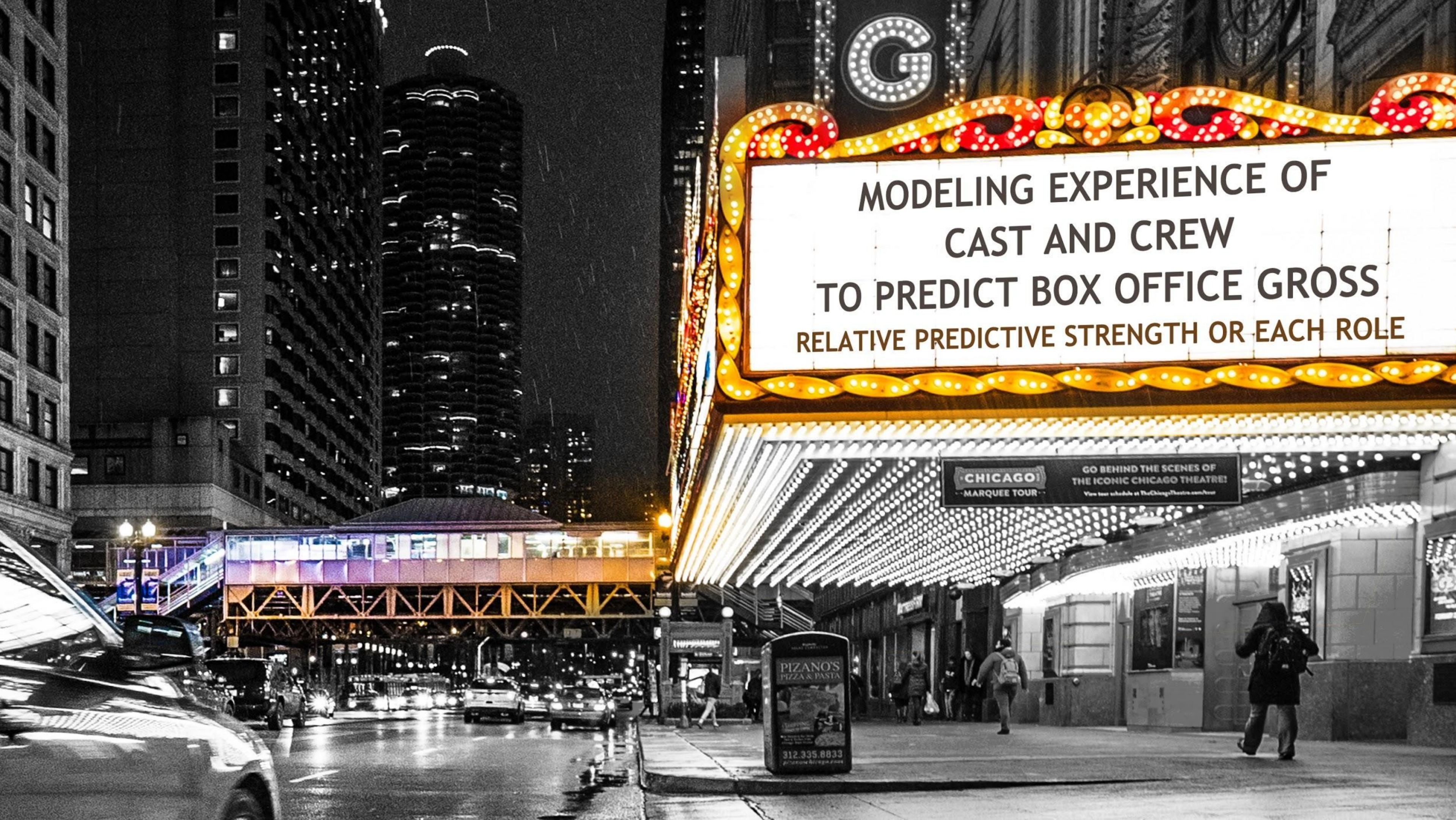


MODELING EXPERIENCE OF CAST AND CREW TO PREDICT BOX OFFICE GROSS

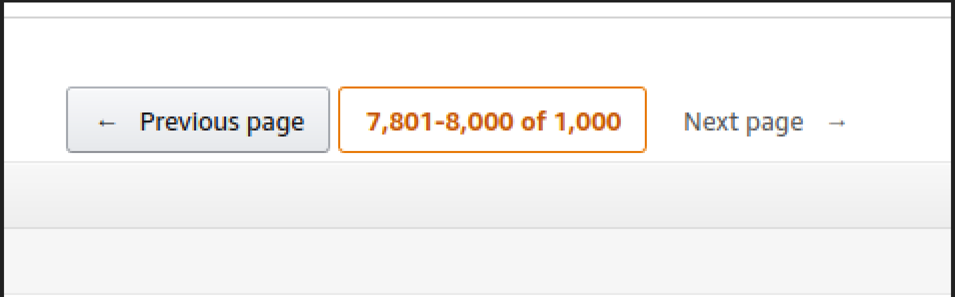
RELATIVE PREDICTIVE STRENGTH OR EACH ROLE

Webscraping

Top 10,000 Movies from Box Office Mojo

Flow:

- Collect list of movies
- Request tabs of cast and crew
- Parse information



Interesting Aspects:

