TV Show Platforms and Ratings

Alain Duplan
Linus Hsu
Corey Kozlovski
Hoseung Baek

Background

Our goal is to use R to determine which streaming service has the highest rated shows according to IMDb and Rotten Tomatoes. Our dataset contains information about shows from Netflix, Hulu, Prime Video, and Disney+

Our models focused on analyzing the relationships between the rating of a show and its availability on different streaming service platforms

Tasks

- Does platform availability have a relationship with the ratings?
- Is there a relationship between IMDb and Rotten Tomatoes scores?
- Is it possible to predict ratings based on platform availability?

Data Source & Preprocessing

- https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney
- Dataset was found on Kaggle
- Preprocessing
  - Removed any rows with empty values
  - Added a column named “Index” that assigns a number to each TV show so it’s easily identifiable during analysis

	Index	Title	Year	Age	IMDb	Rotten.Tomatoes	Netflix	Hulu	Prime.Video	Disney.	type
1	0	Breaking Bad	2008	18+	9.5	0.96	1	0	0	0	1
2	1	Stranger Things	2016	16+	8.8	0.93	1	0	0	0	1
3	2	Money Heist	2017	18+	8.4	0.91	1	0	0	0	1
4	3	Sherlock	2010	16+	9.1	0.78	1	0	0	0	1
5	4	Better Call Saul	2015	18+	8.7	0.97	1	0	0	0	1
6	5	The Office	2005	16+	8.9	0.81	1	0	0	0	1

Global Environment

Data	
movies.df	931 obs. of 8 variables
testSet	372 obs. of 2 variables
trainingSet	559 obs. of 2 variables
Values	
predictions	Factor w/ 2 levels "0","1": 2 2 2 2 ...
testOutcomes	Factor w/ 2 levels "0","1": 1 1 1 1 ...
trainingOutc...	Factor w/ 2 levels "0","1": 2 2 2 2 ...

	Year	Age	IMDb	R_Tomatoes	Netflix	Hulu	Prime.Video	Disney.
1	2008	18+	9.5	9.6	1	0	0	0
2	2016	16+	8.8	9.3	1	0	0	0
3	2017	18+	8.4	9.1	1	0	0	0
4	2010	16+	9.1	7.8	1	0	0	0
5	2015	18+	8.7	9.7	1	0	0	0
6	2005	16+	8.9	8.1	1	0	0	0
7	2011	18+	8.8	8.3	1	0	0	0
8	2005	16+	8.4	9.3	1	0	0	0
9	2013	18+	8.8	9.2	1	0	0	0
10	2005	7+	9.2	10.0	1	0	0	0
11	2010	18+	8.2	8.1	1	0	0	0
12	2017	16+	8.7	9.4	1	0	0	0
13	2017	18+	8.4	8.1	1	0	0	0

