predicting THE OSCARS®

By Jason Chan

Brown University DSI
Dec 5, 2019
Github Repo: shorturl.at/kowBV

Introduction

- Academy Awards - world's most prestigious film ceremony
- Type of Problem: Binary Classification
  - Can be scaled to multi-class model
- Origin of Data
  - shorturl.at/juzIN

Recap

- Contains Numerical Info and 20 other Film Awards Info
  - E.g. Golden Globes, BAFTA, Directors Guild, etc.
- Before Preprocessing Shape: (1235, 69)
  - Containing films from 2000-2018
  - Total Oscar Best Picture Winners: 18
- Genre, Nomination and Won Categories
  - Data in the form of
    - “Action|Adventure|Sci-Fi”
    - “Best Song|Best Composer|Best Director|Best Picture“
  - Split data, one-hot encoded

Cross Validation

- One Hot Encoding Before CV pipeline
  Preprocessing
  - Standard Scaler
  - Multivariate Imputation

- Three models
  - Logistic Regression (`C`)
  - Random Forest (`max_depth, min_samples_split`)
  - Imputed XGBoost (`colsample_bytree, max_depth, min_child_weight`)
  - No impute XGBoost (`colsample_bytree, max_depth, min_child_weight, learning_rate, subsample`)

- Stratified 4-fold split, run 6 times for each model
- Evaluation without accuracy → precision, recall, AUPRC, F1

Results

Table 4.1.1 Logistic Regression Results, best hyperparameters: `C = 1.0`

Metric	Mean	Standard Deviation
Average Precision	0.35	0.2
Precision	0.64	0.18
Recall	0.5	0.2
F1	0.54	0.2

Table 4.1.2 Random Forest Results, best hyperparameters: `max_depth = 5.0, min_samples_split = 2.0`

Metric	Mean	Standard Deviation
Average Precision	0.1	0.12
Precision	0.33	0.47
Recall	0.08	0.12
F1	0.13	0.19

Table 4.1.3 XGBoost *with* Imputation, best hyperparameters: `colsample_bytree = 0.75, max_depth = 4, min_child_weight = 2`

Metric	Mean	Standard Deviation
Average Precision	0.3	0.25
Precision	0.67	0.37
Recall	0.33	0.24
F1	0.43	0.27

Table 4.1.4 XGBoost *without* Imputation, best hyperparameters: `colsample_bytree = 1.0, max_depth = 2, min_child_weight = 2, learning_rate = 0.01, subsample = 1.0`

Metric	Mean	Standard Deviation
Average Precision	0.39	0.11
Precision	0.69	0.16
Recall	0.58	0.19
F1	0.59	0.12

⇐ BEST MODEL

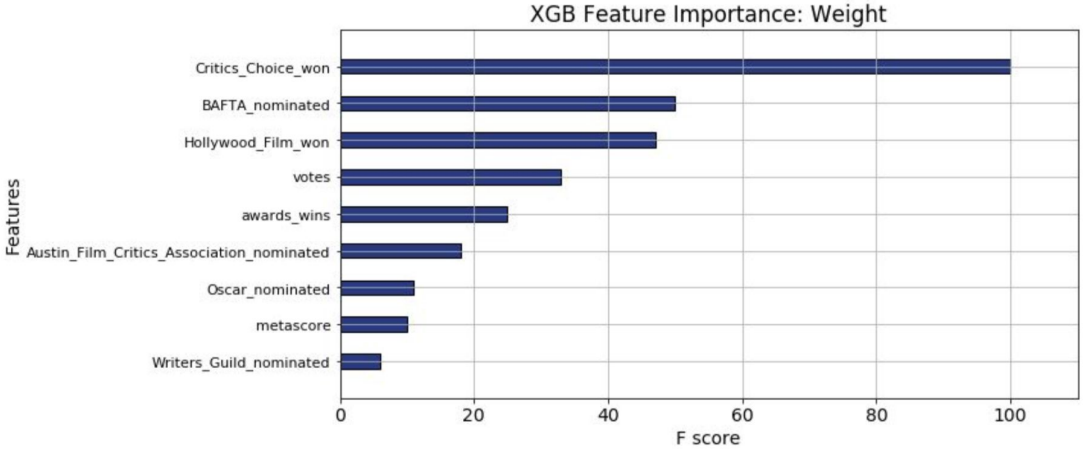
Feature Importances

Weight

the percentage representing the relative number of times a particular feature occurs in the trees of the model (# of times a feature occurs in a split).

