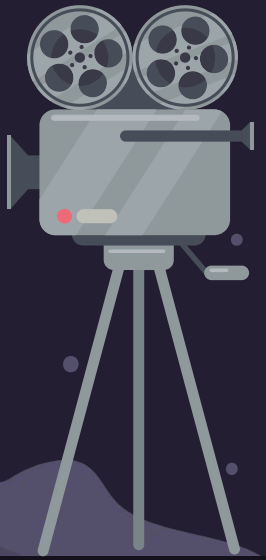
Predicting Box Office Success

A Data-Driven Approach to Movie Revenue Forecasting

By: Jagruta A., Mallika S., Alex S., and Vishwa P.



Is there a “formula for movie success”?

In 2023, the film industry generated an estimated **$77 billion**, highlighting its ever-growing popularity. With high stakes, studio bosses and producers constantly seek ways to ensure big box office returns. To assist, we're offering metadata on over 3,000 films from the Movie Database to help predict worldwide box office revenue. Data points include **cast, crew, keywords, budget, posters, release dates, languages, production companies, and countries**.

Credit to : https://www.behance.net/diegocarbonell

Exploratory Data Analysis: Data Summarization

```
$ id                    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30…
$ belongs_to_collection <chr> "[{'id': 313576, 'name': 'Hot Tub Time Machine Collection', 'poster_path': '/iEhb00TGPucF0b4joM1ieyY026U.jpg'…
$ budget                <int> 14000000, 40000000, 3300000, 1200000, NA, 8000000, 14000000, NA, NA, 6000000, 1000000, NA, 15000000, 53000000…
$ genres                <chr> "[{'id': 35, 'name': 'Comedy'}]", "[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, …
$ homepage              <chr> NA, NA, "http://sonyclassics.com/whiplash/", "http://kahaanithefilm.com/", NA, NA, "http://www.thepossessionm…
$ imdb_id               <chr> "tt2637294", "tt0368933", "tt2582802", "tt1821480", "tt1380152", "tt0093743", "tt0431021", "tt0391024", "tt01…
$ original_language     <chr> "en", "en", "en", "hi", "ko", "en", "en", "en", "en", "en", "en", "en", "en", "en", "en", "sr", "en", "en", "…
$ original_title        <chr> "Hot Tub Time Machine 2", "The Princess Diaries 2: Royal Engagement", "Whiplash", "Kahaani", "마린보이", "Pin…
$ overview              <chr> "When Lou, who has become the \"father of the Internet,\" is shot by an unknown assailant, Jacob and Nick fir…
$ popularity            <dbl> 6.575393, 8.248895, 64.299990, 3.174936, 1.148070, 0.743274, 7.286477, 1.949044, 6.902423, 4.672036, 14.77406…
$ poster_path           <chr> "/tQtWuwvMf0hCc2QR2tkolwl7c3c.jpg", "/w9Z7A0GHEhIp7etpj0vyKOeU1Wx.jpg", "/lIv1QinFqz4dlp5U4lQ6HaiskOZ.jpg", "…
$ production_companies  <chr> "[{'name': 'Paramount Pictures', 'id': 4}, {'name': 'United Artists', 'id': 60}, {'name': 'Metro-Goldwyn-Maye…
$ production_countries  <chr> "[{'iso_3166_1': 'US', 'name': 'United States of America'}]", "[{'iso_3166_1': 'US', 'name': 'United States o…
$ release_date          <chr> "2/20/15", "8/6/04", "10/10/14", "3/9/12", "2/5/09", "8/6/87", "8/30/12", "1/15/04", "2/16/96", "4/16/03", "1…
$ runtime               <int> 93, 113, 105, 122, 118, 83, 92, 84, 100, 91, 119, 98, 122, 118, 145, 97, 85, 111, 96, 87, 130, 95, 116, 92, 8…
$ spoken_languages      <chr> "[{'iso_639_1': 'en', 'name': 'English'}]", "[{'iso_639_1': 'en', 'name': 'English'}]", "[{'iso_639_1': 'en',…
$ status                <chr> "Released", "Released", "Released", "Released", "Released", "Released", "Released", "Released", "Released", "…
$ tagline               <chr> "The Laws of Space and Time are About to be Violated.", "It can take a lifetime to find true love; she's got …
$ title                 <chr> "Hot Tub Time Machine 2", "The Princess Diaries 2: Royal Engagement", "Whiplash", "Kahaani", "Marine Boy", "P…
$ Keywords              <chr> "[{'id': 4379, 'name': 'time travel'}, {'id': 9663, 'name': 'sequel'}, {'id': 11830, 'name': 'hot tub'}, {'id…
$ cast                  <chr> "[{'cast_id': 4, 'character': 'Lou', 'credit_id': '52fe4ee7c3a36847f82afae7', 'gender': 2, 'id': 52997, 'name…
$ crew                  <chr> "[{'credit_id': '59ac067c92514107af02c8c8', 'department': 'Directing', 'gender': 0, 'id': 1449071, 'job': 'Fi…
$ revenue               <int> 12314651, 95149435, 13092000, 16000000, 3923970, 3261638, 85446075, 2586511, 34327391, 18750246, 117235147,
```

The Dataset has 3,000 rows and 23 columns.
Revenue is our main variable of interest

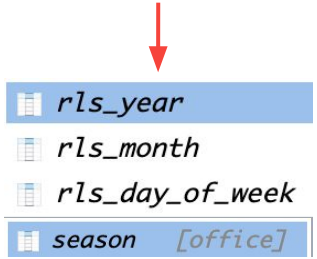
Exploratory Data Analysis: Data Cleaning

```
$ id                  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31…
$ collection_name     <chr> "Hot Tub Time Machine Collection", "The Princess Diaries Collection", NA, NA, NA, NA, NA, NA, "The Muppet Collect…
$ language            <chr> "en", "en", "en", "en", "ko", "en", "en", "ar", "en", "en", "en", "en", "en", "en", "en", "sr", "en", "en", "de",…
$ mgenre              <chr> "Comedy", "Comedy", "Drama", "Thriller", "Action", "Animation", "Horror", "Documentary", "Action", "Comedy", "Dra…
$ main_cast           <chr> "Rob Corddry", "Anne Hathaway", "Miles Teller", "Vidya Balan", "Kim Kang-woo", "Scott Grimes", "Jeffrey Dean Morg…
$ prod_country_code   <chr> "US", "US", "US", "IN", "KR", NA, "US", NA, "US", "US", "US", "US", "US", "US", "US", "RS", "US", "GB", "AT", "FR…
$ prod_company        <chr> "Paramount Pictures", "Walt Disney Pictures", "Bold Films", NA, NA, NA, "Ghost House Pictures", NA, "Walt Disney …
$ budget              <int> 14000000, 40000000, 3300000, 1200000, 500000, 8000000, 14000000, 400000, 14000000, 6000000, 1000000, 6000000, 150…
$ homepage            <dbl> 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0,…
$ imdb_id             <chr> "tt2637294", "tt0368933", "tt2582802", "tt1821480", "tt1380152", "tt0093743", "tt0431021", "tt0391024", "tt011711…
$ original_title      <chr> "Hot Tub Time Machine 2", "The Princess Diaries 2: Royal Engagement", "Whiplash", "Kahaani", "마린보이", "Pinocch…
$ overview            <chr> "When Lou, who has become the \"father of the Internet,\" is shot by an unknown assailant, Jacob and Nick fire up…
$ popularity          <dbl> 6.575393, 8.248895, 64.299990, 3.174936, 1.148070, 0.743274, 7.286477, 1.949044, 6.902423, 4.672036, 14.774066, 1…
$ poster_path         <chr> "/tQtWuwvMf0hCc2QR2tkolwl7c3c.jpg", "/w9Z7A0GHEhIp7etpj0vyKOeU1Wx.jpg", "/lIv1QinFqz4dlp5U4lQ6HaiskOZ.jpg", "/aTX…
$ release_date        <chr> "2/20/15", "8/6/04", "10/10/14", "3/9/12", "2/5/09", "8/6/87", "8/30/12", "1/15/04", "2/16/96", "4/16/03", "11/21…
$ runtime             <int> 93, 113, 105, 122, 118, 83, 92, 84, 100, 91, 119, 98, 122, 118, 145, 97, 85, 111, 96, 87, 130, 95, 116, 92, 87, 9…
$ status              <chr> "Released", "Released", "Released", "Released", "Released", "Released", "Released", "Released", "Released", "Rele…
$ tagline             <chr> "The Laws of Space and Time are About to be Violated.", "It can take a lifetime to find true love; she's got 30 d…
$ title               <chr> "Hot Tub Time Mac                                                                                              Bo
$ Keywords            <chr> "[{'id': 4379, 'n                                                                                              '},
$ cast                <chr> "[{'cast_id': 4,                                                                                               7,
$ crew                <chr> "[{'credit_id': '                                                                                              o
$ revenue             <int> 12314651, 9514943                                                                                              351
$ budget_imp          <lgl> FALSE, FALSE, FAL                                                                                              FALSE,
$ collection_num      <dbl> 1, 1, 0, 0, 0, 0,                                                                                              1, 0,
$ language_num        <dbl> 1, 1, 1, 1, 0, 1,                                                                                              0, 1,
```