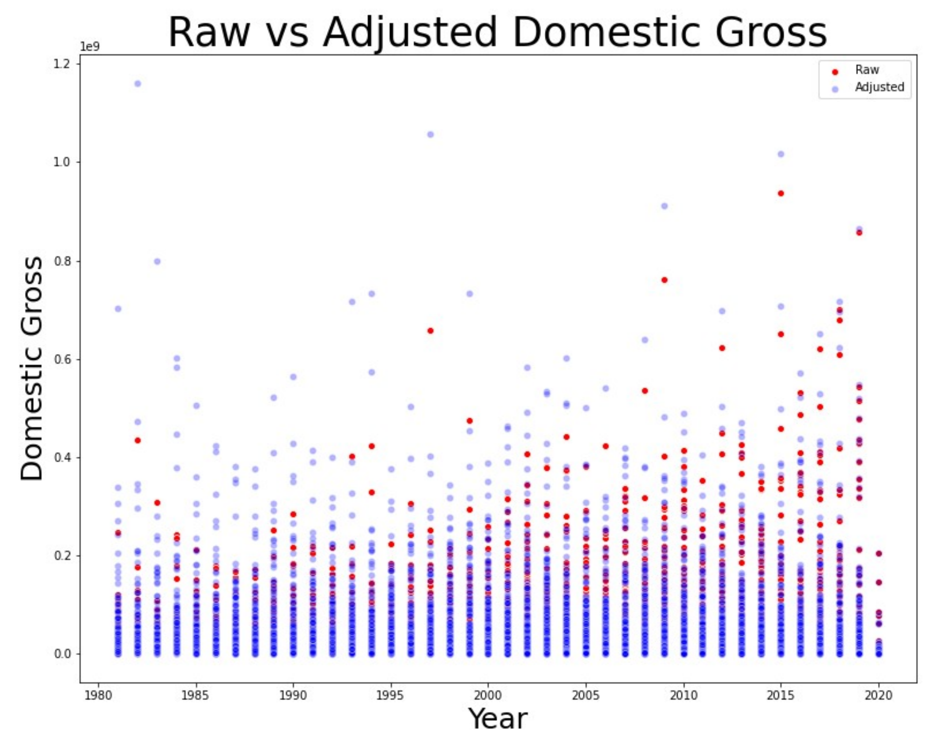
- Going out of bounds
- Inconsistent/missing information

Data Cleaning

Correcting and Dropping Inconsistencies

Issues:

- No dates
  - Drop rows
- No budgets
  - Drop budget column
- Wrong lengths
  - Check for min/hr
- Inflation adjustment
  - CPI data from BLS



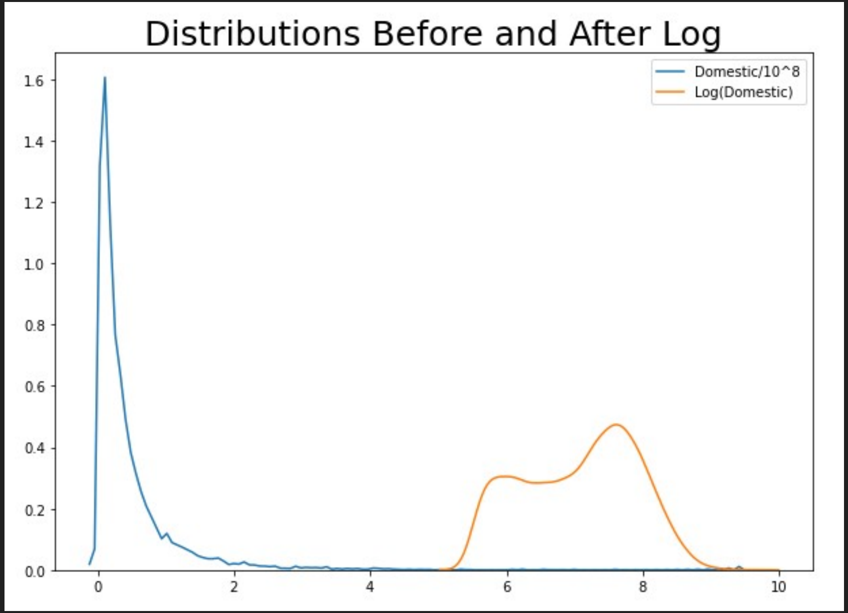
Feature Engineering

How to deal with so much categorical data?

- 1000s of actors, directors, etc
  - Sort by date
  - Find experience levels by date
  - Record maximum experience
- Multiple genres
  - Dummies
  - Form combinations
  - Disregard strange genres

What about distributions?