1. Critics_Choice_won
2. BAFTA_nominated
3. Hollywood_Film_won

Figure 4.2.1 XGBoost Feature Importance Plot: Weight Importance



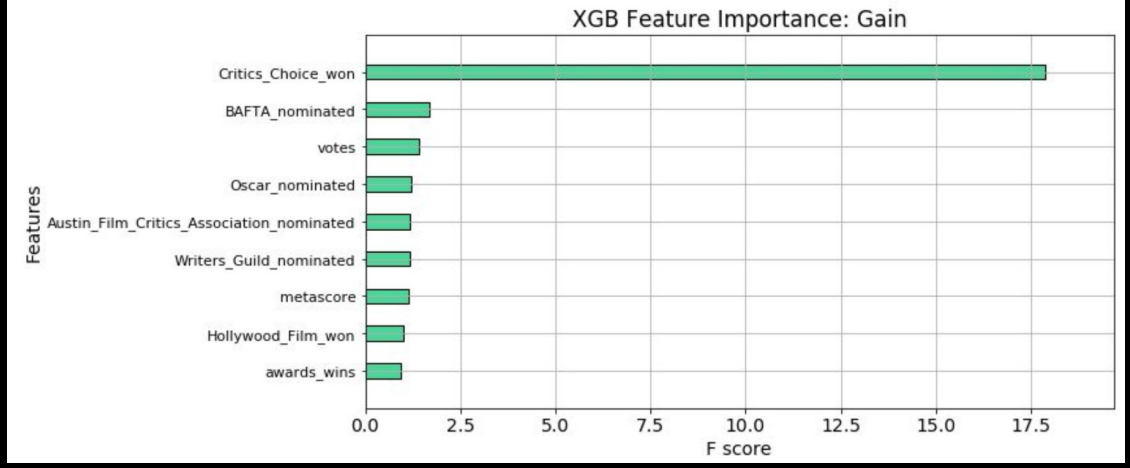
Coverage

the relative number of observations split/related to this feature.

