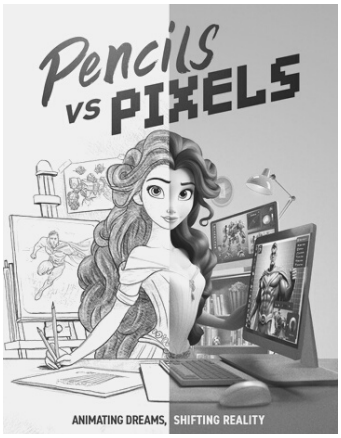
MIDTERM PROJECT

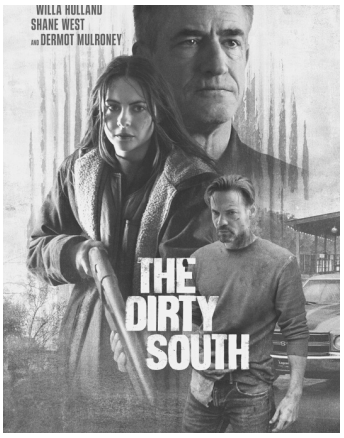
THE 2023 IMDB PREDICTION CHALLENGE

The 2023 IMDb Prediction Challenge

This November, twelve blockbusters will be released...

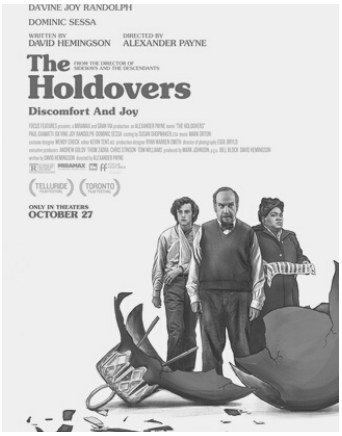


Pencils vs Pixels
(Nov 7)

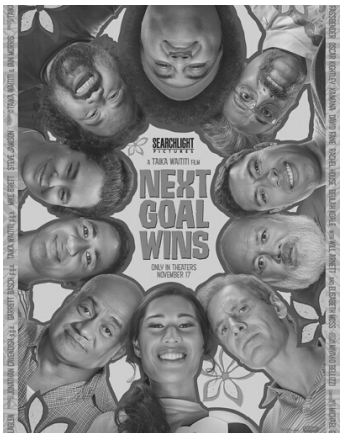


The Dirty South
(Nov 10)

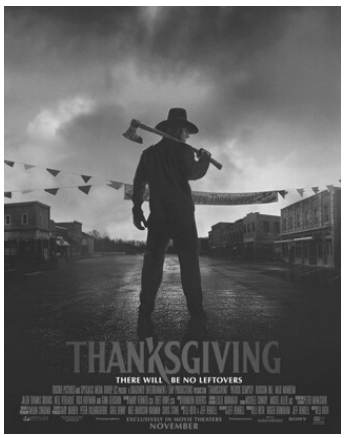
The Marvels
(Nov 10)



The Holdovers
(Nov 10)

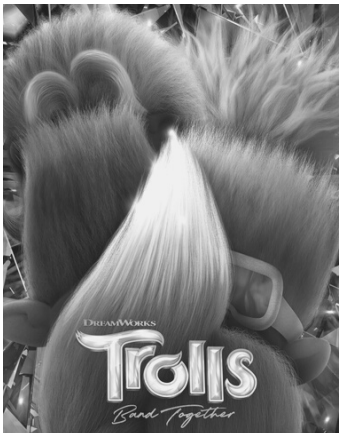


Next Goal Wins
(Nov 17)



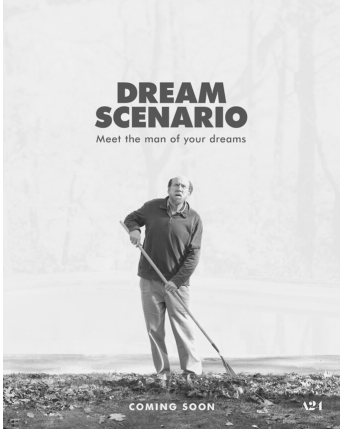
Thanksgiving
(Nov 17)

