75
Aliens (1986)	0.32	31	[0.02, 0.06, 0.12, 0.05, 0.07, 0.05, 0.12, 0.03, 0.22, 0.20]	0.22
Knight and Day (2010)	0.42	64	[0.03, 0.18, 0.15, 0.01, 0.08, 0.03, 0.07, 0.05, 0.14, 0.13]	0.35
Crossroads (2002)	0.31	36	[-0.02, 0.19, 0.21, 0.09, 0.06, -0.01, 0.11, 0.07, 0.14, -0.01]	0.31
Man on Fire (2004)	0.26	34	[0.05, 0.28, 0.17, 0.05, 0.17, 0.03, -0.07, 0.04, 0.07, 0.02]	0.49
The Truman Show (1998)	0.52	67	[0.08, 0.12, 0.24, 0.07, 0.09, 0.07, 0.06, 0.06, 0.08, 0.11]	0.37
Gone in Sixty Seconds (2000)	0.30	41	[-0.03, 0.24, 0.11, 0.06, 0.13, 0.04, 0.07, 0.02, 0.03, 0.14]	0.36
Memento (2000)	0.43	74	[0.06, 0.04, 0.08, 0.07, 0.10, 0.11, 0.09, 0.05, 0.13, 0.09]	0.79
Blues Brothers 2000 (1998)	0.32	47	[0.02, 0.09, 0.06, 0.05, 0.14, 0.02, 0.19, 0.06, 0.15, 0.01]	0.48
The Green Mile (1999)	0.31	78	[0.09, 0.04, 0.15, 0.05, 0.00, 0.06, 0.12, 0.04, 0.13, 0.07]	0.98
The Transporter (2002)	0.45	31	[0.14, 0.19, 0.14, -0.01, 0.02, 0.09, -0.06, 0.04, 0.13, 0.12]	0.53
Child's Play (1988)	0.35	73	[0.07, 0.08, 0.19, 0.14, 0.08, 0.04, 0.02, 0.08, 0.08, 0.12]	0.11
You're Next (2011)	0.35	44	[0.06, 0.07, 0.12, 0.04, 0.09, 0.03, 0.12, 0.01, 0.15, 0.11]	0.42
Full Metal Jacket (1987)	0.38	13	[-0.03, 0.19, 0.16, 0.10, 0.22, 0.02, -0.08, -0.01, 0.35, 0.09]	0.06
Angels in the Outfield (1994)	0.39	59	[0.09, 0.12, 0.17, 0.04, 0.05, 0.02, 0.12, 0.03, 0.17, 0.13]	0.06

Table 3.1.1 Ridge Regression for Middle 30 Movies

Predicted Movie	RMSE	Alpha	Beta	Intercept
The Thing (1982)	0.37	0.0066	[0.0, 0.11, 0.17, 0.09, 0.20, 0.05, 0.01, 0.03, 0.15, 0.04]	0.26
12 Monkeys (1995)	0.38	0.0081	[0.11, 0.13, 0.12, 0.03, 0.15, 0.03, 0.0, 0.0, 0.09, 0.04]	0.82
Bad Boys (1995)	0.40	0.0096	[0.08, 0.01, 0.14, 0.01, 0.01, 0.05, 0.0, 0.01, 0.20, 0.21]	0.91
Armageddon (1998)	0.40	0.0036	[0.0, 0.13, 0.15, 0.03, 0.05, 0.03, 0.27, 0.09, 0.15, 0.0]	0.33
Bad Boys 2 (2003)	0.43	0.0036	[0.06, 0.11, 0.02, 0.02, 0.02, 0.06, 0.35, 0.04, 0.04, 0.17]	0.53
The Mist (2007)	0.37	0.0036	[0.10, 0.0, 0.09, 0.00, 0.0, 0.02, 0.26, 0.0, 0.24, 0.14]	0.35
Braveheart (1995)	0.40	0.0046	[0.12, 0.11, 0.06, 0.07, 0.04, 0.04, 0.27, 0.02, 0.07, 0.01]	0.70
The Others (2001)	0.37	0.0046	[0.01, 0.0, 0.11, 0.04, 0.04, 0.03, 0.21, 0.07, 0.28, 0.0]	0.61
Baby Geniuses (1999)	0.41	0.0101	[0.0, 0.27, 0.11, 0.02, 0.13, 0.05, 0.0, 0.05, 0.15, 0.03]	0.28
One Flew Over the Cuckoo's Nest (1975)	0.42	0.0076	[0.06, 0.07, 0.02, 0.07, 0.12, 0.03, 0.09, 0.0, 0.23, 0.10]	0.89
Honey (2003)	0.39	0.0031	[0.01, 0.18, 0.20, 0.05, 0.07, 0.0, 0.0, 0.05, 0.32, 0.03]	0.14
Big Daddy (1999)	0.29	0.0096	[0.14, 0.15, 0.10, 0.0, 0.0, 0.09, 0.29, 0.04, 0.00, 0.07]	0.53
Halloween (1978)	0.42	0.0011	[0.09, 0.05, 0.11, 0.08, 0.04, 0.07, 0.07, 0.01, 0.23, 0.12]	0.37
The Poseidon Adventure (1972)	0.44	0.0041	[0.04, 0.13, 0.18, 0.05, 0.04, 0.02, 0.12, 0.03, 0.16, 0.0]	0.44
Rocky (1976)	0.49	0.0096	[0.12, 0.04, 0.05, 0.09, 0.15, 0.01, 0.13, 0.0, 0.0, 0.39]	0.25
The Intouchables (2011)	0.43	0.0021	[0.16, 0.27, 0.08, 0.06, 0.05, 0.07, 0.0, 0.05, 0.09, 0.03]	0.68
Aliens (1986)	0.33	0.0046	[0.0, 0.04, 0.12, 0.04, 0.06, 0.05, 0.08, 0.02, 0.31, 0.23]	0.23
Knight and Day (2010)	0.43	0.0056	[0.0, 0.24, 0.18, 0.0, 0.06, 0.01, 0.0, 0.03, 0.20, 0.15]	0.28
Crossroads (2002)	0.32	0.0001	[-0.05, 0.24, 0.27, 0.09, 0.04, -0.02, 0.14, 0.06, 0.15, -0.04]	0.21
Man on Fire (2004)	0.26	0.0006	[0.04, 0.36, 0.21, 0.04, 0.21, 0.03, -0.19, 0.03, 0.09, 0.01]	0.40
The Truman Show (1998)	0.51	0.0031	[0.06, 0.13, 0.39, 0.07, 0.10, 0.07, 0.0, 0.05, 0.07, 0.10]	0.25
Gone in Sixty Seconds (2000)	0.30	0.0086	[0.0, 0.30, 0.09, 0.05, 0.13, 0.03, 0.01, 0.01, 0.0, 0.15]	0.43
Memento (2000)	0.43	0.0026	[0.05, 0.0, 0.08, 0.07, 0.12, 0.12, 0.08, 0.04, 0.21, 0.11]	0.67
Blues Brothers 2000 (1998)	0.33	0.0031	[0.0, 0.09, 0.03, 0.04, 0.15, 0.01, 0.30, 0.05, 0.16, 0.0]	0.42
The Green Mile (1999)	0.30	0.0081	[0.08, 0.0, 0.22, 0.04, 0.0, 0.05, 0.15, 0.02, 0.15, 0.04]	0.97
The Transporter (2002)	0.47	0.0001	[0.16, 0.24, 0.16, -0.02, 0.01, 0.10, -0.20, 0.03, 0.21, 0.14]	0.44
Child's Play (1988)	0.36	0.0036	[0.05, 0.06, 0.30, 0.16, 0.07, 0.03, 0.0, 0.07, 0.07, 0.12]	0.00
You're Next (2011)	0.34	0.0031	[0.05, 0.05, 0.13, 0.03, 0.09, 0.02, 0.13, 0.00, 0.20, 0.12]	0.38
Full Metal Jacket (1987)	0.38	0.0056	[0.0, 0.16, 0.13, 0.10, 0.20, 0.01, 0.0, 0.0, 0.34, 0.06]	0.11
Angels in the Outfield (1994)	0.39	0.0066	[0.08, 0.13, 0.22, 0.03, 0.01, 0.0, 0.09, 0.02, 0.25, 0.14]	0.03