1. Awards_won
2. Critics_Choice_won
3. Hollywood_Film_won

Figure 4.2.2 XGBoost Feature Importance Plot: Coverage Importance

Figure 4.2.3 XGBoost Feature Importance Plot: Gain Importance

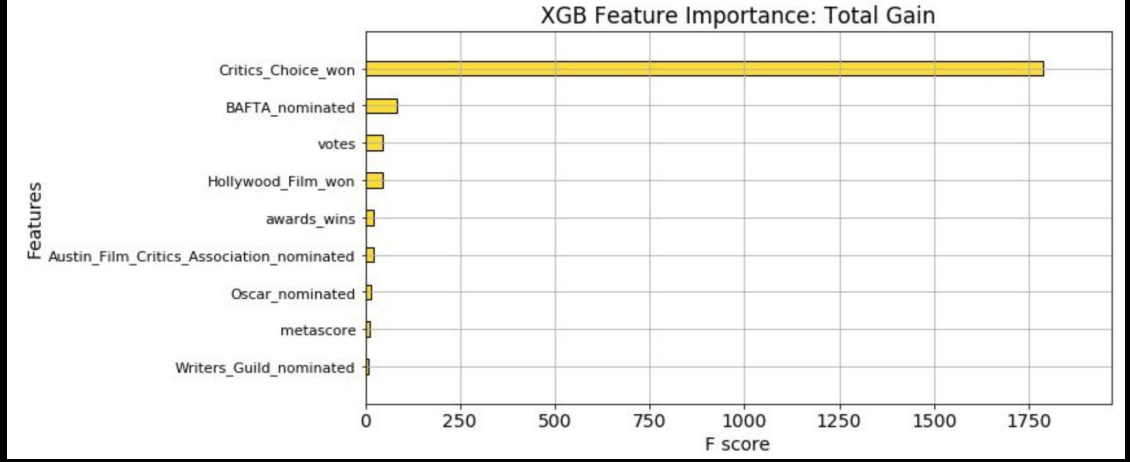


Gain

implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. The average gain across all splits the feature is used in

1. Critics_Choice_won
2. BAFTA_nominated
3. votes

Figure 4.2.4 XGBoost Feature Importance Plot: Total Gain Importance



Total Gain

the total gain across all splits the feature is used in.

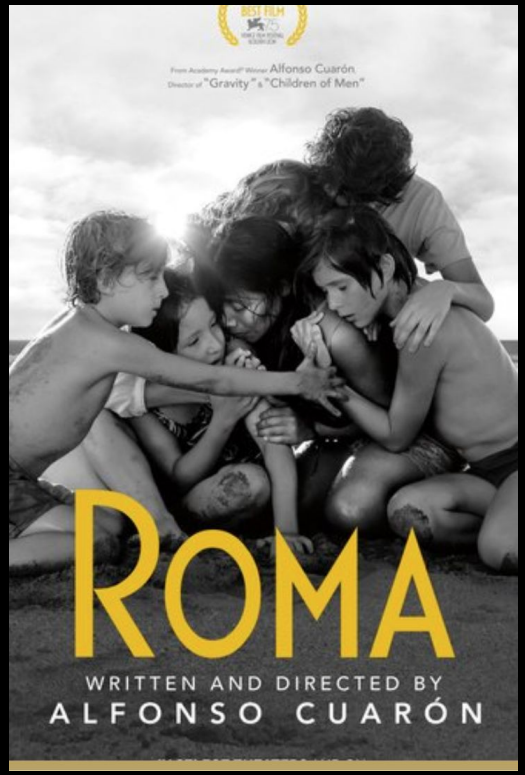
1. Critics_Choice_won
2. BAFTA_nominated
3. votes

Testing my model....

Testing on 2019 Oscars Best Picture....

Movie	Win Score
Roma	45.9
The Favourite	25.1
Black Panther	21.1
A Star Is Born	18.9
Green Book	18.5
Vice	18.5
Bohemian Rhapsody	18.5
BlacKkKlansman	18.5

My prediction



WINNER

Who was right? Who was wrong?
We may never know...

Los Angeles Times

Must Reads: Oscars 2019: 'Green Book' is the worst best picture winner since 'Crash'

VARIETY ≡ FILM TV MUSIC TECH THEATER REAL EST

HOME > FILM > NEWS

FEBRUARY 25, 2019 2:

'Green Book': Twitter Reacts With Shock, Disdain After Oscars Best Picture Win

ACADEMY AWARDS

'Green Book' should never have won the best picture Oscar. Here's why

Brian Truitt USA TODAY

Outlook

- Dimensionality Reduction (>1000 features)
  - PCA, TSNE, SelectKBest (F-Test, Mutual Information
- Resampling Techniques
  - Tried stratified
  - Oversampling/Undersampling? Others? Combination?
- Limited Resource Constraints
  - Time/Computational Power
  - Upgrade → Better hyperparameter tuning
- MCAR test
  - Better method of dealing with missing data? Reduced features model?

Goals

1. Extend to other Oscar awards (Best Director, Actress, Actor...)
2. PREDICT THE 2020 OSCARS WINNERS!

Anyone up for a friendly wager?