The Hunger Games: The Ballad of Songbirds and Snakes
(Nov 17)

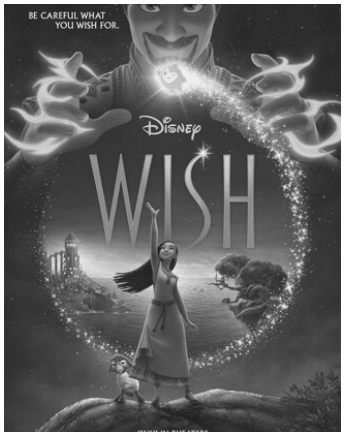


Trolls Band Together
(Nov 17)

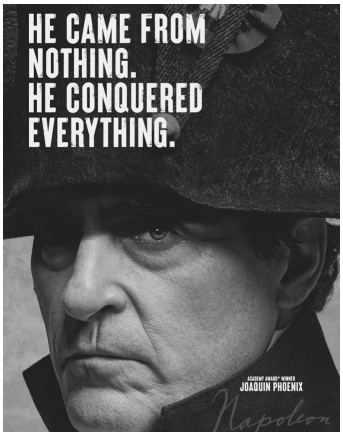
Leo
(Nov 21)



Dream Scenario
(Nov 22)



Wish
(Nov 22)



Napoleon
(Nov 22)

The 2023 IMDb Prediction Challenge

...and it is up to us to predict whether the public will love them or hate them.

Throughout the past lectures, we have learned the statistical foundations of predictive data analytics. We know how to construct a model with enough flexibility to elevate our R-squared to very high levels. We know how to deal with biased estimates, heteroskedasticity, collinearity, and how to confront outliers. But we also know about the dangers of overfitting. It is now up to us to craft a model with colossal predictive power to anticipate the critical reception of these movies.

More explicitly, **our goal is to predict the IMDb ratings of the twelve upcoming blockbusters**. To train our statistical model, I have gathered data from about 2,000 movies on IMDb. For each movie, I have information on the following variables:

DATA DICTIONARY

Note: We provide a data dictionary in myCourses, with detailed information about each variable.

1. Identifiers

- movie_title
- movie_id
- imdb_link

2. Dependent Variable

- imdb_score

3. Film Characteristics

- movie_budget
- release_day
- release_month
- release_year
- duration
- language
- country
- maturity_rating
- nb_news_articles
- colour_film
- nb_faces
- plot_keywords
- dummy variables for all genres
- movie_meter_IMDBpro

4. Cast Characteristics

- actor1
- actor1_star_meter (IMDbPro)
- actor2
- actor2_star_meter (IMDbPro)
- actor3
- actor3_star_meter (IMDbPro)

5. Production Characteristics

- director
- distributor
- cinematographer
- production_company

The 2023 IMDb Prediction Challenge

WHAT ARE WE TRYING TO PREDICT, EXACTLY?

We're trying to predict the IMDb rating of the movies. Note that the IMDb ratings vary and can change over time, but the rating is quite stable after the first few days of the release.

HOW DO WE CRAFT A PERFECT STATISTICAL MODEL?

I want to emphasize this: **statistical modelling is an art**. It is like sculpting. You will need some tools (some of which we have already learned in this course). But ultimately, your creativity, intuition, and statistical knowledge are what will lead you to create a good predictive model. You possess the tools to do some serious predictive modeling—all the material from Lectures 1-6—but there are no written rules when it comes to building a statistical model.

I expect every team to come up with an entirely different statistical model. Even in professional data-science competitions, statistical models are tremendously different. This is what makes our job exciting—that data science forces us to be creative while using technical tools.

DO YOU EXPECT OUR MODEL TO BE WITHOUT ISSUES?

You cannot get rid of all problems (collinearity, heteroskedasticity, insignificance, overfitting, etc.). You cannot build a perfect model. It's a process that works based on intuition and trial and error. In my own research, I typically test more than 300 models, run over 50 tests, and experiment with different predictors.