Cleaned data and extracted useful information from the columns – **belongs_to_collection, spoken_languages, genres, cast, production_countries, production_companies.** We converted certain columns to boolean to simplify information extraction and facilitate regression modeling.

Feature Engineering

```
> head(office$release_date)
[1] "2/20/15"  "8/6/04"   "10/10/14" "3/9/12"
```



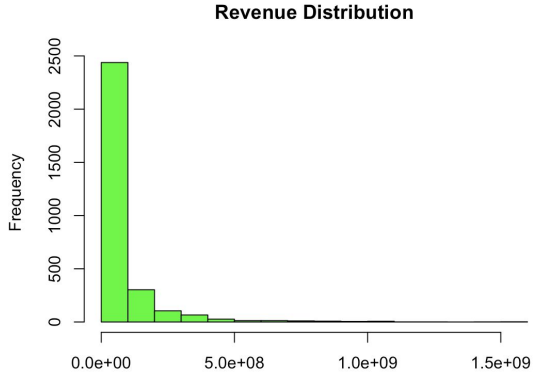

```
> head(office$runtime_category)
[1] Medium Medium Medium Long   Medium Short
Levels: Short Medium Long

> head(office$popularity_category)
[1] Low      Medium   High     Very Low Very Low Very Low
Levels: Very Low Low Medium High
```

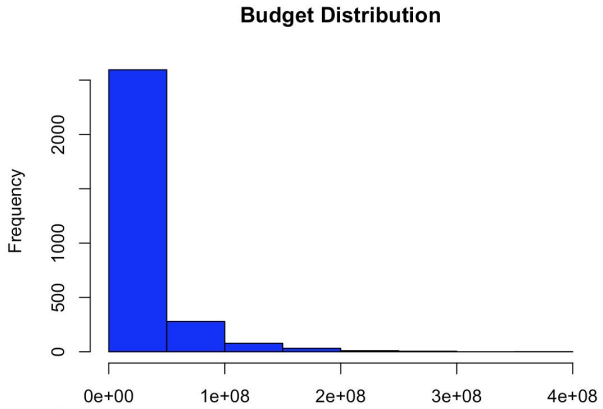
```
> head(office$main_cast)
[1] "Rob Corddry"   "Anne Hathaway" "Miles Teller
> head(office$crew)
[1] "[{'credit_id': '59ac067c92514107af02c8c8', '
```

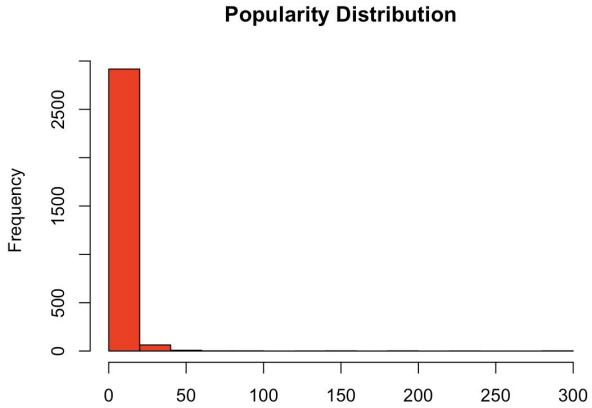
```
> print(cols_with_cast_or_crew)
 [1] "cast"               "crew"               "main_cast"          "num_cast"           "num_crew"
 [6] "unknowngender_cast" "female_cast"        "male_cast"          "unknowngender_crew" "female_crew"
[11] "male_crew"
```

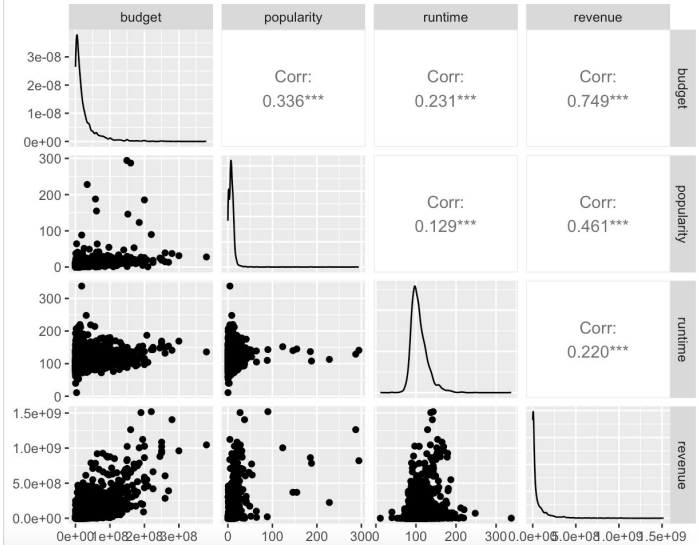
Leveling on categorical columns, Extraction on release dates, cast/crew split on gender and Number, Bool values for binary categorical data, knn impute for budget.