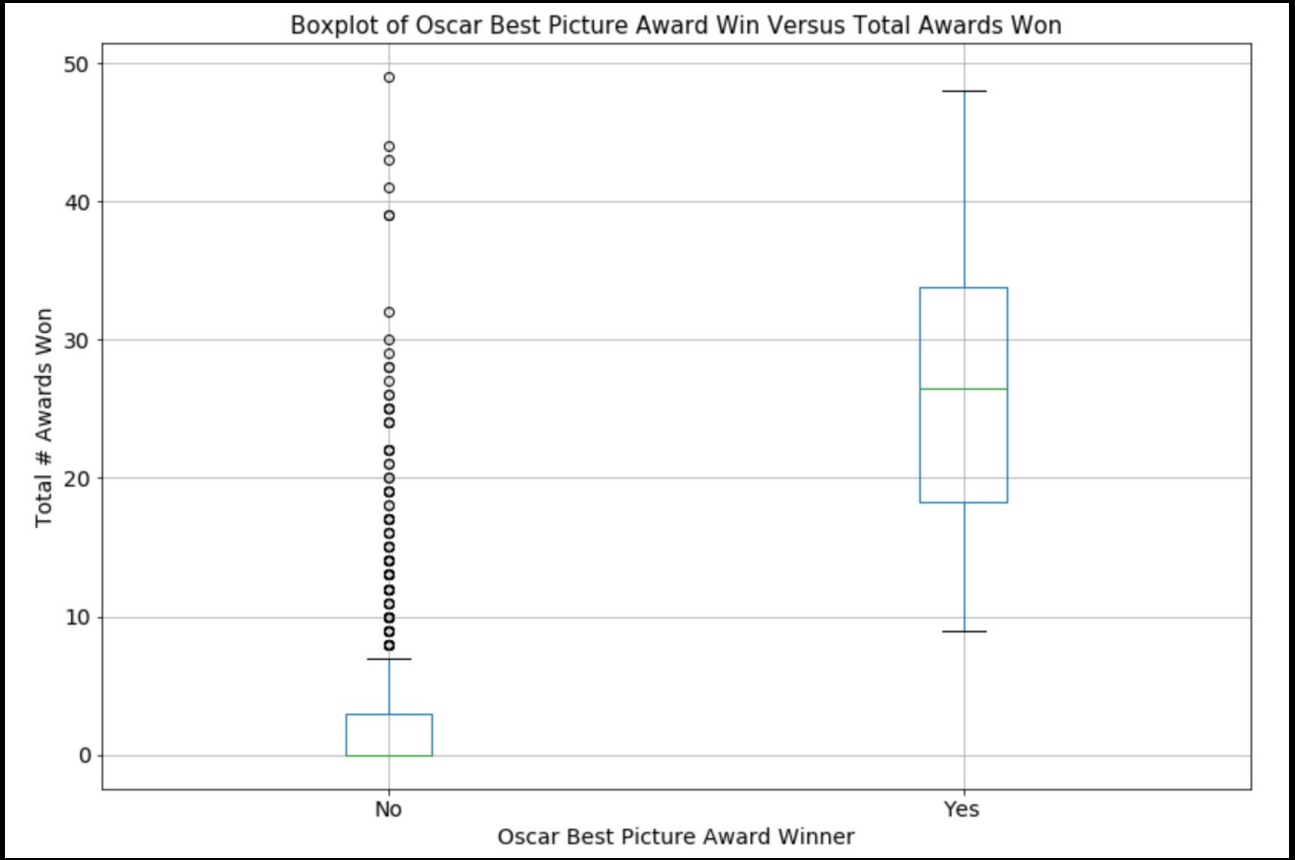
predicting

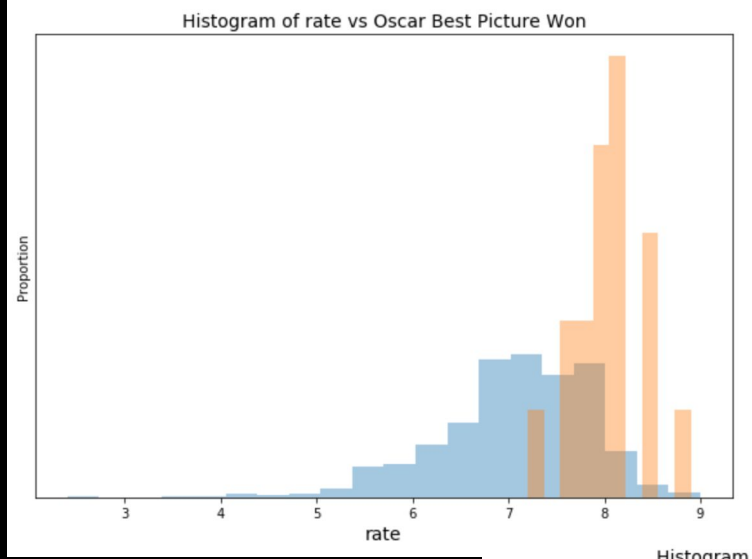
THE

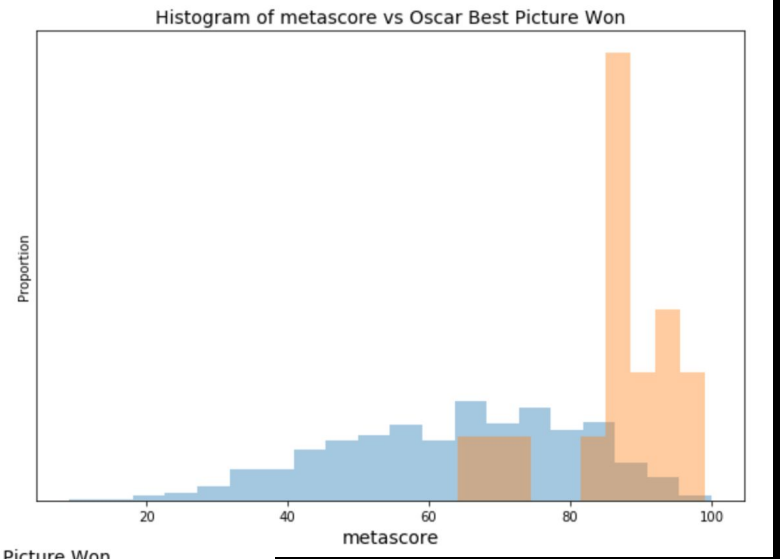
OSCARS®

THANK YOU!