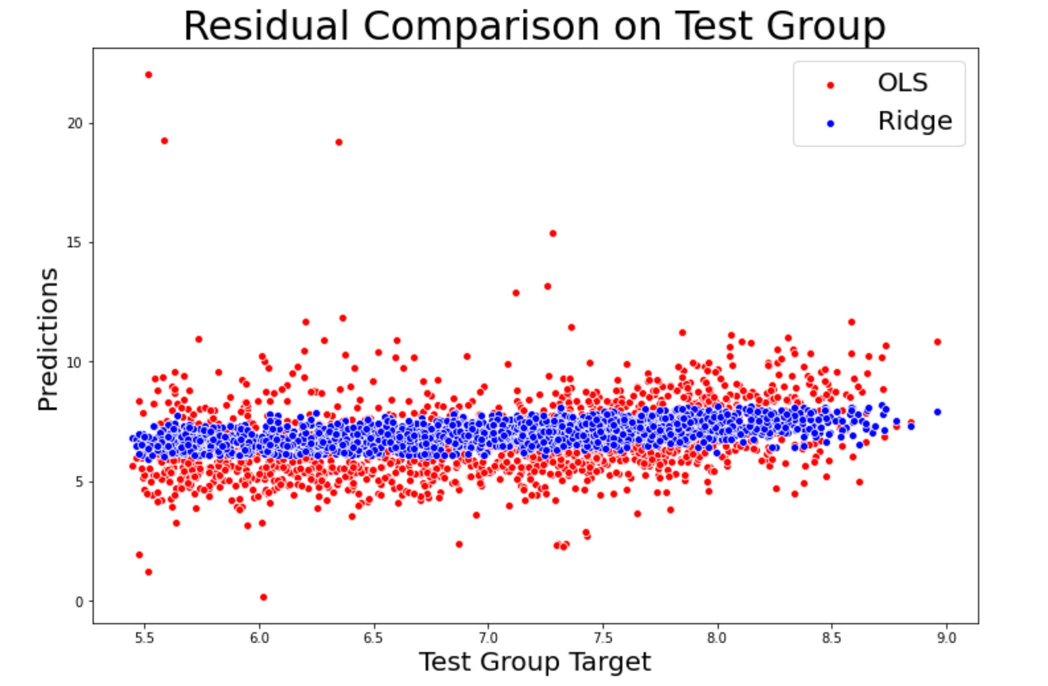
- Target (Domestic Box Office Gross)~
  - log-normal
  - Same for experience levels (not-shown)
- Year and experience
  - Unintended interaction
  - Limit to 1980



Modeling

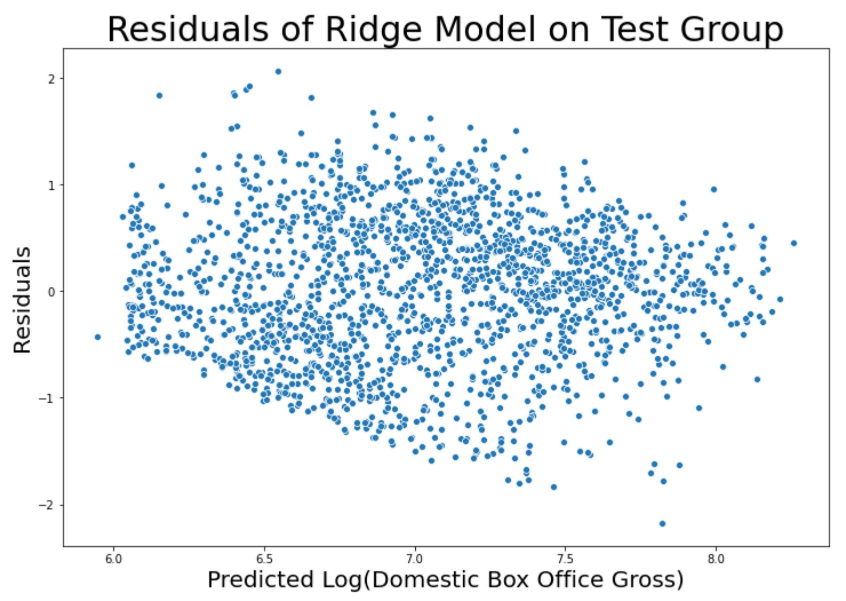
OLS Model

- Quick sense of fit
- Initial reduction of variables
- Ultimately over-fit
  - High R^2 on training
  - Low overall P
  - Higher P for features
  - Variance on Test Set



Ridge Regression Model

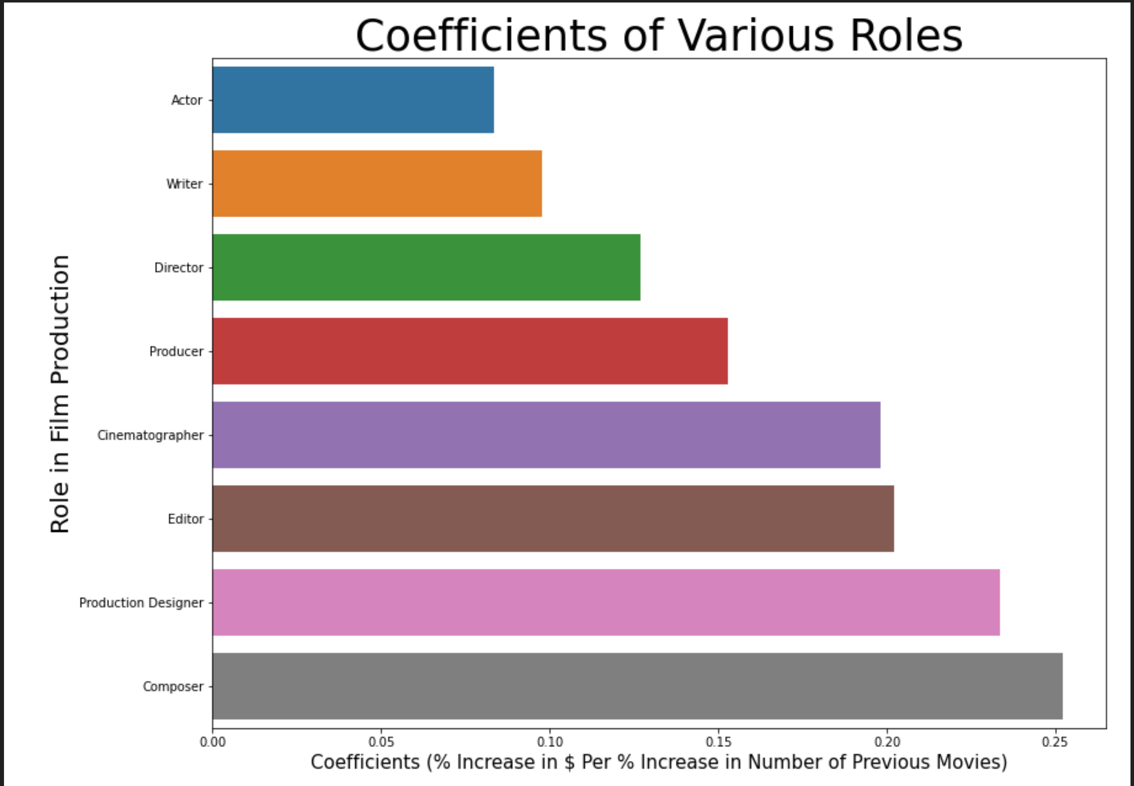
- K folds for alpha
- Reduced correlated coefficients
- Better MAE on test set
  - 0.54 vs 1.04
- Allowed to keep more variables