Univariate Data Visualisation



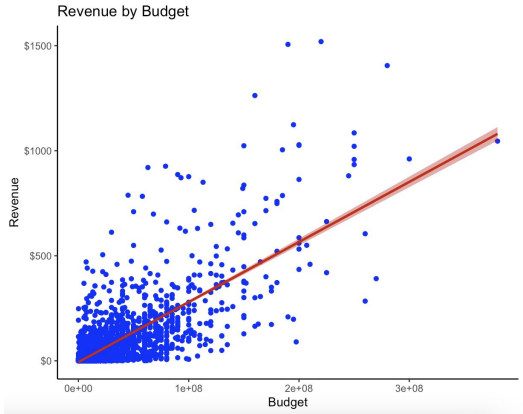


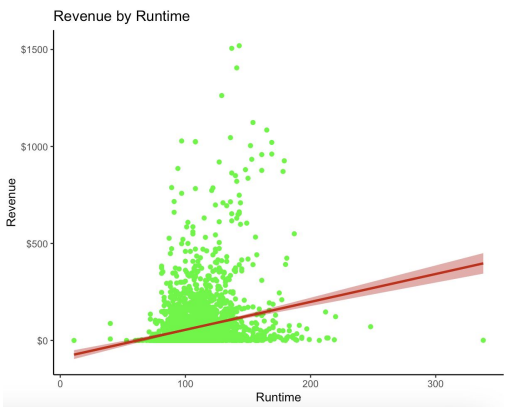


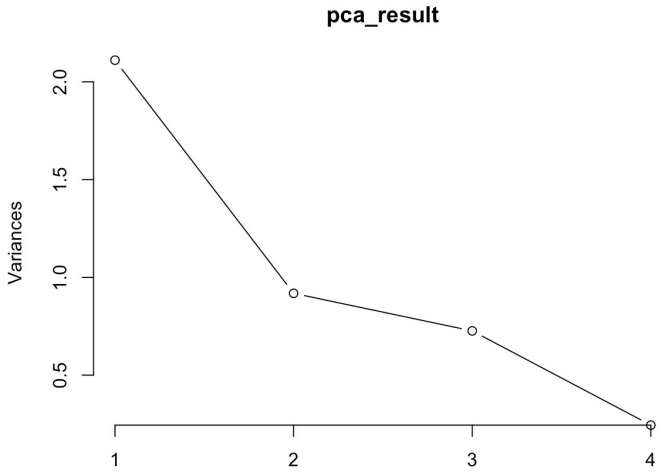


During the Univariate analysis of **revenue, budget** and **popularity** fields. We found that these variables are "**highly right skewed**". The correlation matrix indicates that the correlation of budget and popularity with revenue is higher.

Bivariate/Multivariate Data Visualisation



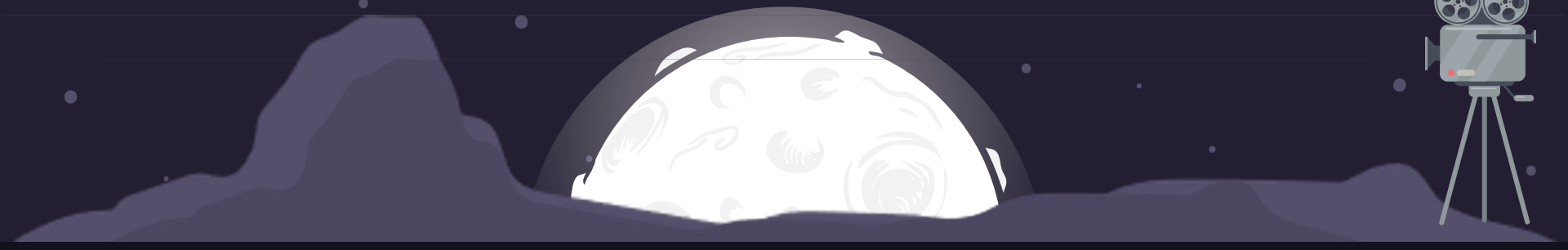




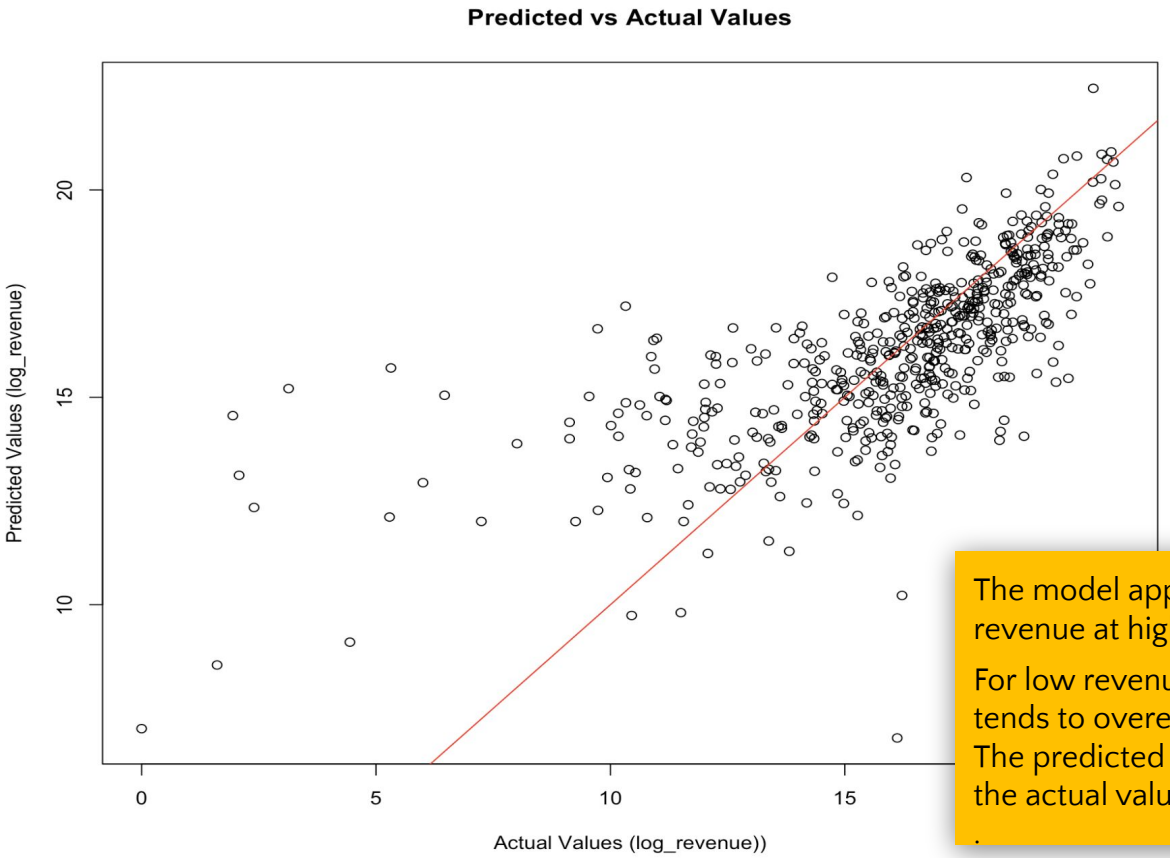
During the Bivariate analysis of **revenue, budget, runtime** and **popularity** fields. We found that budget and popularity are "**positively correlated**" and are better predictors of revenue. While **Principal Component Analysis** Scree Plot of these 4 columns results also suggest the same.