Exploratory Data Analysis

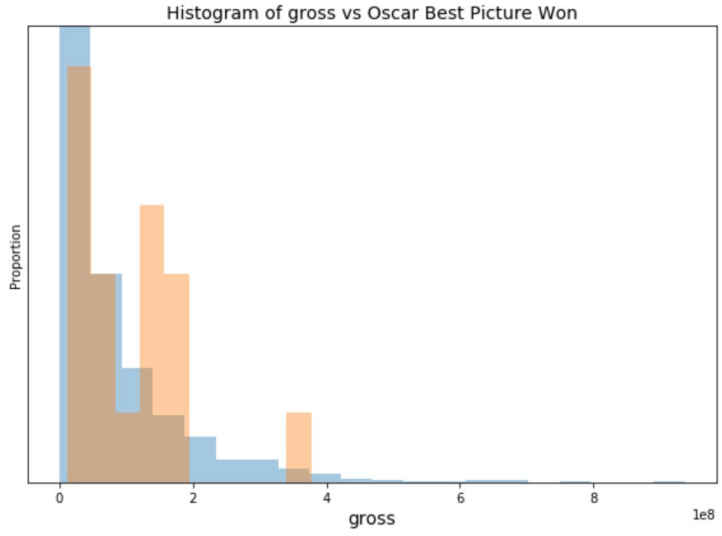






IMDB Rating

Metacritic Rating



Gross ($M)

LEGEND

Oscar Best Picture...
Orange: *WON*
Blue: *DID NOT WIN*