Data Science Meets Hollywood

Jit Seneviratne

Predicting Revenue for Movies

Can Cumulative Worldwide Box Office Revenue be Predicted from Available Quantitative and Qualitative Information?

Pipeline for Data

Budget

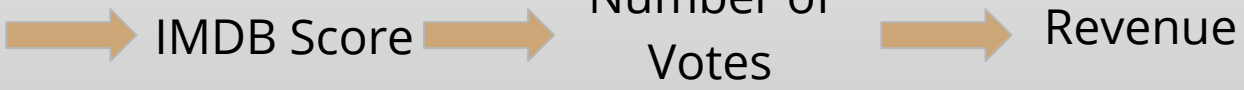
Actor

Actress

Director

Genre

Runtime



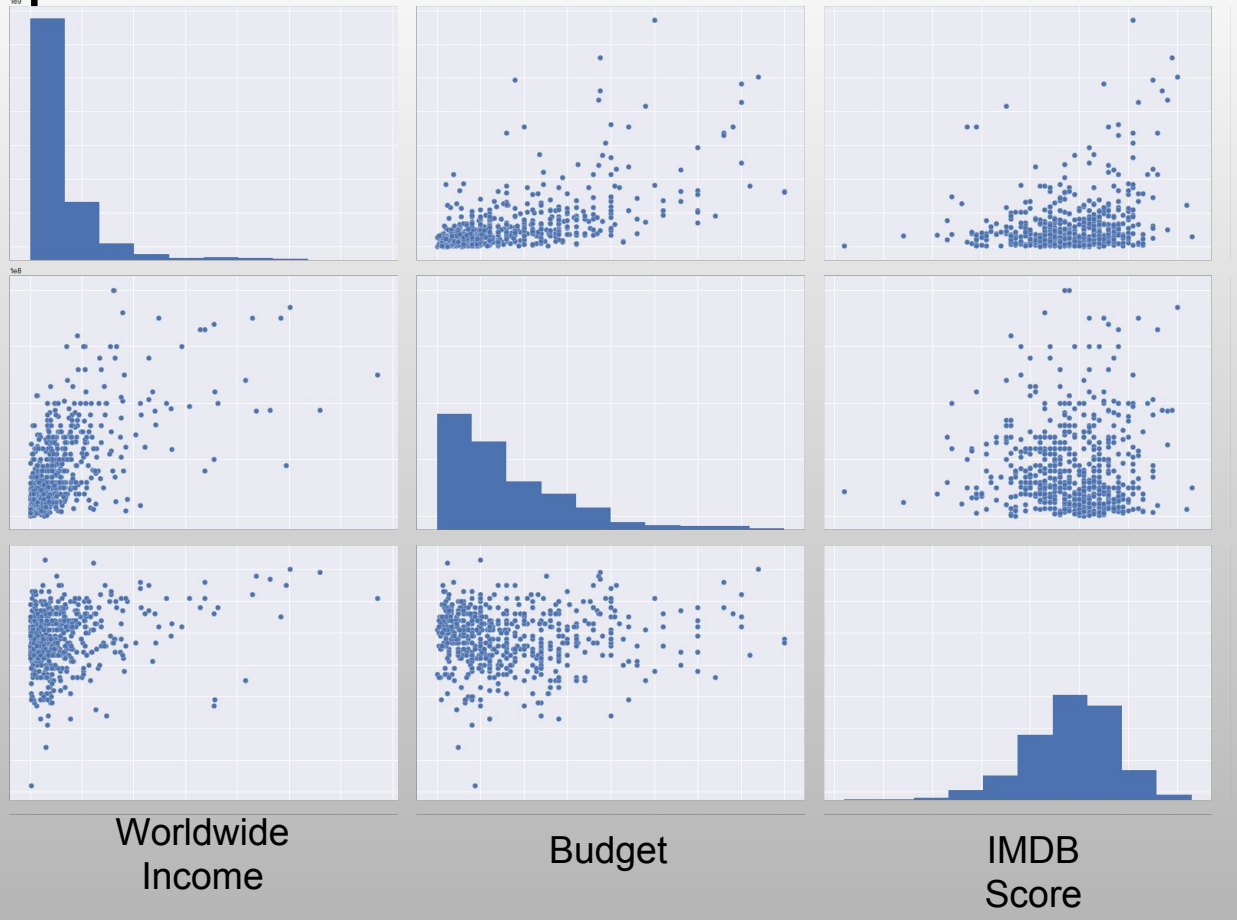
A Few Assumptions

- Star power has the same weight as quality
- Time value of money applies to both cost and revenue streams