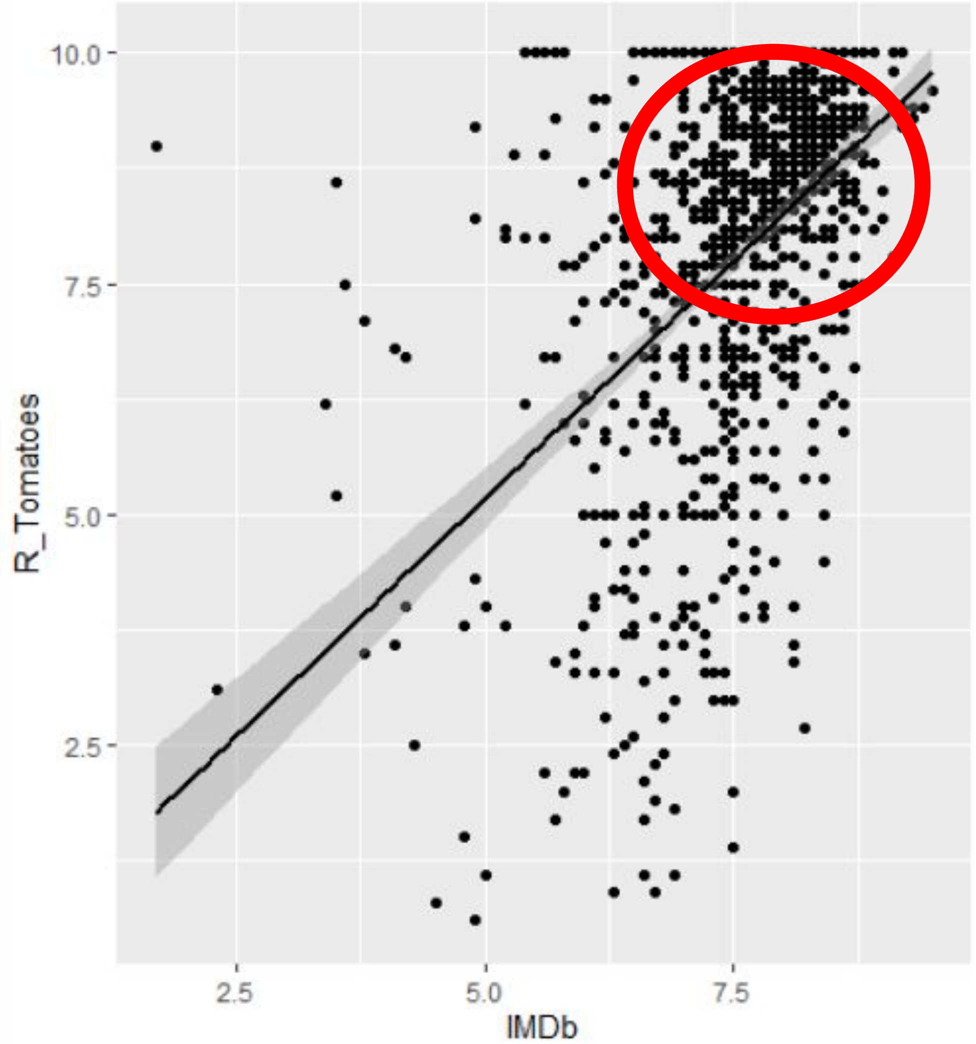
IMDb vs Rotten Tomatoes

Correlation of numerical data:
A check to see if IMDb and Rotten Tomatoes gives similar ratings

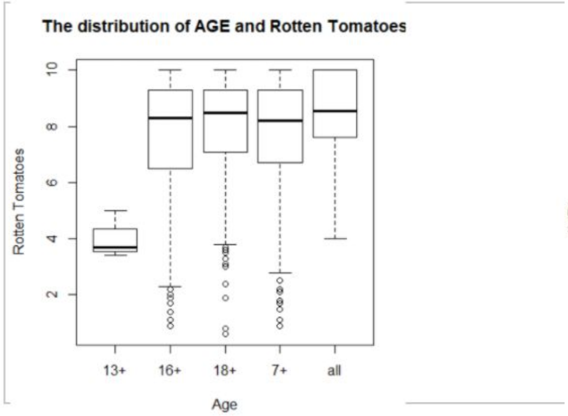
```
ggplot(data = movies.df, mapping = aes(x =IMDb, y = R_Tomatoes)) + geom_point()
+  geom_smooth(method = "lm", color="black", show.legend = FALSE)
```

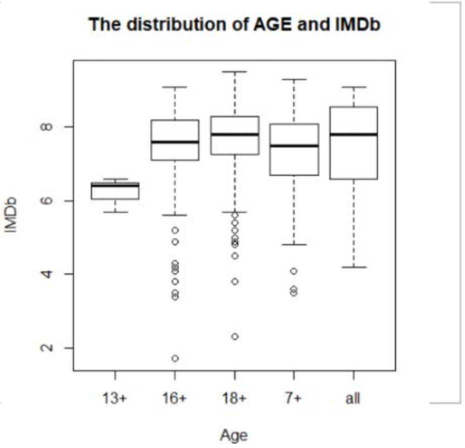
Notable outliers
Strong relationship of two rating systems

```
> table(movies.df$Age)

13+ 16+ 18+  7+ all
  3 359 376 177  16
```

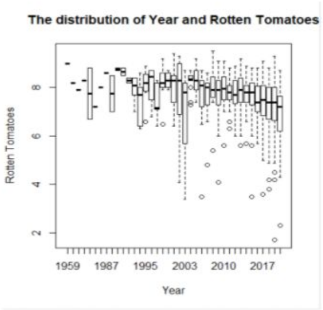


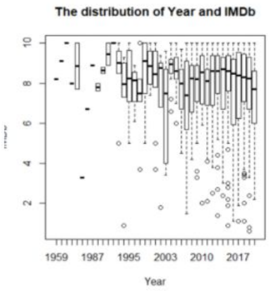
Age and Ratings





Year and Ratings





KNN Predictive Model

- Used R to create a KNN model that predicts platform availability using the two rating systems and its relationship with platform access.