Models We Tried

- Linear Regression
- Bagging
- Random Forest
- Boosting



Multiple Linear Regression

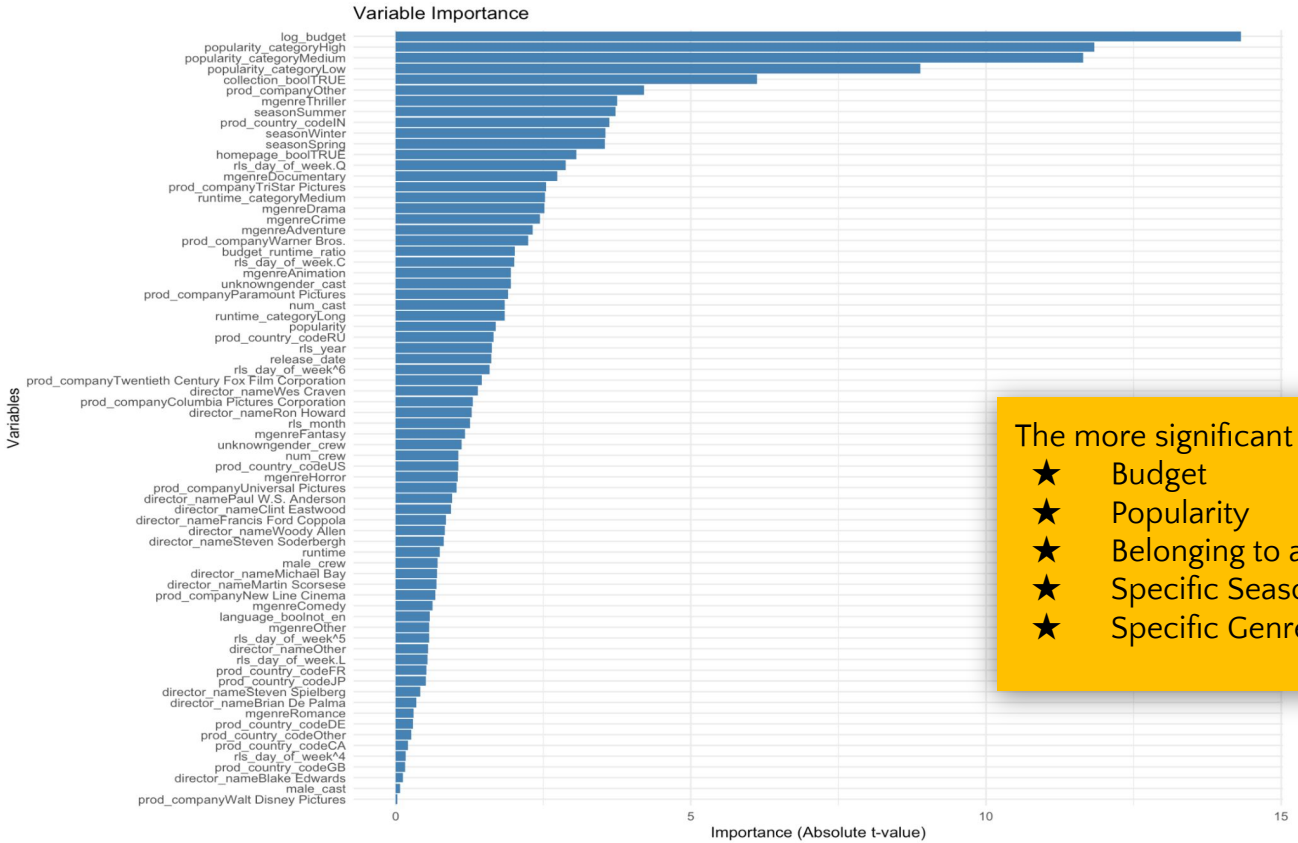


The model appears to better predict revenue at higher values.

For low revenue movies, the model tends to overestimate the revenue. The predicted values are higher than the actual values.

.

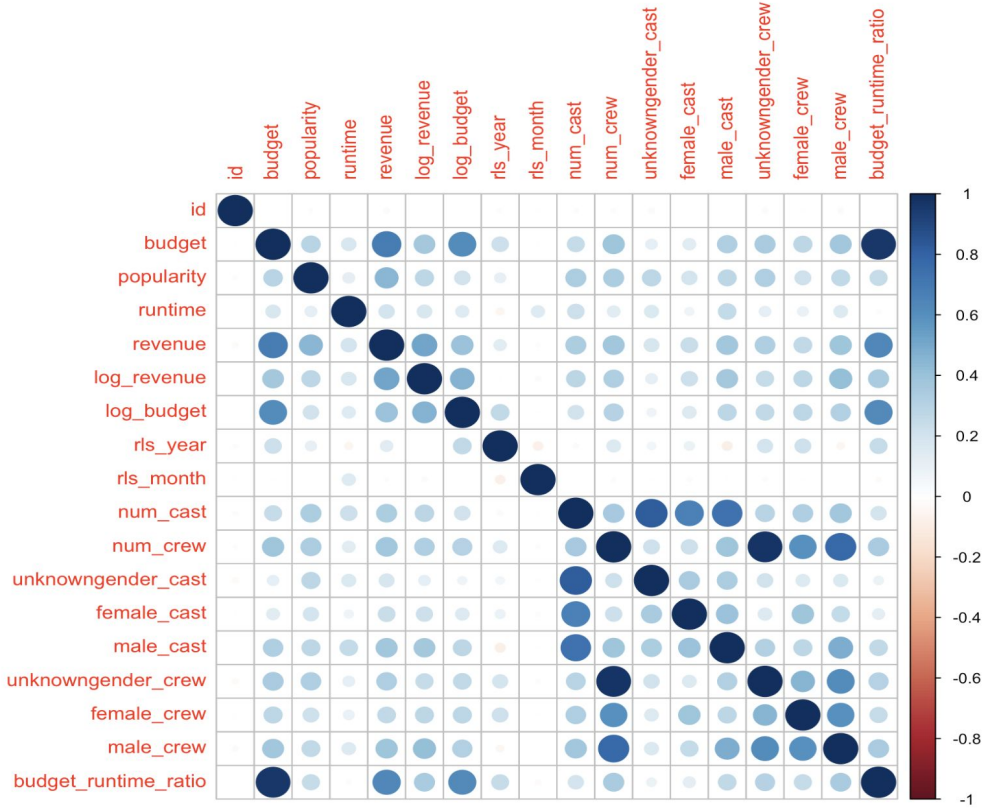
Linear Regression: importance()



The more significant variables are as follows:

- ★ Budget
- ★ Popularity
- ★ Belonging to a collection
- ★ Specific Seasons: Winter, Spring
- ★ Specific Genres: Thriller

Linear Regression: corrplot()



Key (numeric) variables that appear correlated:

- ★ Budget and revenue
- ★ Crew count and budget

Surprising Finds

- ★ Revenue does not appear correlated with numbers of cast and crew working on a movie
- ★ There is a slight correlation between release month and runtime (coincidence?)

Linear Regression: Final looks

Popularity, **Collection Status** and **Budget** were the most significant predictors for **Revenue**.