Process

- Scrape Data from IMDB 2000 - 2016 (600+ movies)
- Compare with award winners (Wikipedia)
- Clean Data
- Explore Data
- Model Data
- Test Data
- Refine Process

EDA - Pairplot



First Fit


Worldwide Income vs Budget
1e9
1.0
0.8
0.6
0.4
0.2
Worldwide Income (billions)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e8
budget (100 millions)

Findings

Reduce ROI to 15

R^2 : 0.3516

For every $ change in budget, expect an increase of $2.5938 in worldwide revenue

Budget with Categorical Variables = 0.44872015523336373

Adding Interactions with Budget = 0.48664962632436037

Adding IMDB Rating = 0.51965927757669672

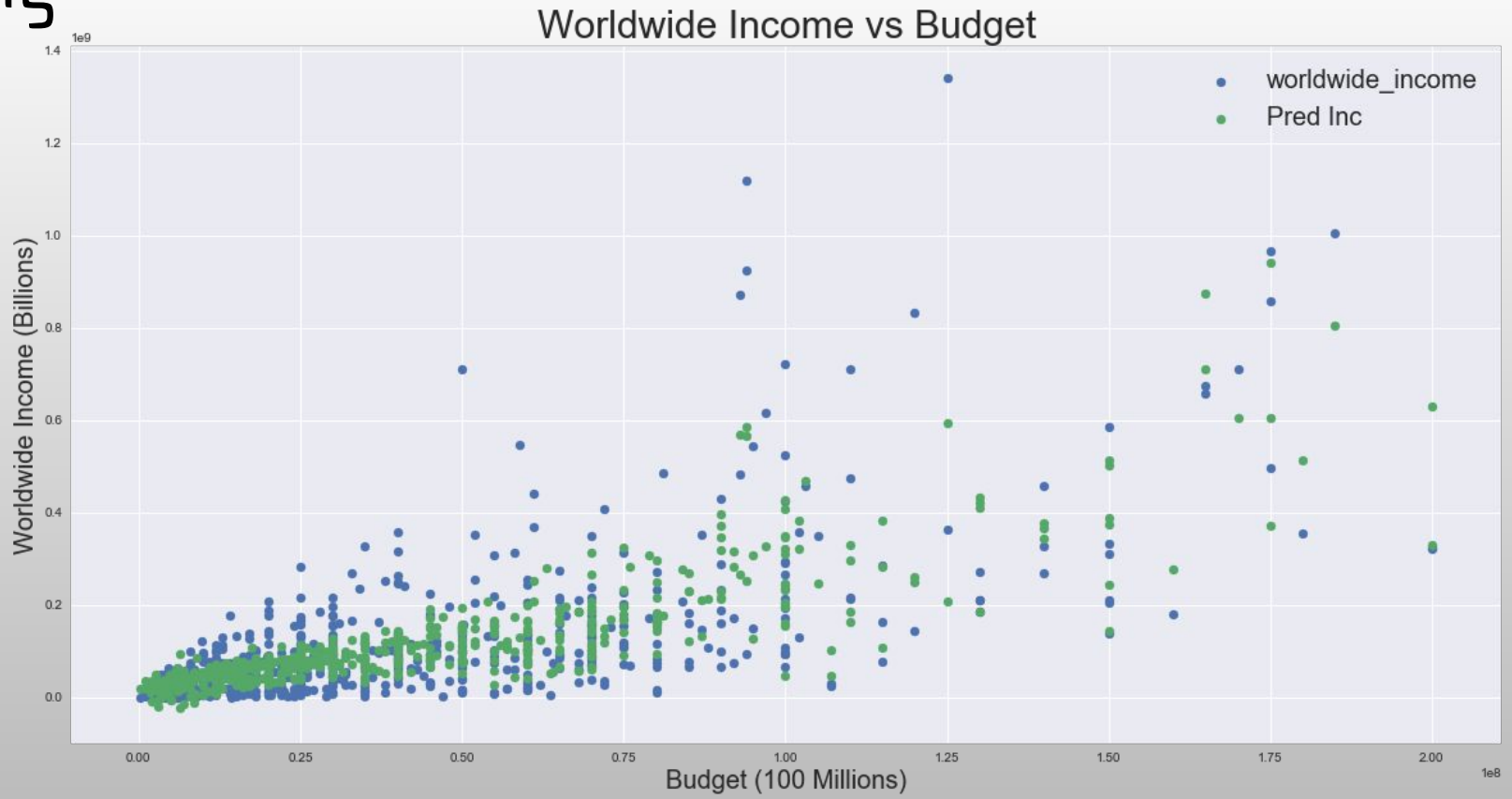
Adding Interaction with Budget = 0.57561391212225865

