MOVIE BOX OFFICE PREDICTIVE MODELING

CAPSTONE FINAL PRESENTATION

Data Sets

Kaggle - The Movies Dataset - (cast, crew, genres, user ratings)
https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

The Movie DB - (revenue, budget)
https://www.themoviedb.org/

Box Office Mojo - (Domestic Box Office sales)
https://www.boxofficemojo.com/

Feature Engineering

Example of messy data provided for each movie

Cast

```
{'id': 31, 'name': 'Tom Hanks'},
{'id': 12898, 'name': 'Tim Allen'},
{'id': 7167, 'name': 'Don Rickles'},
{'id': 12899, 'name': 'Jim Varney'},
{'id': 12900, 'name': 'Wallace Shawn'},
```

Keywords

```
{'id': 931, 'name': 'jealousy'},
{'id': 4290, 'name': 'toy'},
{'id': 5202, 'name': 'boy'},
{'id': 6054, 'name': 'friendship'},
{'id': 9713, 'name': 'friends'},
{'id': 9823, 'name': 'rivalry'},
```

Crew

```
{'id': 7879, 'job': 'Director', 'name': 'John Lasseter'},
{'id': 12891, 'job': 'Screenplay', 'name': 'Joss Whedon'}
{'id': 7, 'job': 'Screenplay', 'name': 'Andrew Stanton'},
{'id': 12892, 'job': 'Screenplay', 'name': 'Joel Cohen'},
{'id': 12893, 'job': 'Screenplay', 'name': 'Alec Sokolow'
{'id': 8, 'job': 'Editor', 'name': 'Lee Unkrich'},
{'id': 1168870, 'job': 'Editor', 'name': 'Robert Gordon'}
```

Feature Engineering

How to measure the quality of each cast member, crew member, etc?

<u>Measures of film quality in the dataset:</u>

Popularity

mean	8.471042
std	12.082205
min	0.000657
25%	3.900239
50%	7.358204
75%	10.829195
max	547.488298

- No description on Kaggle
- No apparent cap on range
- Not clear how this is determined

Vote Average

mean	6.202871
std	0.995378
min	0.000000
25%	5.600000
50%	6.300000
75%	6.900000
max	10.000000

- User generated
- Ranges from 0-10
- Paired with Vote Count

Feature Engineering

Process:

1) Extract Ids for each set, Ex:

	title	Ids_Cast	Ids_Director	Ids_Screenplay	Ids_Editor	Ids_Keywords	genreIds
0	Toy Story	[31, 12898, 7167, 12899, 12900, 7907, 8873, 11...	[7879]	[12891, 7, 12892, 12893]	[8, 1168870]	[931, 4290, 5202, 6054, 9713, 9823, 165503, 17...	[16, 35, 10751]

2) Calculate quality of each id in a set using various methods
   - Past Average
   - Past Vote Average
   - Past Historic Average
   - All Movie Average
   - All Movie Vote Average

3) Average over all ids in a set for each method to create features for that set

Feature Engineering

Cast Specific Feature Generation (set average):

- Total of Averaged Ratings
- Top 3 Averaged Ratings
- Ranked Average Ratings

Feature Engineering

Summary of the 42 Features generated:

	Average Rating	Vote Average	Historic Average	All Movie Average	All Movie Vote Average
Actors	X	X	X	X	X
Directors	X	X	X	X	X
Screenwriters	X	X	X	X	X
Editors	X	X	X	X	X
Genres	X				
Keywords	X				