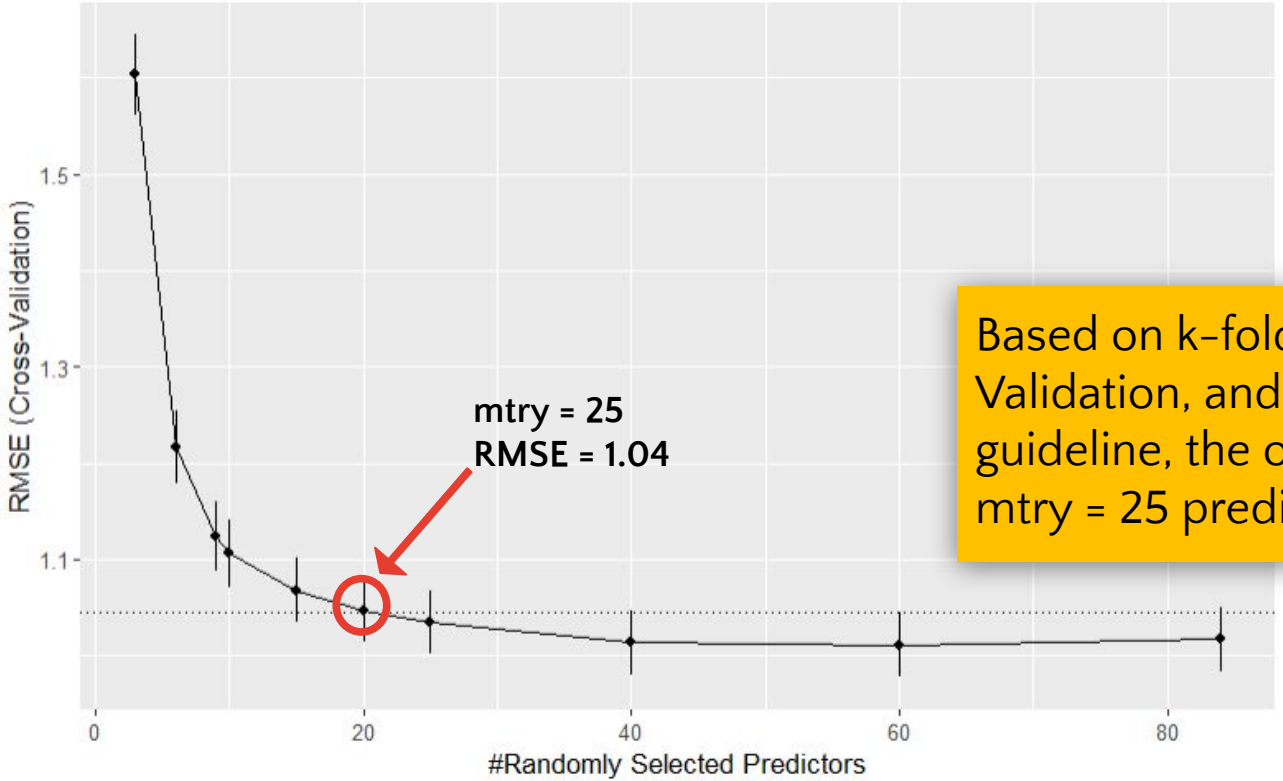
A movie in the 'High' popularity_category can expect an increase of close to 200% in Revenue.

A movie that belongs to a collection can expect an increase of about 75% in Revenue.

There would be a Revenue increase of 53% when the budget is increased.

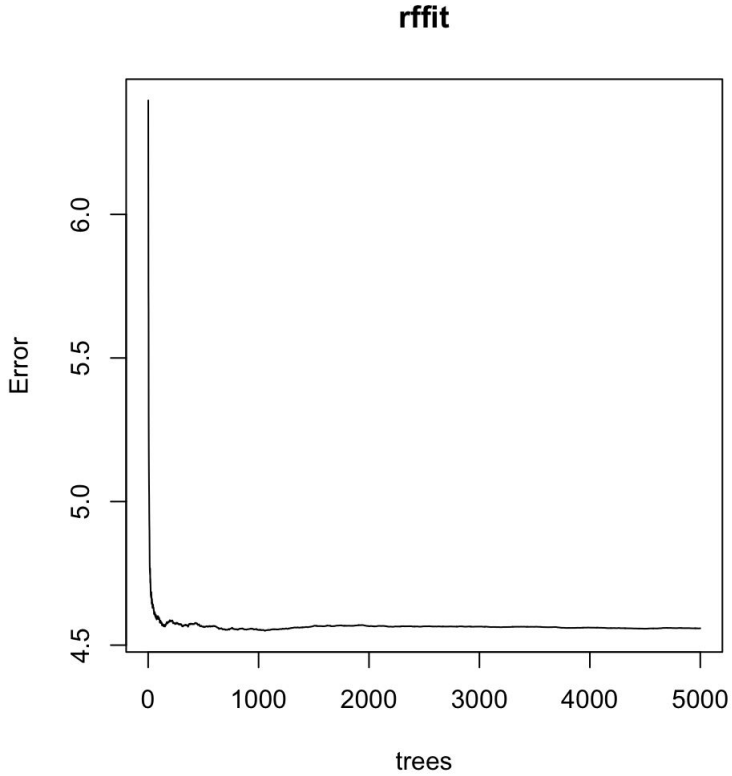
The final RSME for the Multiple Linear Regression was 2.194028.
We sought to improve this value with other methods.

Random Forest: Tuning Number of Variables



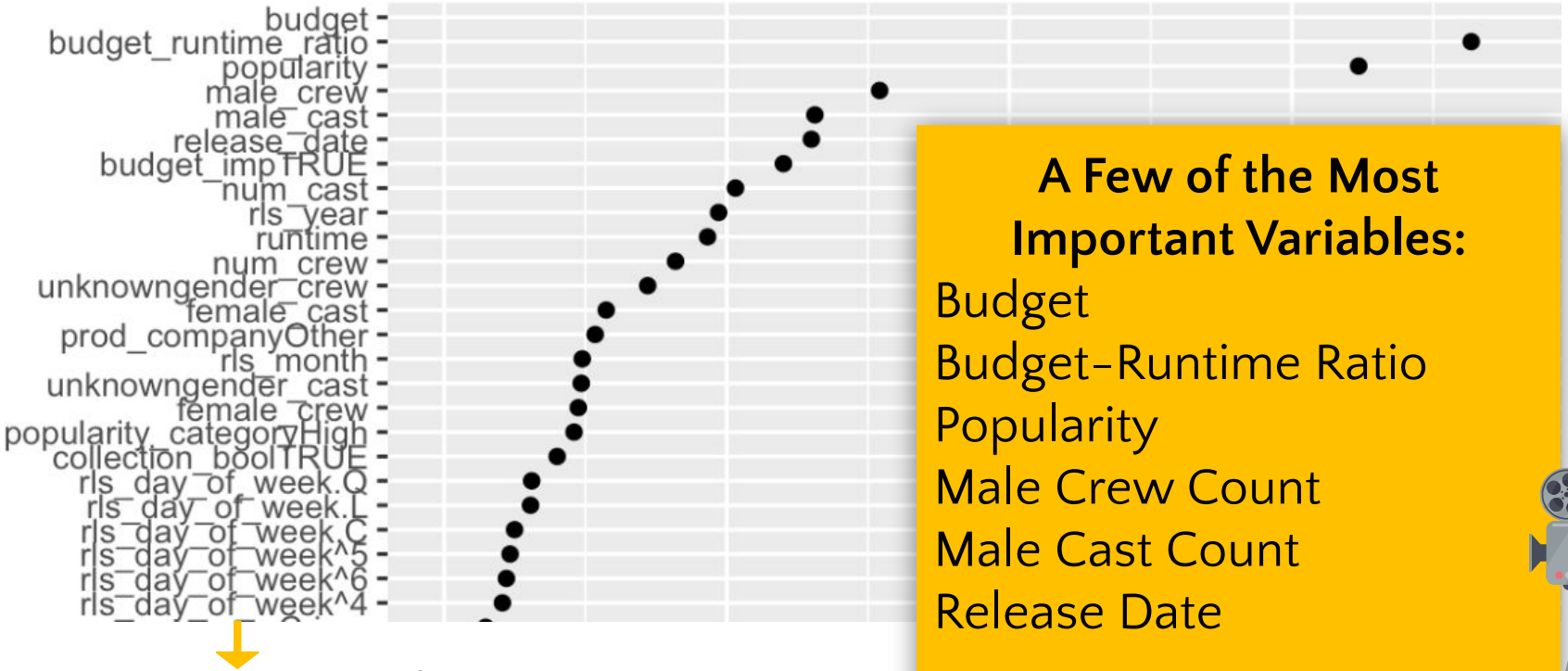
Based on k-fold Cross Validation, and using the 1 SD guideline, the optimal value is mtry = 25 predictors.