Adding Number of Voted Users = 0.6303657583315645

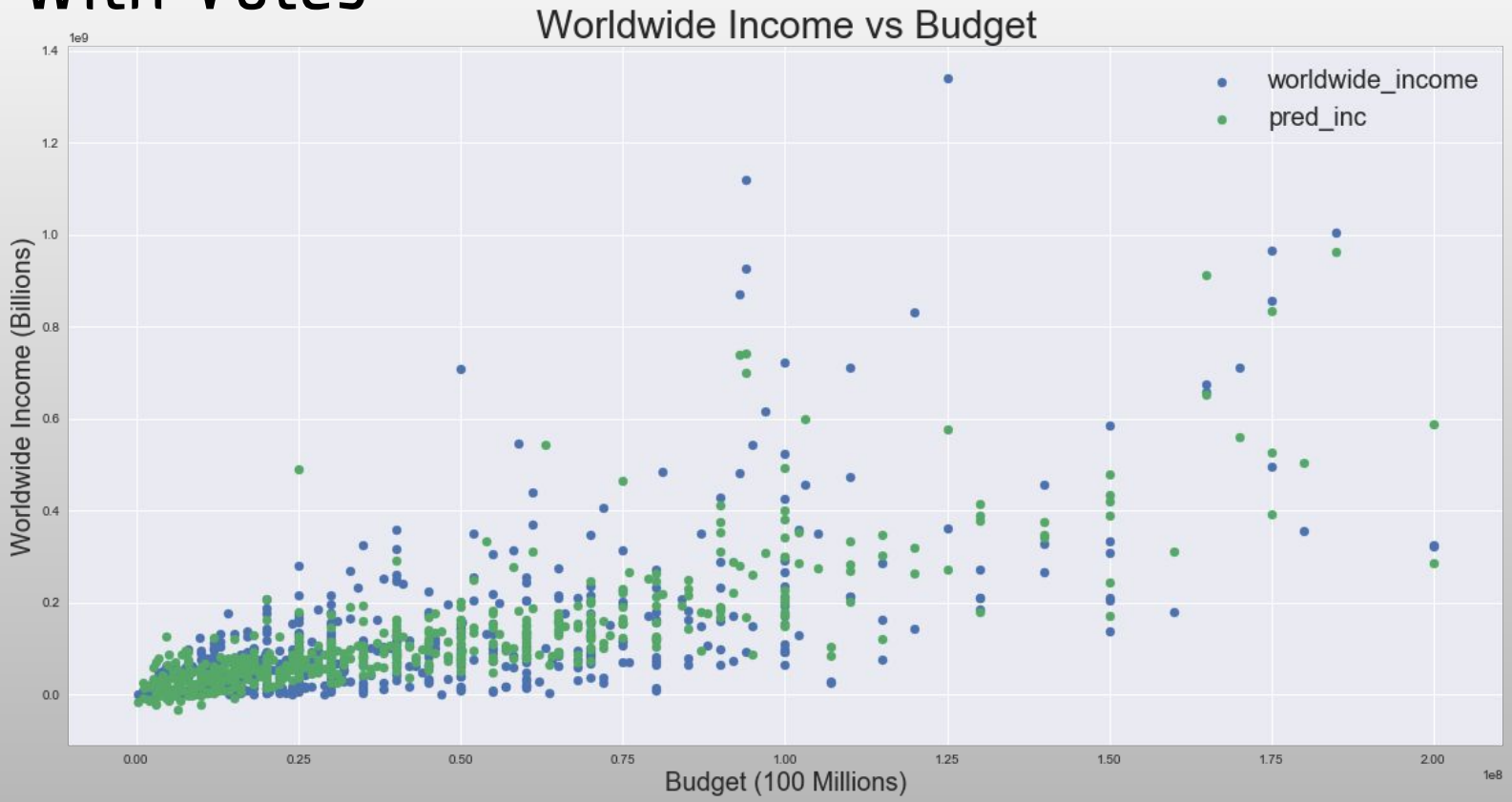
Adding Interaction with Budget = 0.64180328572789025