Results

Feature Coefficients of Interest

- Pair correlations < 0.5
- Three additional features
  - PG-13
  - Drama
- Composer experience top
- Actor experience bottom

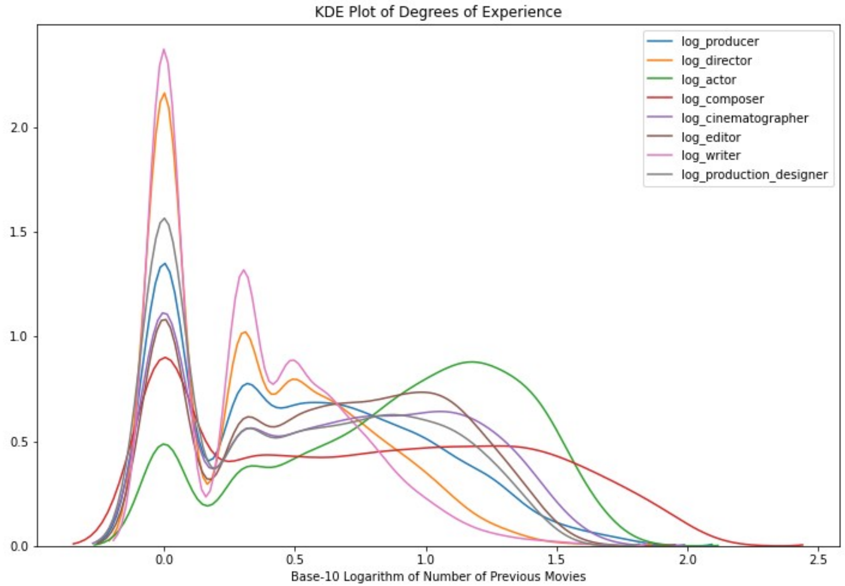


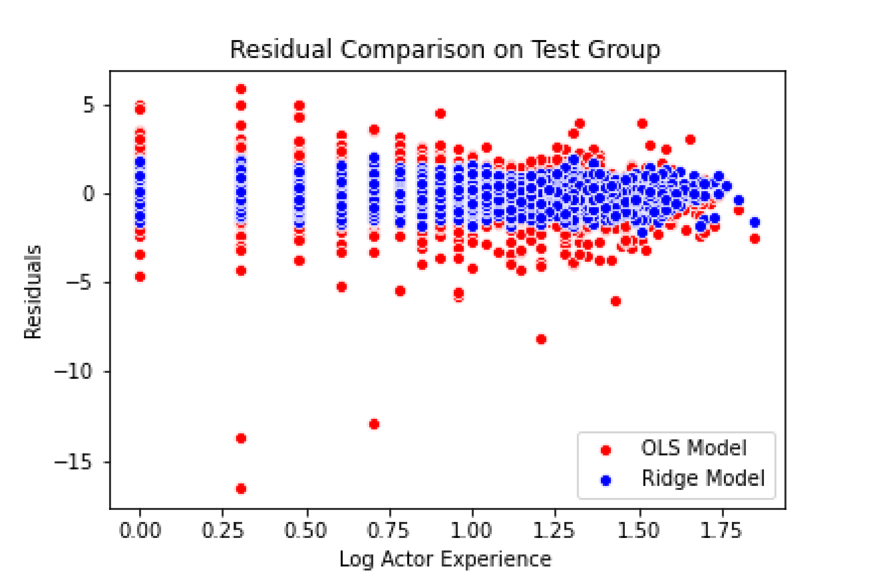
Potential Growth and Improvement

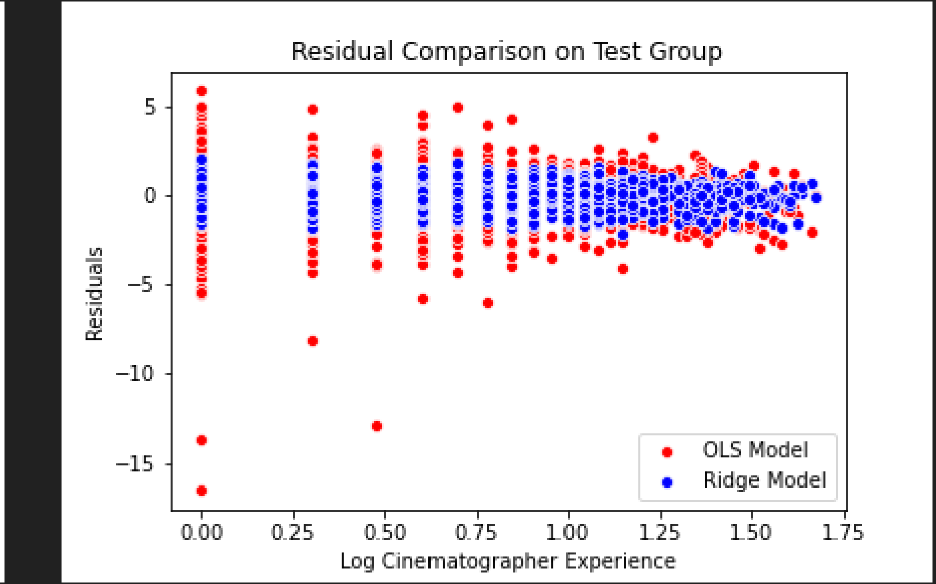
Future Work

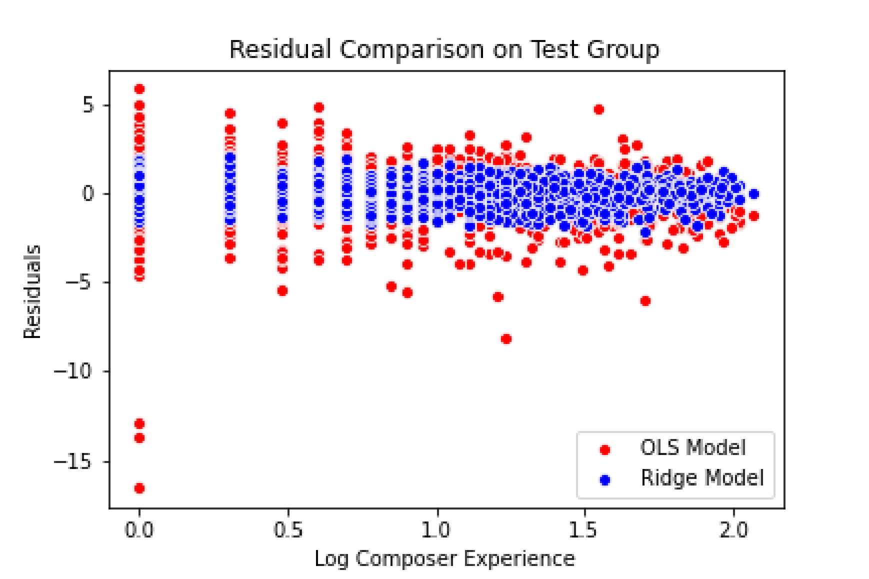
- Attempt to find strange interaction effects
  - Experience and genres
  - Combined experience
- Outside data
  - Demographic
  - Economic
- Domain research