Randomized data

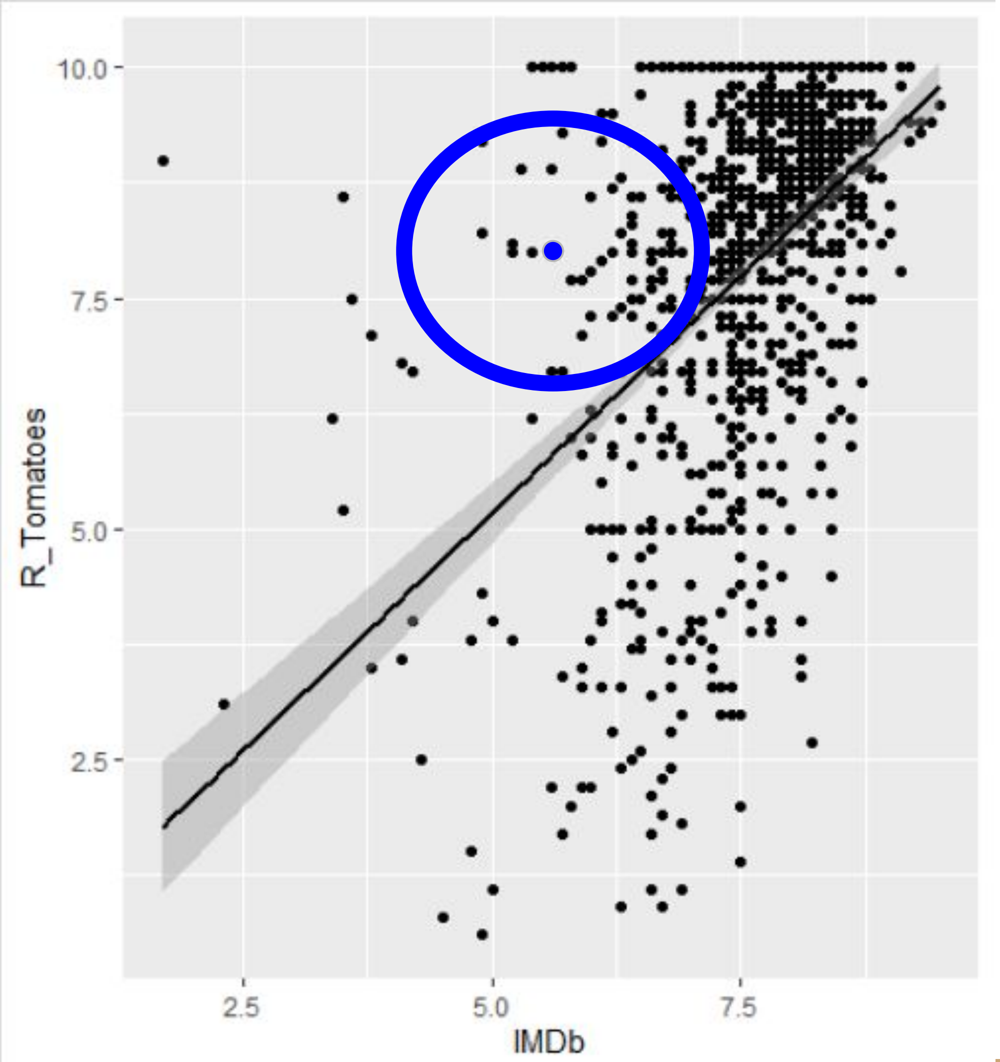
Split Training and Testing Data by 60/40

Based on rows 3:4 (R_Tomatoes/IMDb)

Set k to 30 due to the approximate sq rt of n (count = 931)

```
set.seed(1234)
rows <- sample(nrow(movies.df))
movies.df <- movies.df[rows,]
trainingSet <- movies.df[1:559, 3:4]
testSet <- movies.df[560:931, 3:4]

trainingOutcomes <- movies.df[1:559, 8]
trainingOutcomes <- trainingOutcomes$Disney.


testOutcomes <- movies.df[560:931, 8]
testOutcomes <- testOutcomes$Disney.



library(class)
predictions <- knn(train = trainingSet, cl = trainingOutcomes, k = 30 ,test = testSet)

table(testOutcomes, predictions)
```

Basic KNN Overview



```
             predictions
testOutcomes   0   1
           0 153  47
           1 130  42
```

1.
195 shows predicted correctly
52.4% accuracy model (Netflix)
42/(42+130) = 24.4% Actual positive - Recall

```
             predictions
testOutcomes   0   1
           0 206  16
           1 143   7
```