The more you work in statistics, the more you will begin to get a "feel for the data." This is your first trial and, as such, I want you to tackle this task with an open mind. The important thing is to lose yourself and start playing with the data. **You are not expected** to use all the data available. You are not expected to include all predictors in the model. You are not expected to use all the tools we learned in class.

Yes, the goal is to come up with the most powerful predictive model but it is also your first time doing serious statistical modelling, and I want you to be creative and develop your data-scientist style. Above all, have fun and take this opportunity as a way to determine if data analytics is a good career path for you.

On the following pages, I will guide you with some rough steps that I typically apply when building a statistical model.

The 2023 IMDb Prediction Challenge

1. BUILDING THE MODEL: WHICH STEPS SHOULD WE FOLLOW?

Note: these steps are just suggestions. You're free to follow your own approach and deviate

STEP 1. GET A FEEL OF THE DATASET

First of all, try to see the dataset, look at the variables, understand what they measure, read the data dictionary, and analyze the observations. Normally, you would have to spend weeks cleaning the data, but we've saved you time. We spent a generous budget on RAs, who cleaned this dataset and collected the variables. Thus, the data should be in good shape. However, you might still find incorrect data entries, so take a couple of hours to look over the data

STEP 2. EXPLORE THE VARIABLES INDIVIDUALLY

- What are the distributions of the variables? (Use Box Plots/Histograms)
- Which variables are skewed?
- What about the correlation amongst the variables?
- Are there any observations that display unusual behaviour/outliers?
- Is there any collinearity among variables?

STEP 3. EXPLORE VARIABLE RELATIONSHIPS

- Examine the correlation coefficient between Y and each predictor, xi;
  - Is it positive or negative?
  - Is it weak or strong?
- Look at a scatter plot between Y and each xi, and run a non-constant variance test. Is heteroskedasticity present?
- Flag potential heteroskedastic predictors.
- Run simple linear regressions between Y and each predictor xi.
- Look at the p-values and the r-squared of each regression.
- This should give you a sense about which variables have more linear predictive power.
- Examine correlations between all predictors:
  - Use a correlation matrix and look at the correlation coefficients.
  - Take note of possible collinearity.

The 2023 IMDb Prediction Challenge

STEP 4. TEST NON-LINEARITY AND FIT

- Again, look at the relationship between each Y and xi
- You should test the linearity assumption of the data and see if the relationships should be modeled non linearly.
- Try different polynomial functions and play with the degree to determine which one gives you a higher r-squared and out-of-sample performance.
- Determine if a spline functional form can improve the fit of a predictor. Start playing with the knots and the degrees of the splines.

STEP 5. BUILD A REGRESSION MODEL

- Once you have determined the relationship between Y and each xi, make a rough ranking of the identified predictors to see which are the most significant.
- Begin by bundling, one-by-one, the predictors that are most powerful.
- Start seeing if you should add interactions between predictors.
- Start deciding which dummy variables to include.
- Start running diagnostics for each model.
- Again, there are no rules here. You need to use your intuition and play with different versions of the model.

Warning: More predictors or a more complicated function f() is not synonymous with a better model. A well-crafted linear regression with two predictors can do wonders. I won't penalize for having a simple function with a few predictors, if you can justify this choice. I often see data scientists who run into the temptation of adding more complexity into the model for the sake of complexity. I have come across models with 100 predictors that have poor predictive power. Don't fall into this trap.

STEP 6. TEST OUT-OF-SAMPLE PERFORMANCE

The probability of overfitting a model will increase when you: add more predictors, add interactions, increase the polynomial degree, or increase the knots in a spline. Decide which test you'll apply (validation-set test, K-fold, or LOOCV test). If the model has poor out-of-sample performance, you may want to simplify it. If the model has the potential to gain more predictive power, keep adding complexity.

The 2023 IMDb Prediction Challenge

2. DELIVERABLES: DELIVERING YOUR RESULTS

Just building a good model is not enough. After all, you are scientific communicators. A well-drafted report is essential for you to communicate your model to the world. This is as important as having a powerful model. A poorly presented report will drive your model into obscurity, as no one will read it.