Runtime and other interactions failed to add to the relationship

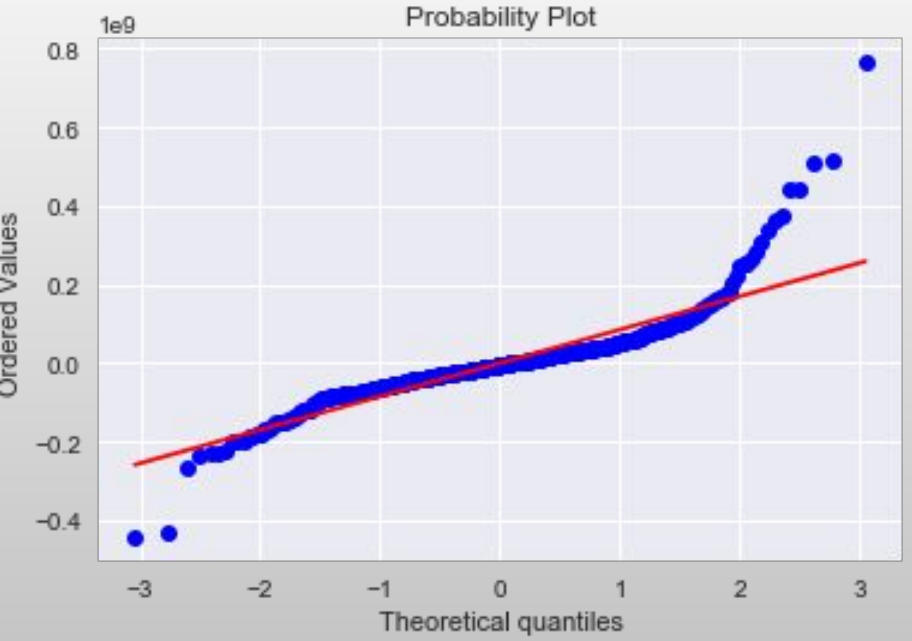
Working Model

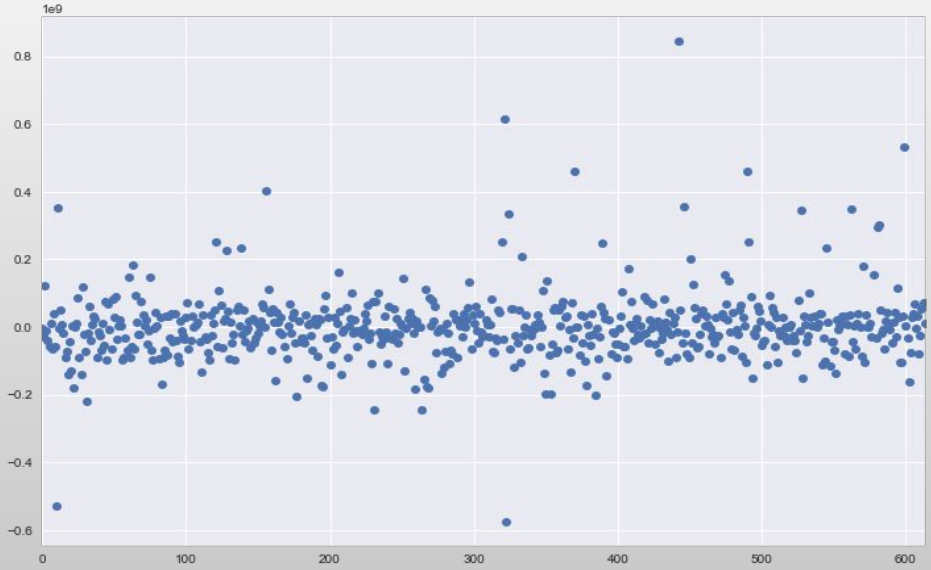


Model with Votes



Residuals


Probability Plot
1e9
Ordered Values
Theoretical quantiles
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−3
−2
−1
0
1
2
3