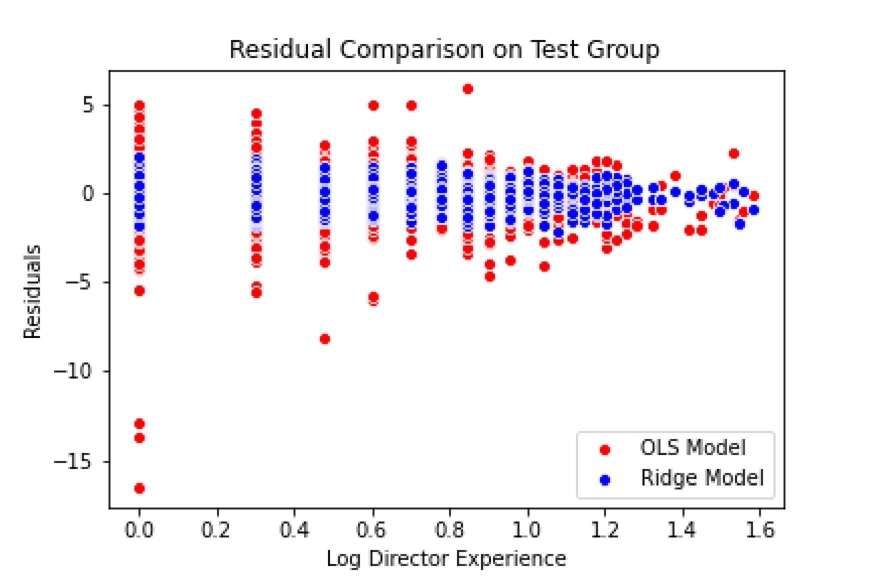
Appendix

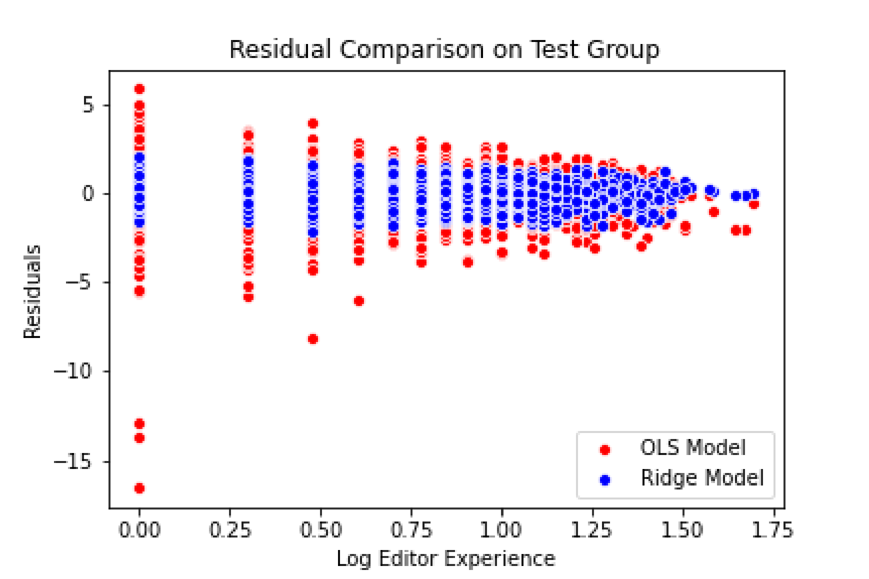

KDE Plot of Degrees of Experience
log_producer
log_director
log_actor
log_composer
log_cinematographer
log_editor
log_writer
log_production_designer
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
Base-10 Logarithm of Number of Previous Movies

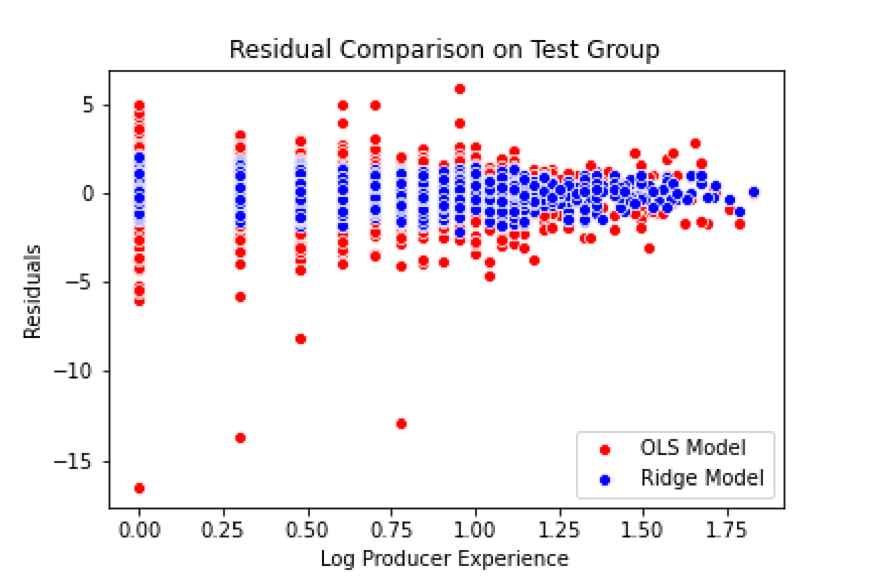

Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Log Actor Experience
OLS Model
Ridge Model

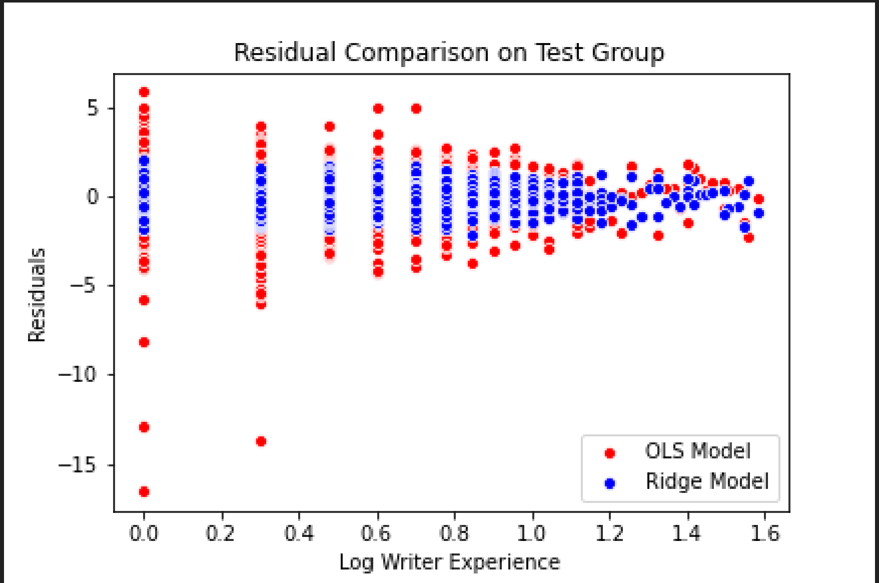

Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Log Cinematographer Experience
OLS Model
Ridge Model


Residual Comparison on Test Group
5
0
−5
−10
−15
Residuals
0.0
0.5
1.0
1.5
2.0
Log Composer Experience
● OLS Model
● Ridge Model