2.
213 shows predicted correctly
57.2% accuracy model (HULU)
7/(7+143) = 4.66% Actual positive - Recall

```
             predictions
testOutcomes   0   1
           0 295   0
           1  77   0
```

3.
295 shows predicted correctly
79.3% accuracy model (Amazon Prime.Video)
0/(0+77) = 0% Actual positive - Recall

```
             predictions
testOutcomes   0   1
           0 365   0
           1   7   0
```

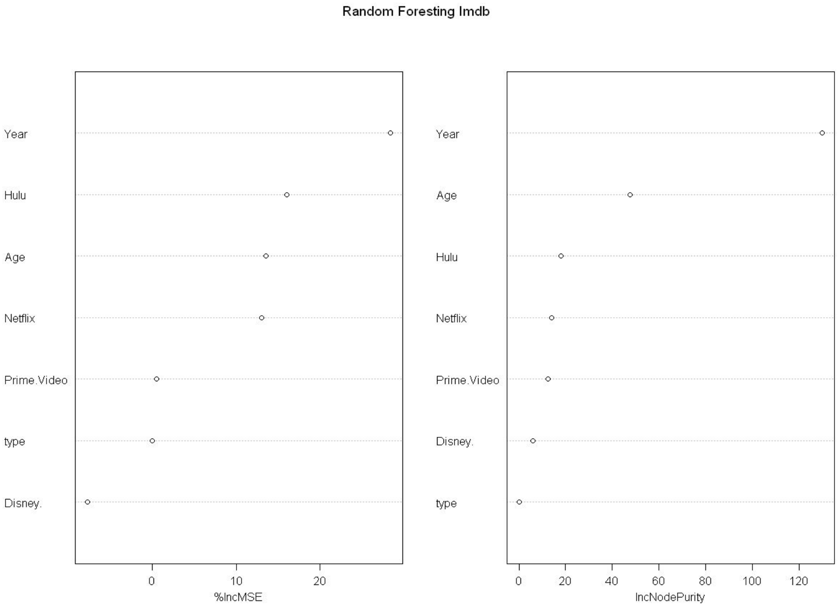
4.
365 shows predicted correctly
98% accuracy model (Disney +)
0/(0+7) = 0% Actual positive - Recall

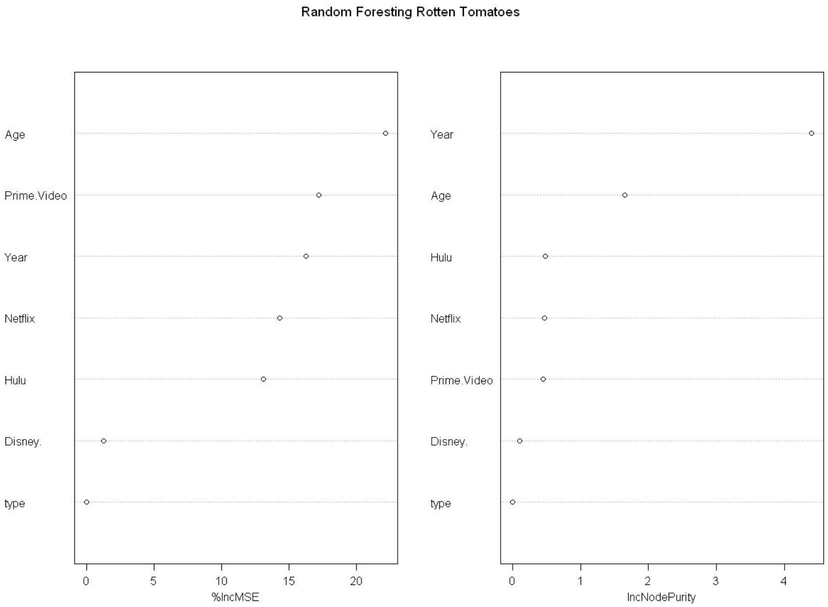
Random Forest Predictive Model

- Used R to create a random forest model (a collection of decision trees) to predict both IMDb and Rotten Tomatoes scores using Age, Year, and platform availability.
- Randomly selected rows into testing and training set with a 40/60 split
- Code(RandomForest Library):
  - Models:
    - Imdb.forest <- randomForest(IMDb~., data=train[,-4], mtry = 4, importance =T, na.action=na.omit)
    - Rotten.forest <- randomForest(Rotten.Tomatoes~., data=train[,-3], mtry = 4, importance =T, na.action=na.omit)
  - 4 variables randomly sampled as candidates at each split
  - importance of predictors to be assessed
  - 500 trees created(default)