The report should be typed, clear, and aesthetically pleasing. The length of the report should be between 6 and 9 pages (1.5 spaced). Up to additional 10 pages may be appended for exhibits (e.g., extra tables, figures). Please use the following framework to organize your paper (Feel free to deviate slightly if you feel it improves the flow).

1

Introduction (1/2 - 1 page)

Here you provide a summary of the project, the goals, etc.

2

Data Description (2 - 3 pages)

Here you describe the distribution of the dependent variable and independent variables, and the relationship between these variables,

3

Model Selection (1/2 - 1 page)

- Tell us about the methodology you used to build your model.
- Explain your rationale for modelling predictors as linear, quadratic, splines, etc. You should also justify why you decided to add X many knots to a spline, etc.
- Tell us about your rationale for including or excluding each predictor. To this end, you should discuss model issues, such as heteroskedasticity, collinearity, underfitting, overfitting, etc.

The 2023 IMDb Prediction Challenge

4

Results (2 - 3 pages)

Present the result for your final model. Tell us about your predictions for the 12 movies, the r-square of the model, the predictive power (i.e., out-of-sample performance), and the significance of each predictor.

5

Appendices (max 10 pages)

All tables and exhibits should be after the conclusions. All tables should be labeled and named.

6

Code

Please attach your R code at the end.

TIP 1: THE RATIONALE IS KEY

You should NOT tell me everything you did, or how hard you worked. Save it for your memoirs. As a reader, I am interested in seeing that you approached this problem using scientifically-rigorous techniques, and your rationale behind them. I don't need to learn about the 450 models you didn't end up using. I also don't need to see 50 different scatter plots or graphs if they don't contribute to my understanding of the model.

TIP 2: PROPER LABELS

Make sure your variables are properly labelled in the table, including a caption. For example, if a variable is called mov_rat_IMDB you should probably rename it to "IMDB ratings" in the paper's tables.

TIP 3: BEWARE OF PARAPHRASING

Your text shouldn't paraphrase your tables. Whatever I can see in the regressions, I don't need to hear from you. So, save your space and avoid telling me that the p-value of regression X is equal to 0.005 if I can see that from the table.

The 2023 IMDb Prediction Challenge

TYPESETTING SOFTWARE

Technical papers, books, and most professional materials aren't typeset with Microsoft Word. Most book publishers, scientists, professors, and agencies use a typesetting language called **LaTeX**. LaTeX is a language that uses coding to typeset documents. After coding a document, you compile it and it will produce a beautiful report. This makes typing mathematical equations incredibly easy—as opposed to using Microsoft Equation Editor. You can also export your stargazer tables directly into LaTeX code. Look at the difference:

Word

sequence (in any order). Formally, we say that a rule $I_a \Rightarrow I_b$ *occurs in a sequence* $s = \langle I_1, I_2, \ldots I_n \rangle$ if and only if there exists an integer k such that $1 \leq k < n$, $I_a \subseteq \bigcup_{i=1}^{k} I_i$ and $I_b \subseteq \bigcup_{i=k+1}^{n} I_i$.

Latex

same sequence (in any order). Formally, we say that a rule $I_a \Rightarrow I_b$ occurs in a sequence $s = \langle I_1, I_2, \ldots I_n \rangle$ if and only if there exists an integer k such that $1 \leq k < n$, $I_a \subseteq \bigcup_{i=1}^{k} I_i$ and $I_b \subseteq \bigcup_{i=k+1}^{n} I_i$.

- The wonderful thing about LaTeX is that it is an open-source language that is free. Like RStudio, there are programs that make it easier to use LaTeX. My favourite one is **LyX***, which you can download here: https://www.lyx.org/Download.

- Although you are not required to typeset your documents in LaTeX—and I will not grade you based on this factor—I highly recommend you to download the software and give it a try. Learning LaTeX will be a great way to impress your employers, professors, and grad-school committees. LaTeX allows you to create gorgeous CVs, slide presentations, letters, etc. (that's the reason most publishers use it to print their books!).