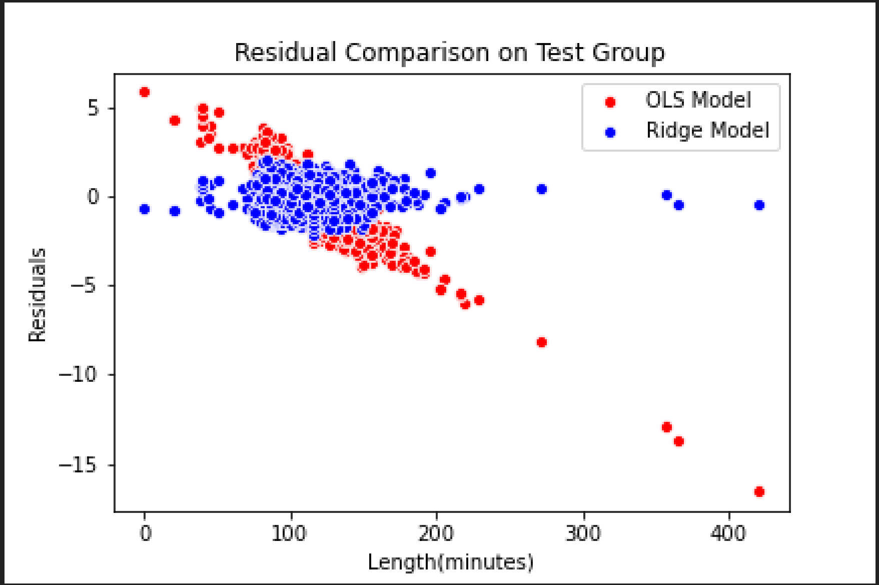

Residual Comparison on Test Group
5
0
−5
−10
−15
Residuals
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Log Director Experience
● OLS Model
● Ridge Model


Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Log Editor Experience
● OLS Model
● Ridge Model

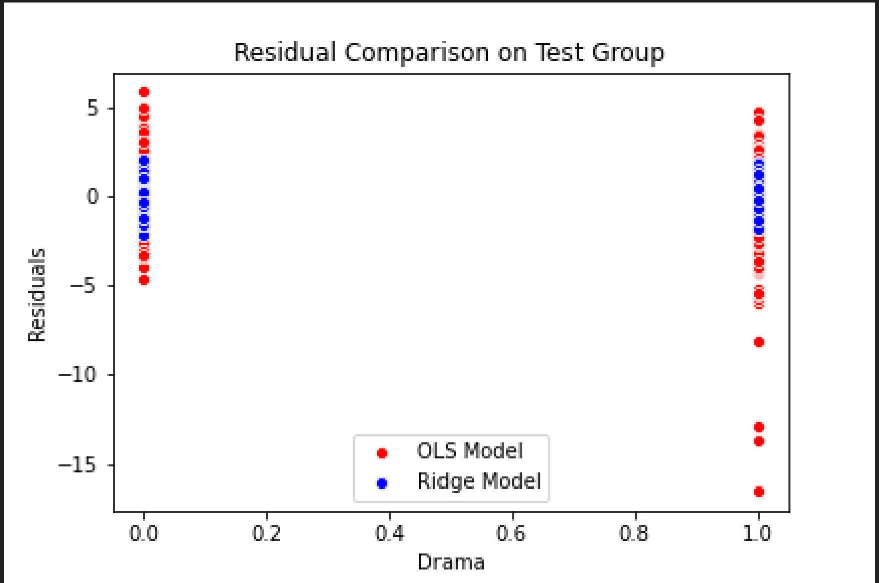

Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
Log Producer Experience
● OLS Model
● Ridge Model