The Model

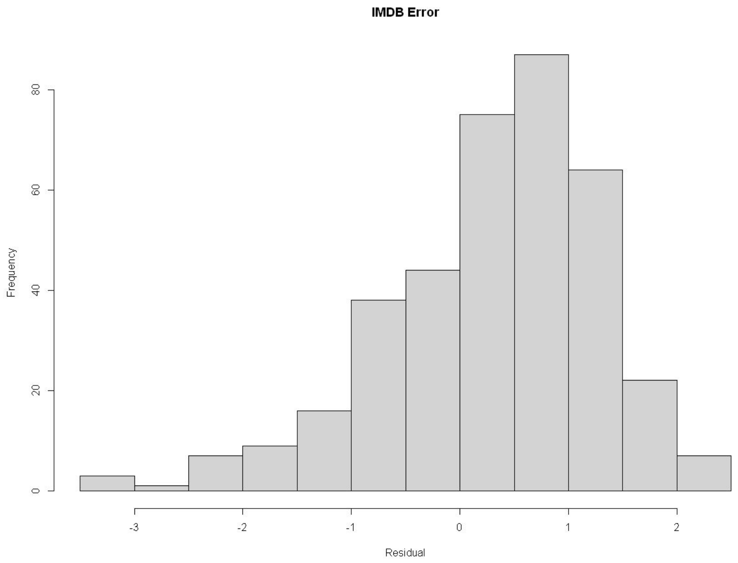
- %INCMSE = percent increase in mean squared error from variable being permuted
- IncNodePurity = finds the average split which has a high inter node 'variance' and a small intra node 'variance'



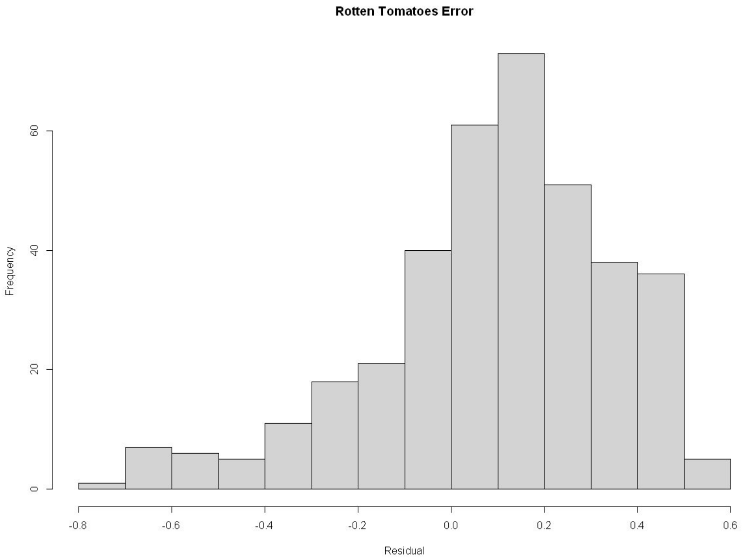


Results and Accuracy

- RMSE
  - IMDb = 1.036 stars
  - Rotten Tomatoes = 27.18%



- Average Residual
  - IMDb = .30112 stars
  - Rotten Tomatoes = 9.385%



Conclusion

- There is a positive correlation between both scores, which was expected
- We weren’t able to accurately predict the scores using the random forest model
  - Using RMSE: 1 star is a big difference in ratings, rotten tomatoes is worse as 27% is too big ignore
  - We cannot say at this time that there is a relationship between availability and rating
- From using KNN predictive models, we are able to see prediction accuracy for the 4 platforms, however a closer look at our data shows that high prediction does not necessarily measure accuracy.
  - Most shows are not on Disney+, but the model predicted the highest accuracy based on prevalence of non-present shows in a database.
  - Some streaming platforms have significantly more shows than others
  - Recall is consistently low between evaluations/predictions
  - This most likely occurred because of a small dataset and not much range of usable variables

Summary | Discussion | Questions?

1. Important to understand how the data was processed
   - Finding correlation between two variables shapes the data
   - Using factor data to identify variables by category (general clustering)
2. Not all predictive models are accurate
   - Low accuracy rate, not enough data, low correlation etc.
   - Other metrics can be more telling of a model than accuracy (recall)
3. Couldn’t effectively predict with either of our models - meaning that the variables we selected are not significant in prediction of platform availability
4. Real world this makes sense - many values in our dataset are locked due to exclusivity and not necessarily the ratings of the show itself