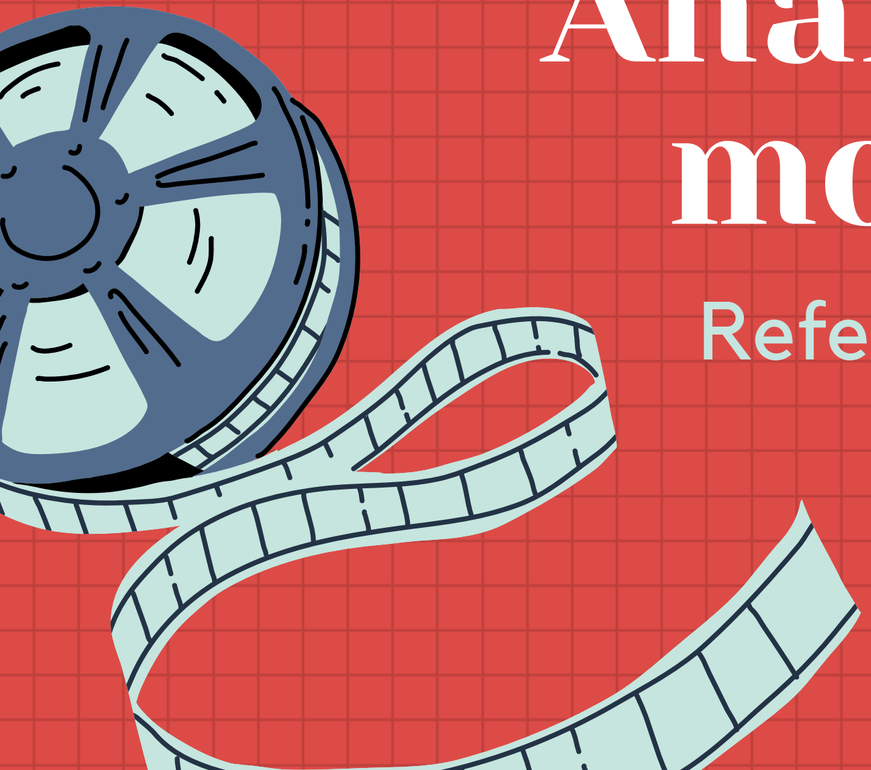
Statistical Methods for High Dimensional Data

Analysis of Italian movies success

Reference years: 2022 and 2023



Features

Number of Movies: 382

1. MovieName
2. Director
3. Cast
4. Duration
5. ReleaseDate
6. Genre
7. Production
8. Rating_MyMovies
9. Rating_Critic
10. Rating_Public
11. Rating_Average
12. Revenue
13. Nominations
14. Awards
15. Wikipedia_trends
16. Max_instafollowers
16. Time_Setting
17. Place_Setting
18. YTtrailer_views
19. Suggested_audience
20. Themes