Cast Specific Features

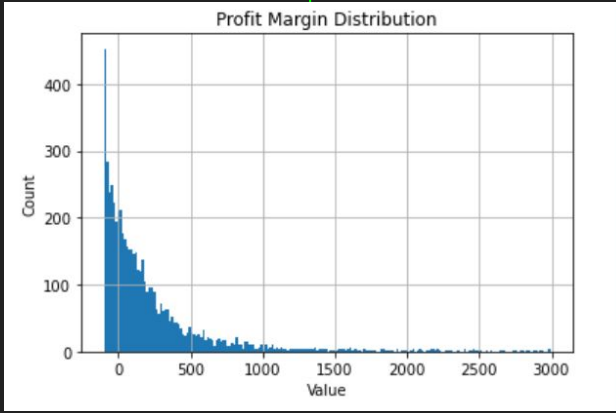
	Total	Top 3	Ranked Average	Ranked Total
Average Rating	X	X	X	X
Vote Average	X	X	X	X
Historic Average	X	X	X	X
All Movie Average	X	X	X	X
All Movie Vote Average	X	X	X	X

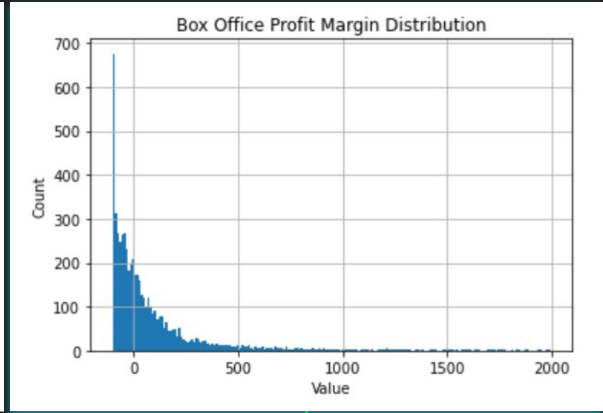
Modeling

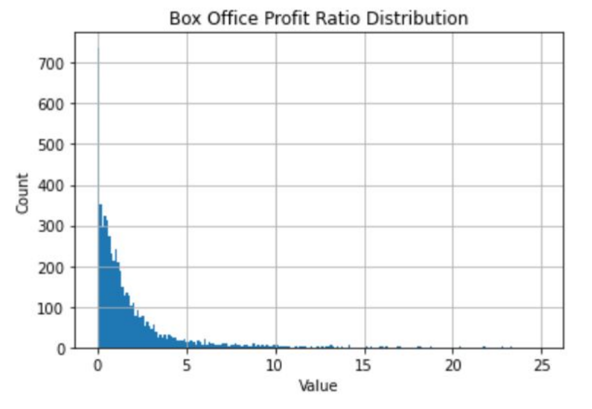
Independent Feature Selection

$$Profit\ Margin = \frac{(Revenue - Budget)}{Budget}$$

$$Box\ Office\ Profit\ Ratio = \frac{Domestic\ Box\ Office\ Earnings}{Budget}$$





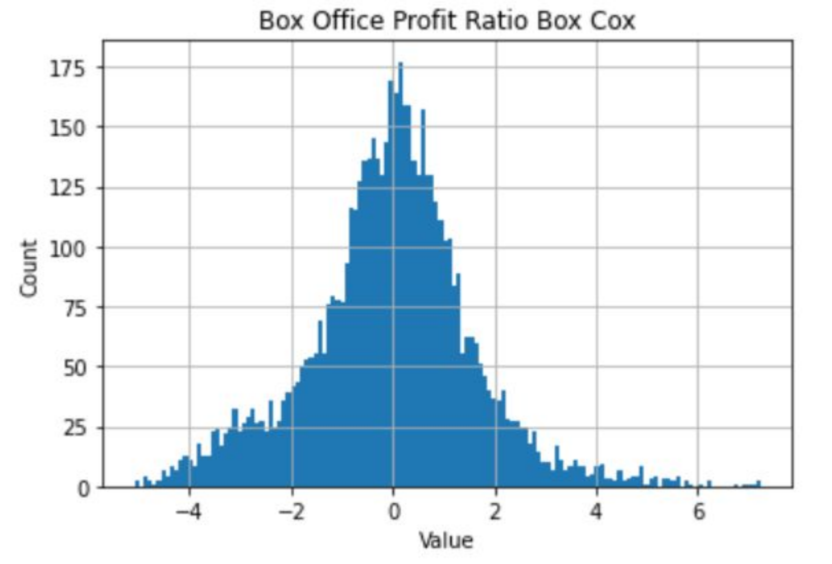


$$Box\ Office\ Profit\ Margin = \frac{(Domestic\ Box\ Office\ Earnings - Budget)}{Budget}$$