- Like R, you will face a short learning curve when beginning typesetting in LaTeX. But after a few weeks, you will find it much better than Microsoft Word, or typical typesetting processors.

The 2023 IMDb Prediction Challenge

GRADING

Your grade will be out of 50, and you will be graded on the following criteria:

Criterion	Reasoning	Max Points
Statistical Analysis	Addresses objective of the analysis using rigorous and thorough statistical techniques. Builds a model based on rigorous analysis. Finds a nice balance between a model that isn't overly simplistic nor overly complex.	10
Interpretation & Conclusions & Recommendations	Correctly interprets all analysis, draws appropriate conclusions, makes predictions based on sound interpretations of the model	10
Flow, Organization & Structure	Report well-organized into different sections and clearly structured following the instructions.	10
Visual presentation of data	Tables are neatly organized and presented. Graphs are visually pleasing and well organized. Has enough graphs to make a complete analysis, but not an excessive number of graphs to overburden the reader.	10
Writing: clarity, correctness, creativity, and style	Statistical analysis is clearly explained and in a creative and professional style throughout report, while respecting the word limits.	10

DUE DATE

The report is due on October 30, at 11:59pm. There will be a submission link in myCourses. Clearly indicate your group name. Only one submission per group.

GET A BONUS

Win Free Cinema Tickets

Your grade will not depend on your predictions. But the group with the most accurate predictions will get free movie tickets! To assess the success of your predictions, we will look at the IMDB ratings of all movies on the last class. The group with the lowest MSE (across all movies) will win.

Once you have sent me your report, please go to the link I will open in mycourses (by the submission tab), and write your group's predictions. To play the game, the predictions must match the ones in your report.

The 2023 IMDb Prediction Challenge

IMDB Links

Pencils vs Pixels: https://www.imdb.com/title/tt26918463

The Dirty South: https://www.imdb.com/title/tt9114286

The Marvels: https://www.imdb.com/title/tt10676048

The Holdovers: https://www.imdb.com/title/tt14849194

Next Goal Wins: https://www.imdb.com/title/tt10767052

Thanksgiving: https://www.imdb.com/title/tt1448754

The Hunger Games: https://www.imdb.com/title/tt10545296

Trolls Band Together: https://www.imdb.com/title/tt14362112

Leo: https://www.imdb.com/title/tt5755238

Dream Scenario: https://www.imdb.com/title/tt21942866

Wish: https://www.imdb.com/title/tt11304740

Napoleon: https://www.imdb.com/title/tt13287846

The 2023 IMDb Prediction Challenge

FREQUENTLY ASKED QUESTIONS

HEY, SOME MOVIES IN THE TEST DATA ALREADY HAVE AN IMDB SCORE! CAN WE USE IT TO BUILD OUR PREDICTIONS?

Do not rely on the existing score for your predictions! The score stems from a very small number of users who saw a pre-release in another region of the world. When the movie is released in North America, the IMDb score will change dramatically because North Americans rate movies differently than users from other regions.

WE FOUND MISTAKES IN THE DATA! DO WE HAVE TO CLEAN THE ENTIRE DATASET?

We spent over 100 hours (and a generous budget) collecting the data and cleaning the variables, with the help of RAs. This dataset is one of the most complete movie datasets out there.

Despite this, I wouldn't be surprised if you find some recording mistakes, or errors in the input of the data. But this is common in virtually every large dataset. In fact, we've made it substantially easier for you by doing most of the data cleaning. As long as 99% of the data points are properly labelled, a few data mistakes aren't an issue at all. So the answer is no: you're not expected to clean the entire dataset—this would be counterproductive and won't make a slight difference in your final predictions.

CAN WE ADD MORE VARIABLES TO THE DATASET, OR USE OTHER TECHNIQUES NOT TAUGHT IN CLASS?

Yes you may, sparingly. But I will only evaluate your mastery of the class topics, so try not to rely too heavily on external methods/data. Also, try not to use variables like "number of user votes" or "oscars won" by a movie. Remember that we're trying to predict movies that will be released in the future, so they won't win any awards, and won't be voted on by users before their release.