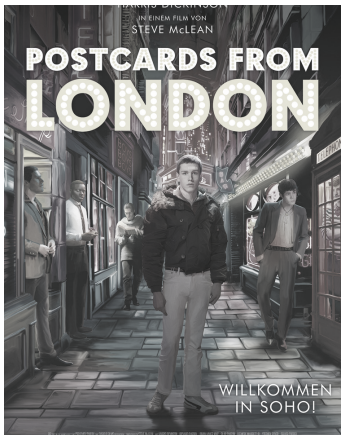
THE IMDB PREDICTION CHALLENGE

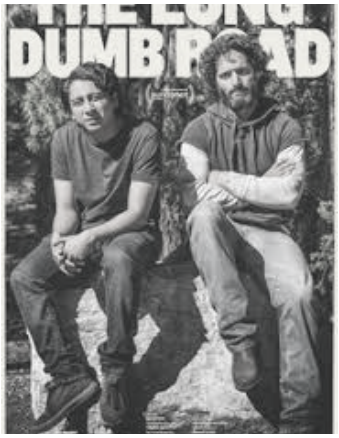
DESCRIPTION

In the fall, 12 blockbuster movies will be coming out:
Please find the link of the iMDB page at the end of the midterm booklet.

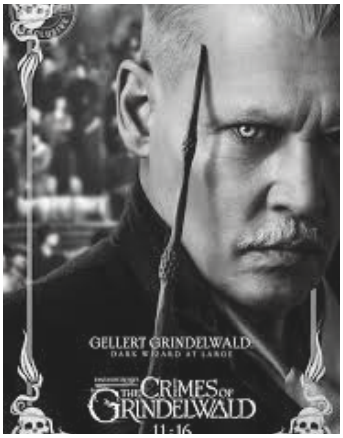
The Grinch
(Nov. 9)



Postcards from London
(Nov 9)

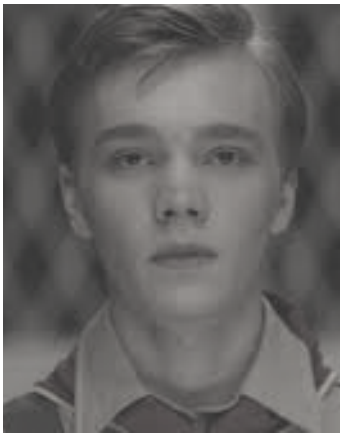


The Long Dumb Road
(Nov 9)

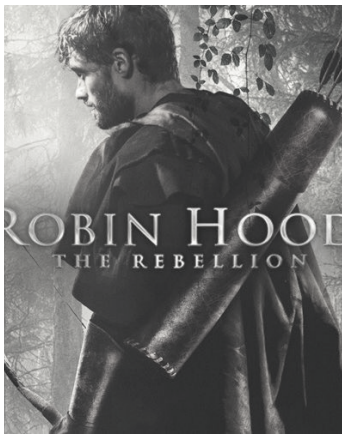


Crimes of Grindelwald
(Nov 16)

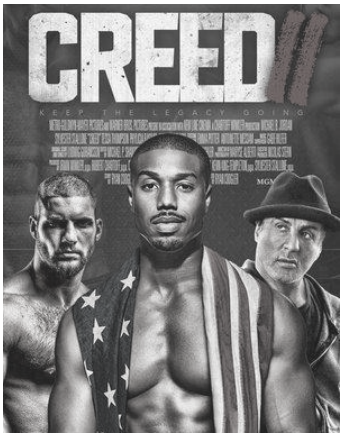
Instant Family
(Nov 16)



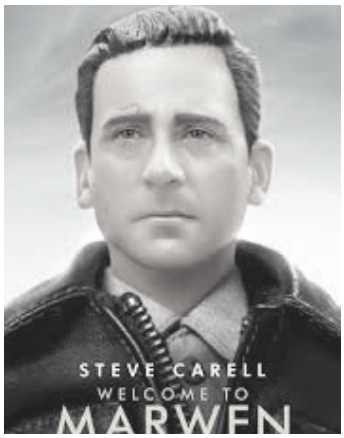
The Clovehitch Killer
(Nov 16)



Robin Hood
(Nov 21)

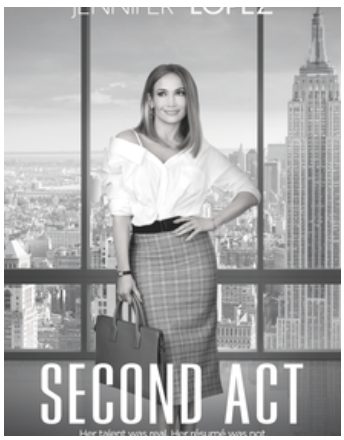


Creed II
(Nov 21)



The Women of Marwen
(Nov 21)

Ralph Breaks the Internet
(Nov 23)



Second Act
(Nov 23)

Becoming Astrid
(Nov 23)

Will people like them? Will they hate them? How will they rate them? Aha! That's for us to predict.