Modeling

Independent Feature Selection

Scipy.Stats.BoxCox - Power transformation, requires positive values



Modeling - Base Regression Model Performances

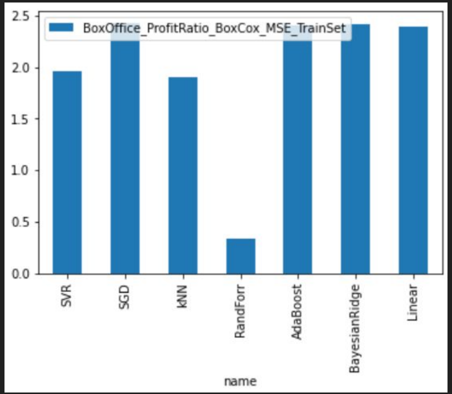
Models Tested

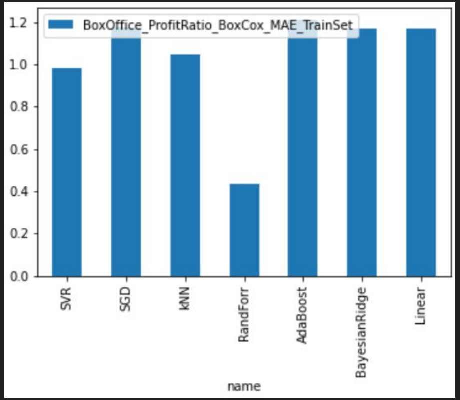
- Epsilon-Support Vector Regression (SVR)
- Stochastic Gradient Descent Regression (SGD)
- K Nearest Neighbors Regressor (kNN)
- Random Forest Regressor (RandForr)
- Ada Boost Regressor (AdaBoost)
- Bayesian Ridge Regressor (BayesianRidge)
- Linear Regression (Linear)

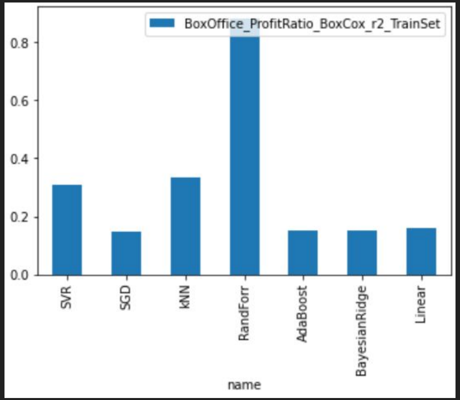
Model Performance Metrics

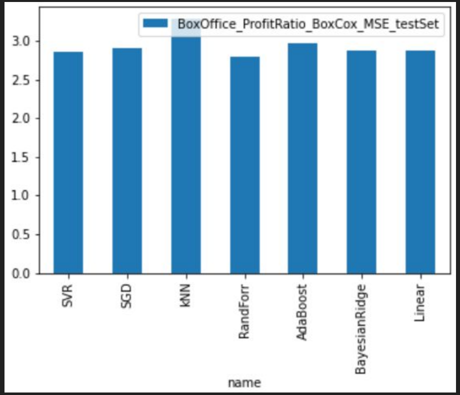
- Mean Squared Error (MSE)
- Mean Average Error (MAE)
- R^2 Score (r2)