Table 3.2.1 Lasso Regression for Middle 30 Movies

Franchise: Star Wars	Average COD by franchise	Average COD by non-franchise
Star Wars: Episode IV - A New Hope (1977)	0.29	0.11
Star Wars: Episode II - Attack of the Clones (2002)	0.20	0.08
Star Wars: Episode V - The Empire Strikes Back (1980)	0.32	0.14
Star Wars: Episode 1 - The Phantom Menace (1999)	0.22	0.12
Star Wars: Episode VII - The Force Awakens (2015)	0.29	0.13
Star Wars: Episode VI - The Return of the Jedi (1983)	0.34	0.15

Franchise: Harry Potter	Average COD by franchise	Average COD by non-franchise
Harry Potter and the Sorcerer's Stone (2001)	0.49	0.12
Harry Potter and the Deathly Hallows: Part 2 (2011)	0.45	0.10
Harry Potter and the Goblet of Fire (2005)	0.48	0.10
Harry Potter and the Chamber of Secrets (2002)	0.52	0.12

Franchise: The Matrix	Average COD by franchise	Average COD by non-franchise
The Matrix Revolutions (2003)	0.27	0.09
The Matrix Reloaded (2003)	0.29	0.13
The Matrix (1999)	0.25	0.14

Franchise: Indiana Jones	Average COD by franchise	Average COD by non-franchise
Indiana Jones and the Last Crusade (1989)	0.28	0.11
Indiana Jones and the Temple of Doom (1984)	0.27	0.13
Indiana Jones and the Raiders of the Lost Ark (1981)	0.22	0.15
Indiana Jones and the Kingdom of the Crystal Skull (2008)	0.14	0.11

Franchise: Jurassic Park	Average COD by franchise	Average COD by non-franchise
The Lost World: Jurassic Park (1997)	0.30	0.12
Jurassic Park III (2001)	0.26	0.09
Jurassic Park (1993)	0.25	0.16

Franchise: Pirates of the Caribbean	Average COD by franchise	Average COD by non-franchise
Pirates of the Caribbean: Dead Man's Chest (2006)	0.32	0.07
Pirates of the Caribbean: At World's End (2007)	0.38	0.13
Pirates of the Caribbean: The Curse of Black Pearl (2003)	0.33	0.13

Franchise: Toy Story	Average COD by franchise	Average COD by non-franchise
Toy Story 2 (1999)	0.46	0.12
Toy Story 3 (2010)	0.45	0.10
Toy Story (1995)	0.44	0.13

Franchise: Batman	Average COD by franchise	Average COD by non-franchise
Batman & Robin (1997)	0.11	0.10
Batman (1989)	0.15	0.16
Batman: The Dark Knight (2008)	0.07	0.11

Table 5.1 Average COD Predicted by Movies from Franchise and Movies Not from Franchise