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 R2 to levels previously unimagined. We know how to deal with biased estimates, with heteroskedasticity, with 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 that can have immense 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 your statistical model, I have gathered data from IMDB, collecting the characteristics from 3,000 movies. For each movie, I have information on the following factors:

1. Labels

movie_id_number
movie_imdb_link
title

2. Dependent Variable

imdb_score

3. Predictors

release_month
release_day
budget
release_year
duration_mins
language
country
content_rating
aspect_ratio
producer
distributor
number_news_articles (available in IMDB Pro)
imdb_rating
director
director_facebook_likes
actor_1_name
actor_1_facebook_likes
actor_1_known_for (available in IMDB Pro)
actor_1_star_meter (available in IMDB Pro)
actor_2_name
actor_2_facebook_likes
actor_2_known_for (available in IMDB Pro)
actor_2_star_meter (available in IMDB Pro)
actor_3_name
actor_3_facebook_likes
actor_3_known_for (available in IMDB Pro)
actor_3_star_meter (available in IMDB Pro)
user_votes_number

critic_reviews_number
user_reviews_number
color
number_professional_critic_reviews_IMDB
genres
cast_total_facebook_likes
number_of_faces_movie_poster
plot_keywords
movie_facebook_likes
action
adventure
scifi
thriller
musical
romance
western
sport
horror
drama
war
animation
crime
sum_total_likes
ratio_movie_cast_likes
movie_meter_IMDB (available in IMDB Pro)
number_of_votes
production_company
plot_summary

*Note: you can get free membership to IMDB Pro for a limited time. You should get the trial
Note 2: Just in case you're wondering, an R-rated movie is not a movie about statistical programming.

How should you go about creating the best predictive statistical model?

I want to emphasise 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, it is your creativity, your intuition, and your statistical experience that will yield a good predictive model. You possess the tools to do some serious predictive models (i.e., all the material from Lectures 1-6), but there are no written rules when it comes to building a statistical model.

No model is perfect!

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 models, I typically test more than 300 models, run over 50 tests and play with different predictors.

The more you work in statistics, the more you will begin to get a "feeling 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.

Each group will come up with a different model, and this is completely normal. 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 venture into it and develop your data-scientist style.

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

Part 1: Crafting the Model

Step 1: Exploring Distributions

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

Step 2: Exploring Relationships between variables

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

Step 3: Test non-linearity & fit

Again, look at the relationship between each Y and xi

1. You should test the linearity assumption of the data and see if the relationships should be modelled in a non-linear fashion.
2. Try different polynomial functions and play with the degree to determine which one gives you a higher r-squared and out-of-sample performance
3. 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 4: Building a multiple regression model

1. Once you have determined the relationship between Y and each xi, make a rough rank of the identified predictors to find the highest to the lowest statistical significance.
2. Begin by bundling, one-by-one, the predictors that are most powerful. Start seeing if interactions make sense.
3. Start seeing if you should add interactions between predictors.
4. Start deciding which dummy variables to include.
5. Start running diagnostics for each model
6. Again, there are no rules here. You need to use your intuition and play with different model versions.

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 5: Test for the model's out of sample performance

The probability of overfitting a model will increase when you:

1. Add more predictors
2. Add interactions
3. Increase the polynomial degree
4. Increase the knots in a spline

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.

Again, the above steps are my own suggestions. Feel free to deviate from them if you have a model-building method that works better for your group.

Part 2: Deliverables: Presenting Your Results

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 correct model. A poorly presented report will drive your model to obscurity, as no one will read it.

The report should be typed, clear, and **aesthetically pleasing**. The length the report should be **between 6 and 9 pages** (1.5 spaced). Exhibits (e.g., extra tables, figures, etc.): up to additional 10 pages may be appended. Please use the following framework to organize your paper (again, just a rough suggestion—feel free to deviate).

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)

- Here, you will tell us which methodology you used to build your model.
- You should explain your rationale for modeling predictors as linear, quadratic, splines, etc. You should also justify why you decided to add X many knots to a spline, etc.
- You should also tell us how your rationale for including or excluding each predictor. To this end, you should discuss model issues, such as heteroskedasticity, collinearity, underfitting, overfitting, etc.

4

Results (2 - 3 pages)

Present the result for your final model. Tell us about 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.

WARNING 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 a 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.

WARNING 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.

WARNING 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.

Typesetting Software

Technical papers, books, and most professional material isn'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 typsetting in LaTeX language. But after a few weeks, you will find it much better than Microsoft Word, or typical typesetting processors.

*Think of LaTeX as R (the language), and think about LyX as Rstudio. If you are using Windows, you can download both as a bundle. If you are using a mac laptop, you will first need to install LaTeX language and then Lyx.

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 Nov 2, at midnight. 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 November 28 at 2:15 pm (end of class). The group with the lowest MSE (across all 9 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!

Midterm project

Link to iMDB page:

The Grinch: https://www.imdb.com/title/tt2709692/
Postcards from London: https://www.imdb.com/title/tt6280608/?ref_=nv_sr_1
The Long Dumb Road: https://www.imdb.com/title/tt4712076/?ref_=nv_sr_1
Fantastic Beasts: The Crimes of Grindelwald: https://www.imdb.com/title/tt4123430/?ref_=nv_sr_1
Instant Family: https://www.imdb.com/title/tt7401588/?ref_=nv_sr_1
The Clovehitch Killer: https://www.imdb.com/title/tt6269368/?ref_=nv_sr_1
Robin Hood: https://www.imdb.com/title/tt4532826/?ref_=nv_sr_1
Creed II: https://www.imdb.com/title/tt6343314/?ref_=nv_sr_1
The Women of Marwen: https://www.imdb.com/title/tt3289724/?ref_=nv_sr_1
Ralph breaks the internet: https://www.imdb.com/title/tt5848272/?ref_=nv_sr_1
Second Act: https://www.imdb.com/title/tt2126357/?ref_=nv_sr_1
Becoming Astrid: https://www.imdb.com/title/tt6433456/?ref_=nv_sr_1