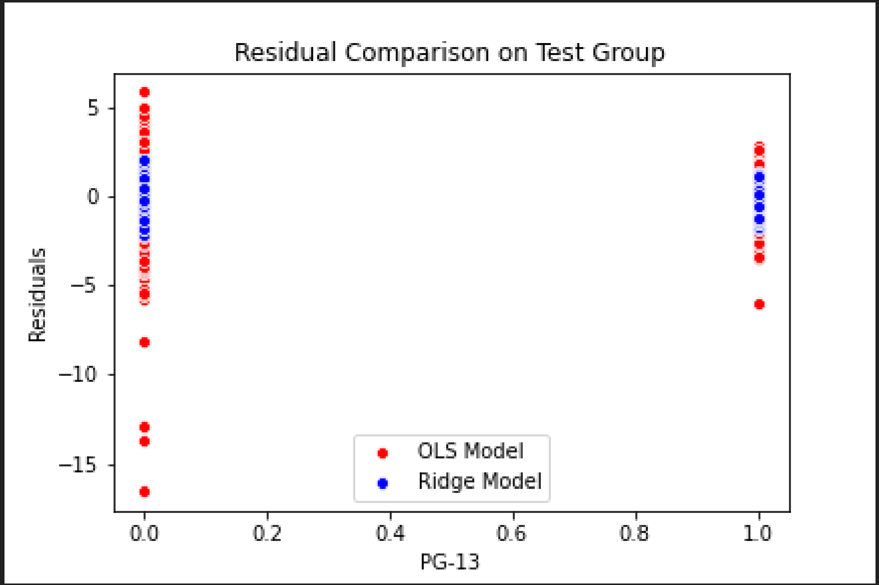


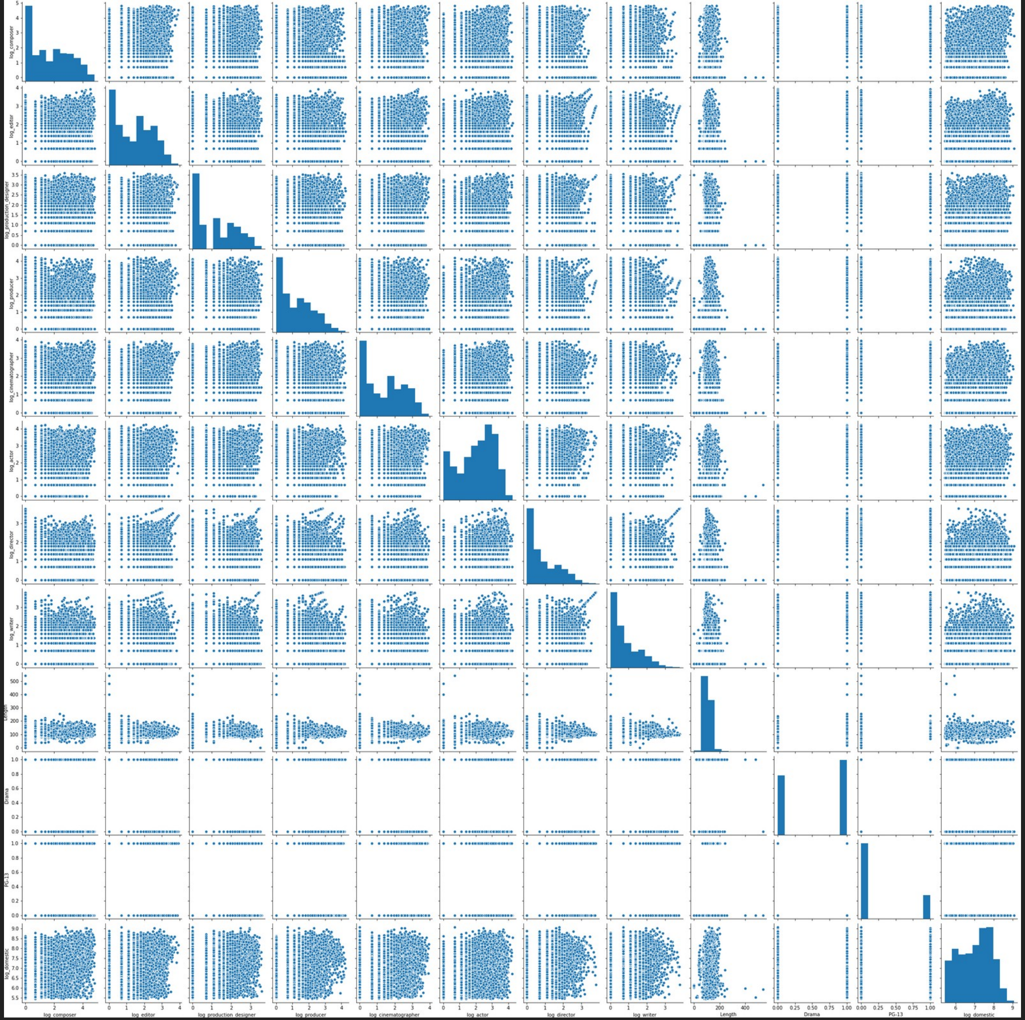
Residual Comparison on Test Group

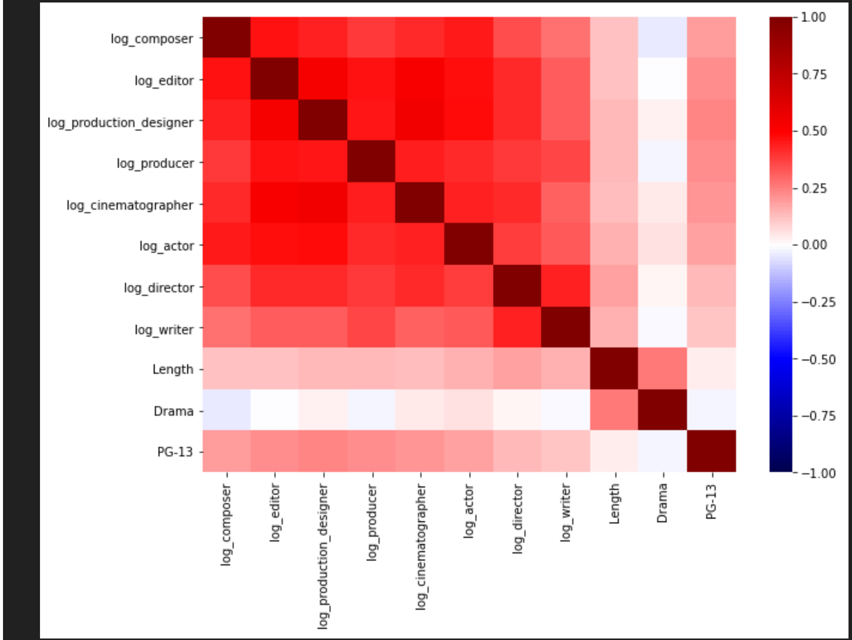


Residual Comparison on Test Group


Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.0
0.2
0.4
0.6
0.8
1.0
Drama
● OLS Model
● Ridge Model


Residual Comparison on Test Group
Residuals
5
0
−5
−10
−15
0.0
0.2
0.4
0.6
0.8
1.0
PG-13
● OLS Model
● Ridge Model


log_composer
log_editor
log_production_designer
log_producer
log_cinematographer
log_actor
log_director
log_writer
Length
Drama
PG-13
log_domestic


log_composer
log_editor
log_production_designer
log_producer
log_cinematographer
log_actor
log_director
log_writer
Length
Drama
PG-13
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Acknowledgements

Movie data: boxofficemojo.com
Inflation data: US Bureau of Labor Statistics