Random Forest: Number of Trees



As expected, roughly 500 was the optimal number of trees at which the RMSE evened out.

Random Forest: Variable Importance

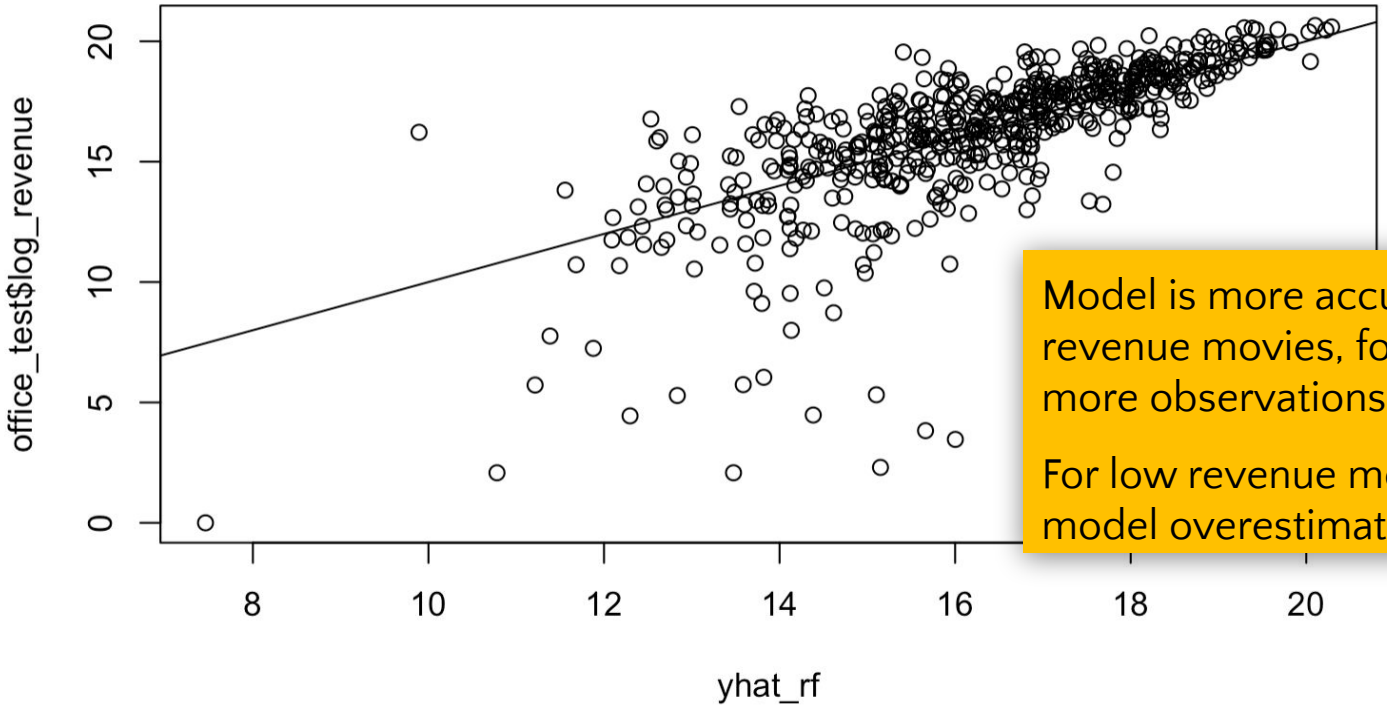


80+ covariates created by feature engineering

A Few of the Most Important Variables:

Budget
Budget-Runtime Ratio
Popularity
Male Crew Count
Male Cast Count
Release Date

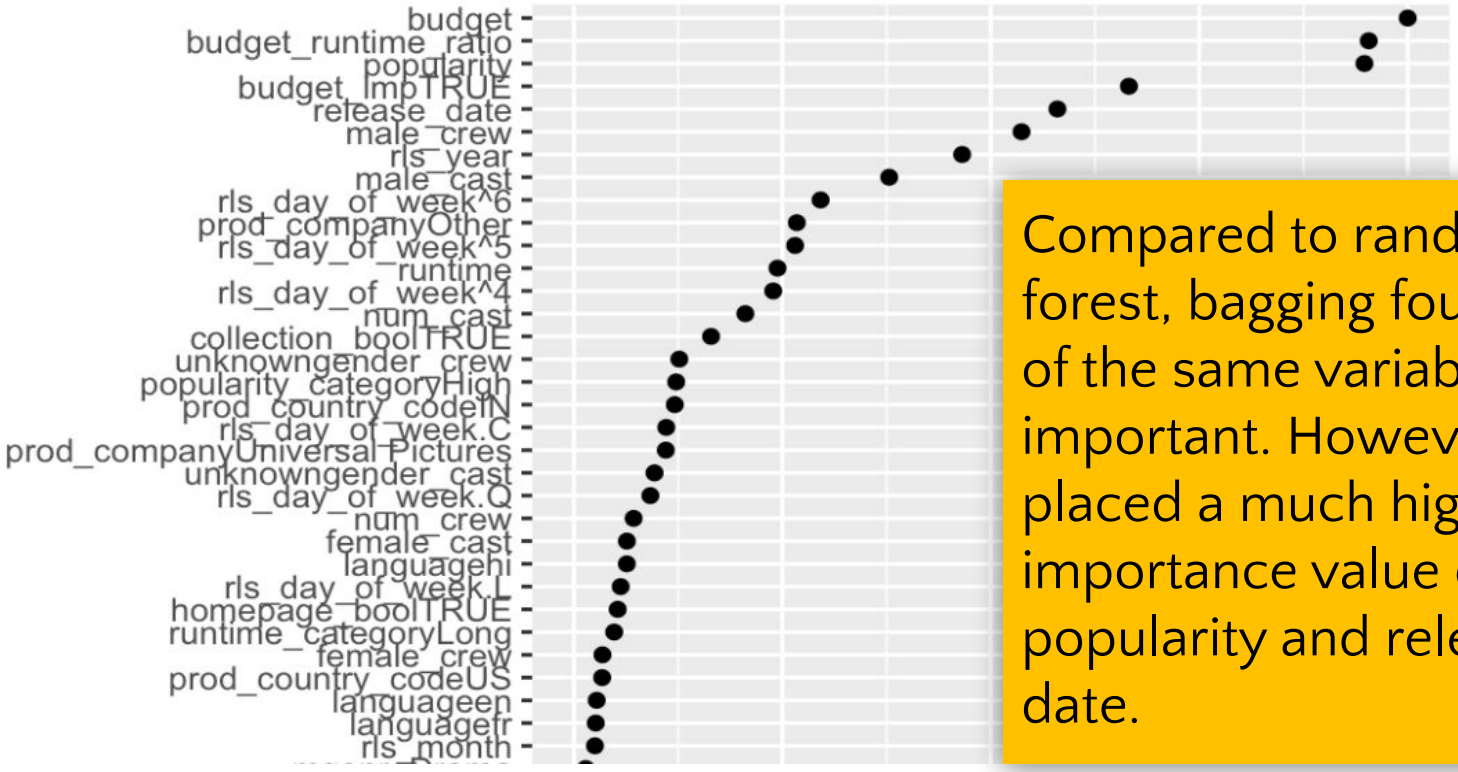
Random Forest: Comparing Predictions to Actuals Using Test Set



Model is more accurate on high revenue movies, for which we had more observations.