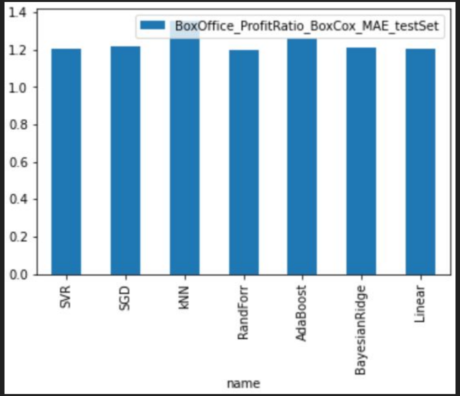
Modeling - Base Regression Model Performances

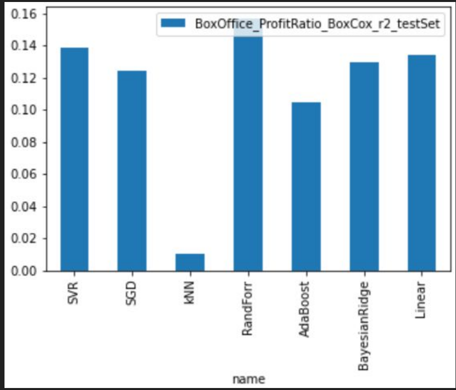












Modeling - Optimization

Two rounds of Bayesian Optimization for Random Forest Model

- Each round using 10 fold cross validation
- Reporting out median of cross validation performance

Round 1:

- 8 metaparameters tested
- Range set around default
- 80 iterations

Round 2:

- 3 metaparameters tested
- Range set around values of Round 1
- 50 iterations

Modeling - Optimization Results

Round 1:

Ranges

n_estimators:	(90 - 110)
max_depth:	(40 - 60)
min_samples_split:	(2 - 10)
min_samples_leaf:	(1 - 10)
min_weight_fraction_leaf:	(0.0 - 0.5)
min_impurity_decrease:	(0.0 - 0.2)
ccp_alpha:	(0.0 - 0.2)
max_features:	(24 - 44)

Values

n_estimators:	101
max_depth:	51
min_samples_split:	4
min_samples_leaf:	5
min_weight_fraction_leaf:	0.0
min_impurity_decrease:	0.0
ccp_alpha:	0.0
max_features:	25

Best Score

0.167

Round 2:

Ranges

n_estimators:	(125 - 250)
max_depth:	(37 - 55)
max_features:	(10 - 40)

Values

n_estimators:	160
max_depth:	39
max_features :	38

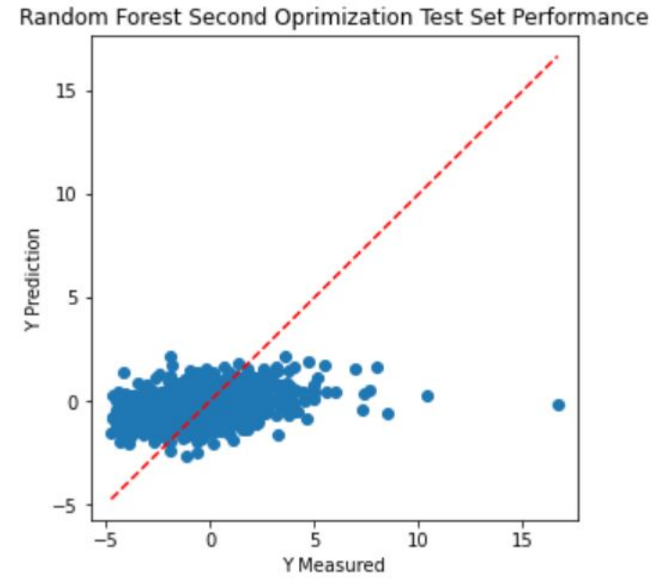
Best Score

0.161

Modeling - Optimization Results

Optimized Model (Round 2) Output


```
MSE : 0.3343961152603857
MAE : 0.4316499854236619
R^2 : 0.882417335490463
```