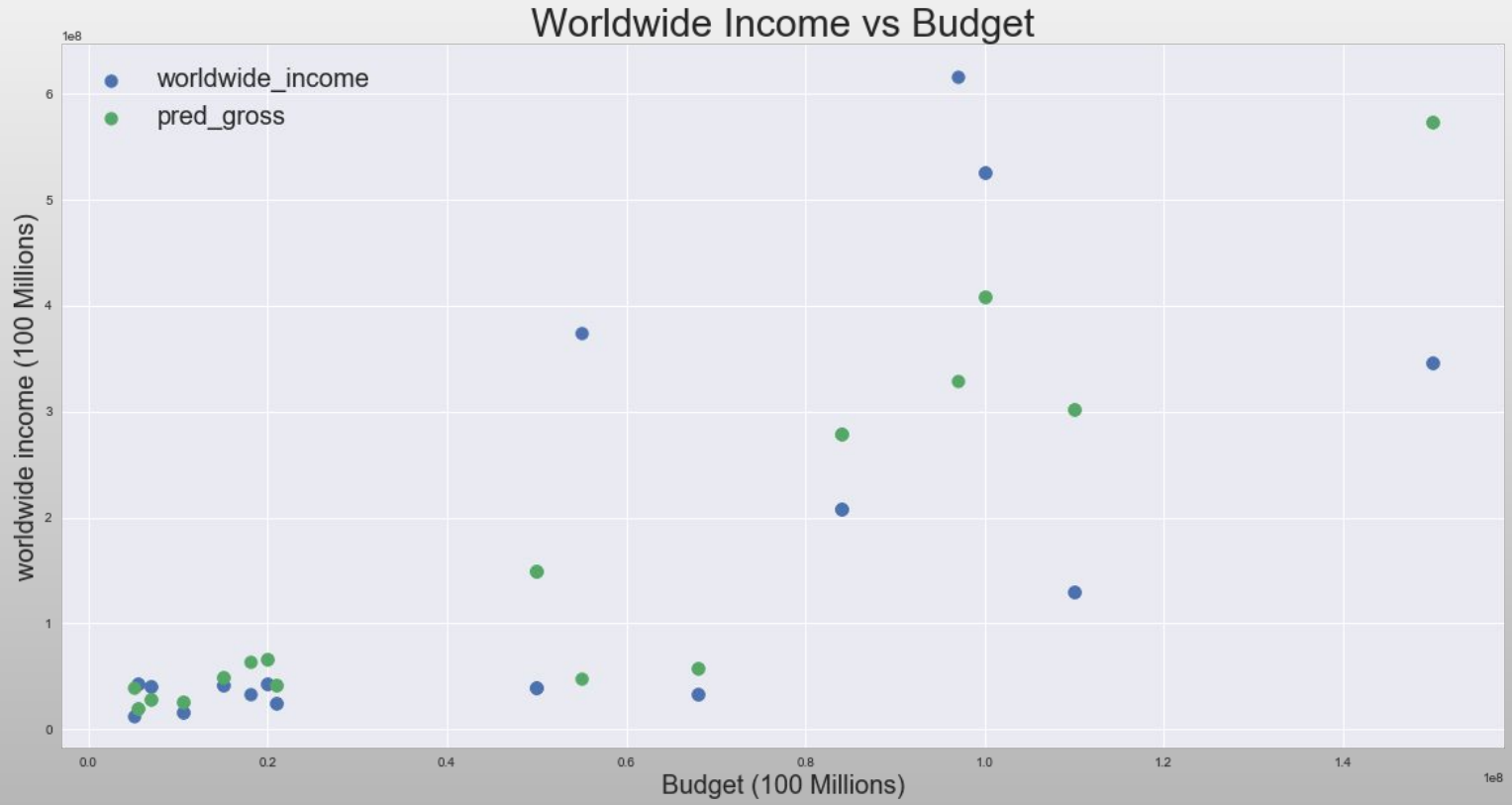

1e9
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
100
200
300
400
500
600

Cross Validation

Test Scores	Train Scores
0.5979239541538607 9	0.5560746980831594 6
0.453437645897113	0.5922876735602817 8
0.5647046456038311 6	0.5584406611201764 3
0.5124447276186848 7	0.5962581288006855 2
0.4795833332279676 5	0.590726470586850 9
0.5688823734349550 7	0.5305943652116187 3
0.4861016704476853 7	0.5925387670358555 1
0.4586023935692225	0.5988502775591171 3
0.5764571544238077 4	0.5521260562289778 1
0.6438139139237923 9	0.5223572978998725 1

Testing on Unseen Data

R^2 = 0.55



Features

budget

director_nominee

actor_nominee

actress_nominee

comedy

biography

romance

sport

thriller

crime

Adventure

budget&director_nominee

budget&actor_nominee

budget&actress_nominee

budget&comedy

budget&biography

budget&romance

budget&sport

budget&thriller

budget&crime

budget&adventure

imdb_score

budget&imdb

Feature Ranking

Budget
Budget & IMDB
Budget & Actor Nominee
Budget & Adventure
IMDB Score
Budget & Sport
Budget and Thriller
Budget & Director Nominee
Budget & Actress Nominee
Budget & Biography

Remaining...

Next Steps:

- Gather more samples!
- Split Personnel Achievements (Between and After)

Thank you