For low revenue movies, the model overestimates y hat.

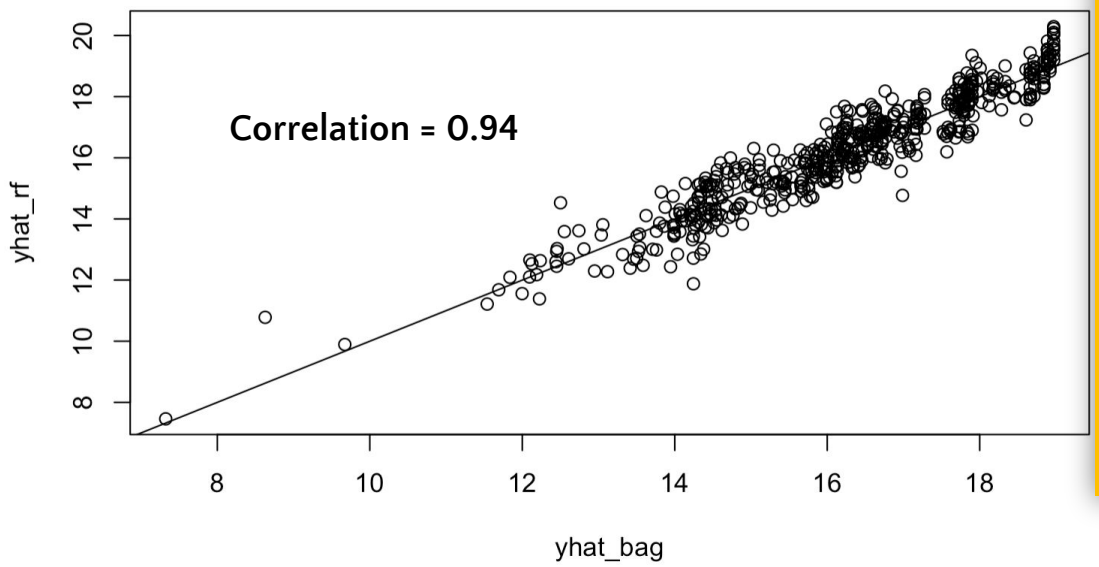
Bagging: Variable Importance



Compared to random forest, bagging found most of the same variables to be important. However, it placed a much higher importance value on popularity and release date.

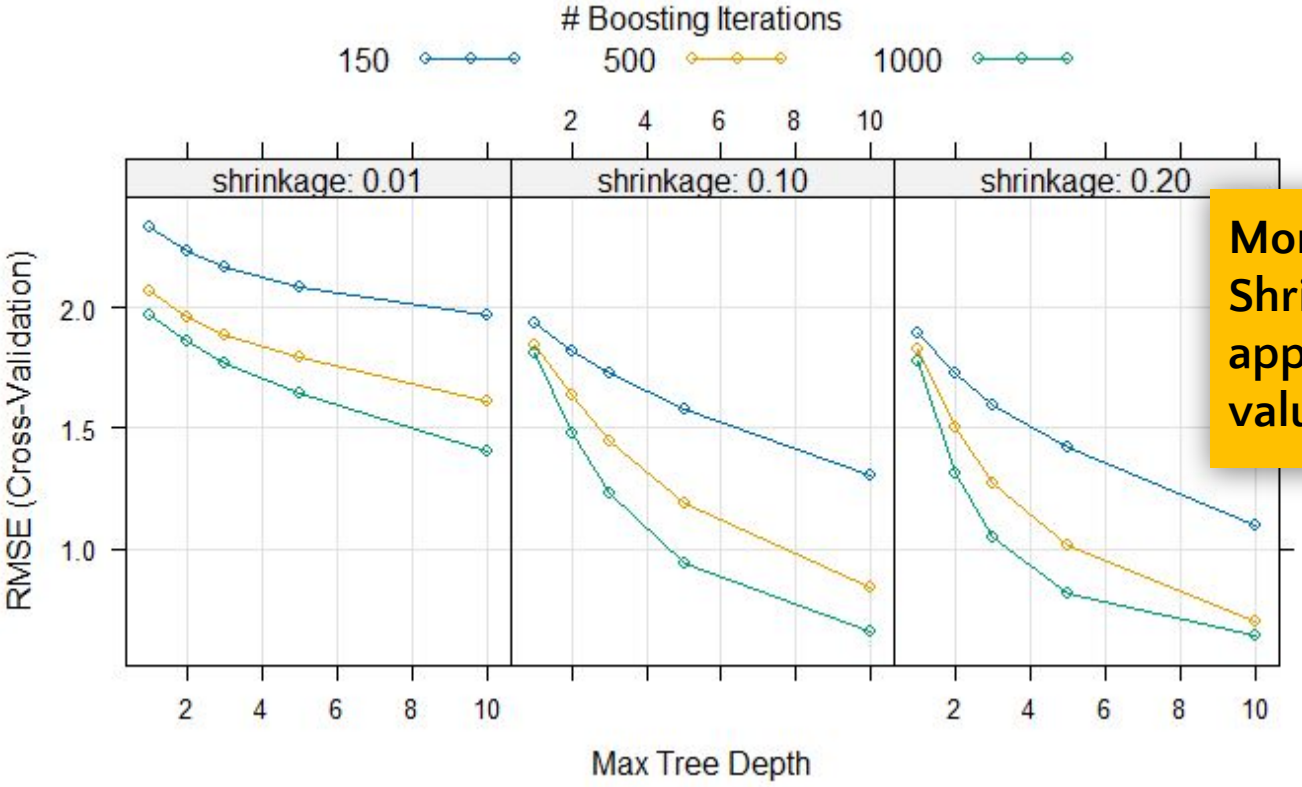
Comparing Similar Methods: Random Forest vs. Bagging

Random Forest Test RMSE: 2.07 (train = 1.04)
Bagged Test RMSE: 2.19 (train = 2.15)