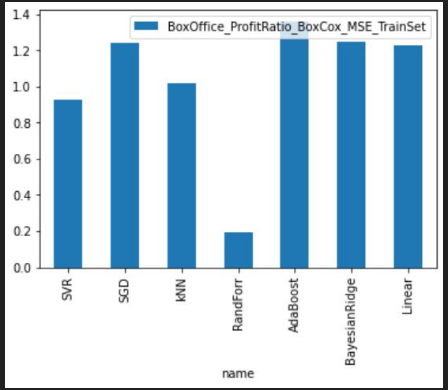

```
MSE : 2.762038502829518
MAE : 1.1942630273074142
R^2 : 0.16630947367785076
```

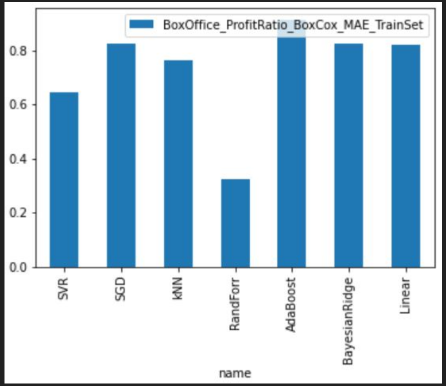
Modeling - Including Production Companies

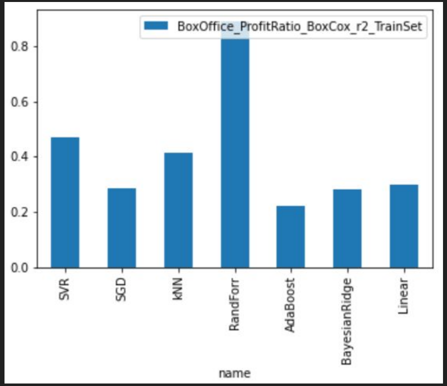
Changes to the Data

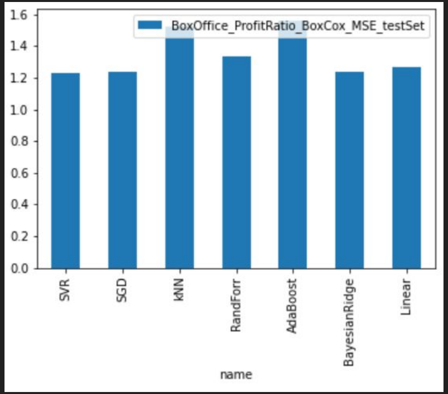
- Added Feature - First Production Company Listed
- Limit dataset to those in top 60 most frequent Production Companies
- Use one hot encoding to add as a categorical variable

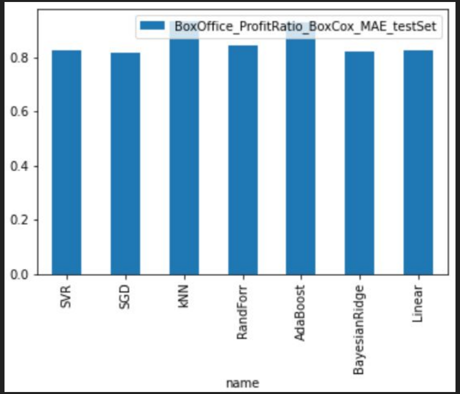
Modeling - Production Companies Base Model Performance

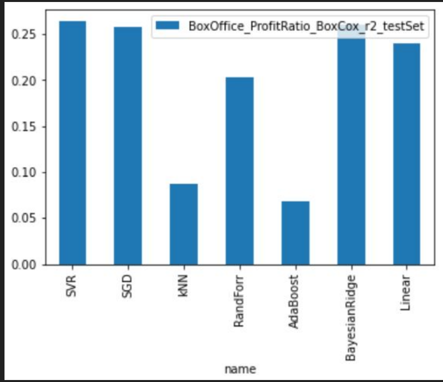












Conclusions

- The features (engineered and original) are not good predictors of the Box Office performance
- The produced models at best overfit the Training set, and cannot generalize to the Test Set
- 42 features engineered are all based on film user rating, a biased measure that does not correlate to film performance, i.e. garbage in, garbage out

Questions?