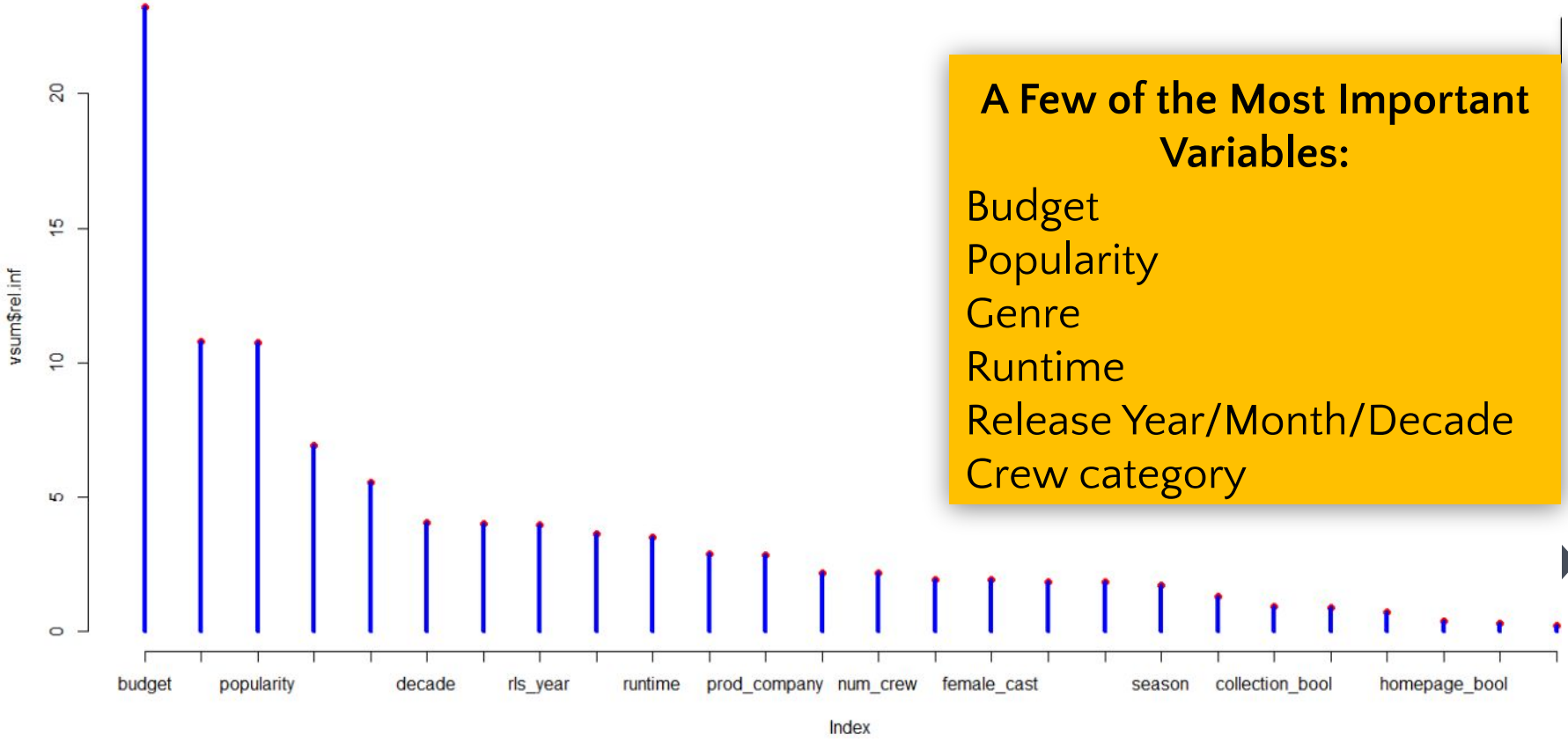
Overall:
Based on RMSE, random forest performs better than bagging on the training set, but about the same on the test set.
This indicates probable overfitting for random forest and need for improvement on feature engineering and tuning parameters

Boosting Iterations and Shrinkage vs RMSE



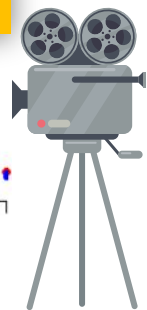
More Boosting iterations and Shrinkage at 0.2 and 0.1 appeared to gives lesser RMSE values

Boosting: Variable Importance

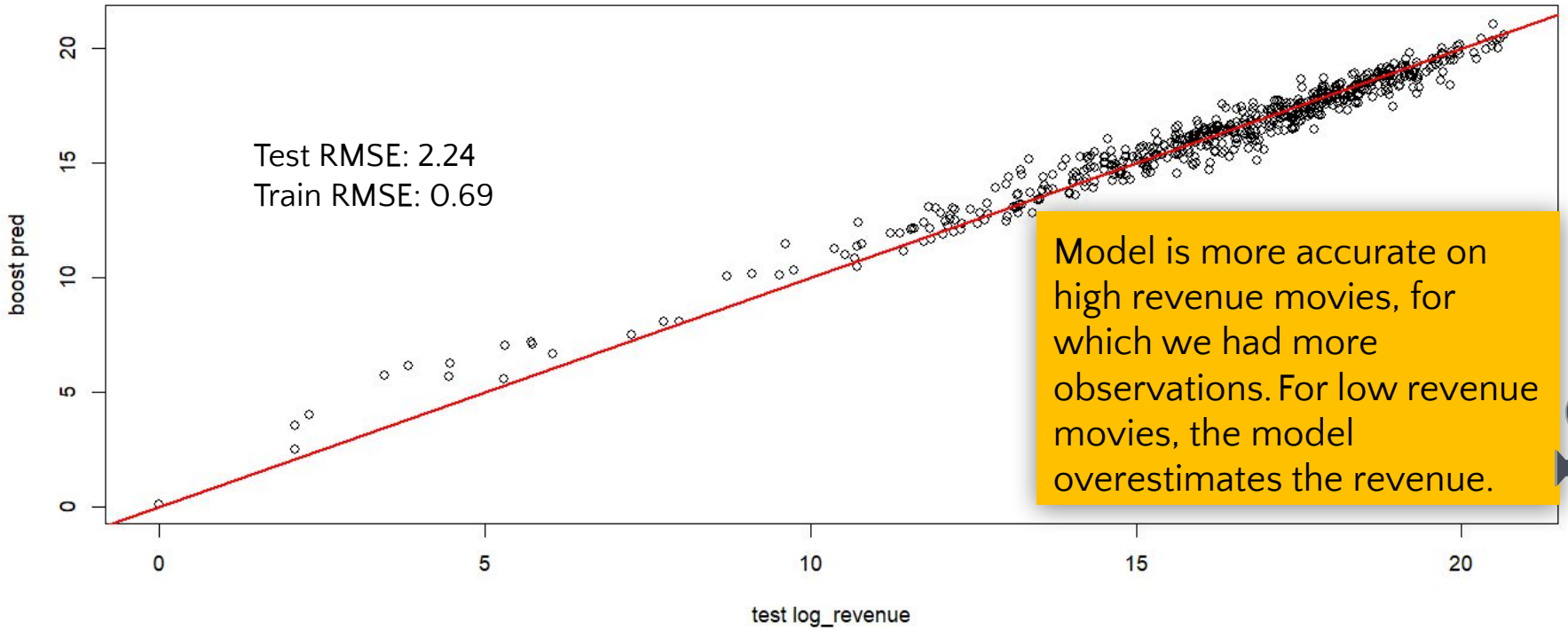


A Few of the Most Important Variables:

Budget
Popularity
Genre
Runtime
Release Year/Month/Decade
Crew category



Boosting: Comparing Predictions to Actuals Using Test Set



Model is more accurate on high revenue movies, for which we had more observations. For low revenue movies, the model overestimates the revenue.

Choosing the Final Model

MODEL	MSE(test)	RMSE(test)	MSE(training)	RMSE (training)
Linear Regression	4.83	2.19	4.15	2.03
Bagging	4.76	2.19	4.61	2.15
Random Forest	**4.27**	**2.07**	**1.08**	**1.04**
Boosting	5.02	2.24	0.47	0.69

Learnings and Improvements

- *What we learned:*
  - *We have observed some overfitting in our models that lead to a higher test RSME*
  - *Feature importance helped us see which variables had a larger impact on predicting revenue*
  - *There is a lot of trial and error in terms of understanding what helps the model predict to its best ability*
- *Given more time, we would incorporate:*
  - *Additional data, we felt our dataset was on the smaller side*
  - *Better/more accurate method for filling the missing budget values – possibly using API or web scraping*
  - *Additional work on feature selection and engineering*

Thank you