Features

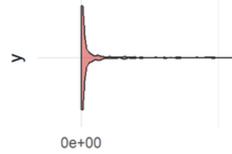
One hot encoding

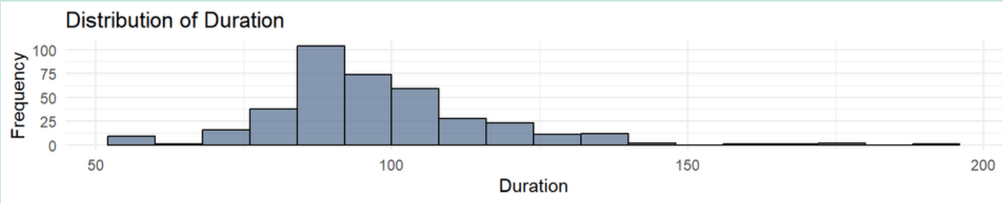
Number of Movies: 382

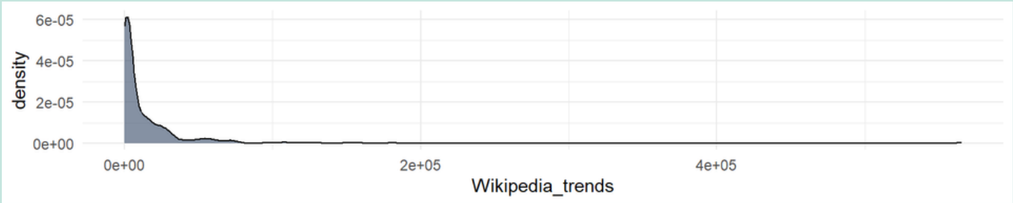
1. MovieName
2. Director
3. Cast
4. Duration
5. ReleaseDate
6. Genre
7. Production
8. Rating_MyMovies
9. Rating_Critic
10. Rating_Public
11. Rating_Average
12. Revenue
13. Nominations
14. Awards
15. Wikipedia_trends
16. Max_instafollowers
16. Time_Setting
17. Place_Setting
18. YTtrailer_views
19. Suggested_audience
~~20~~ Themes

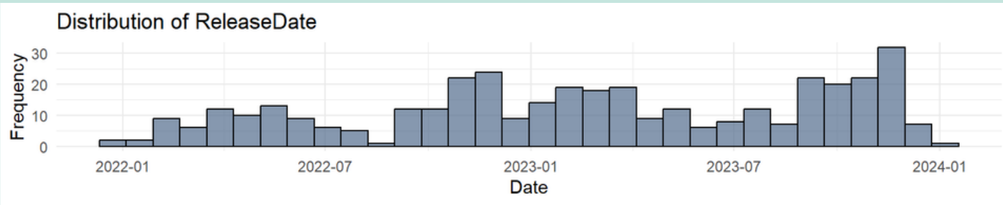
2397

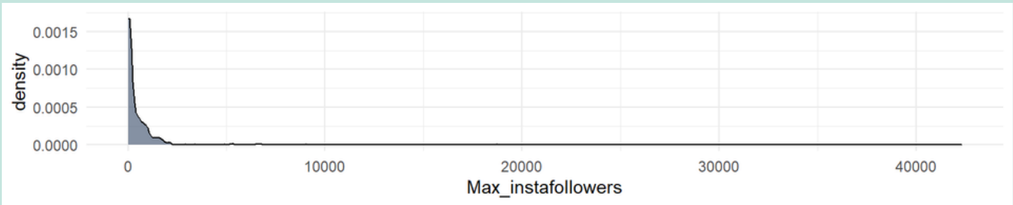
EDA

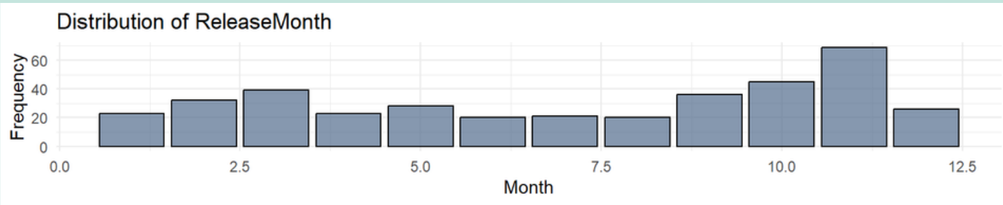


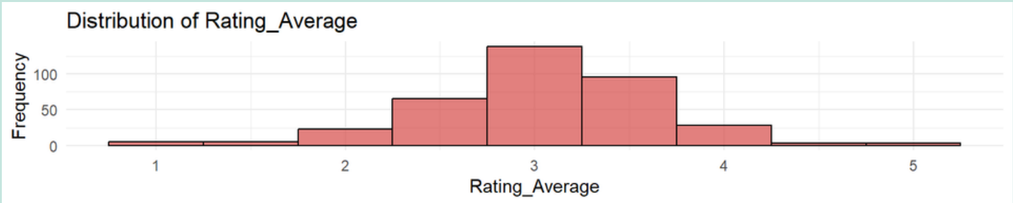


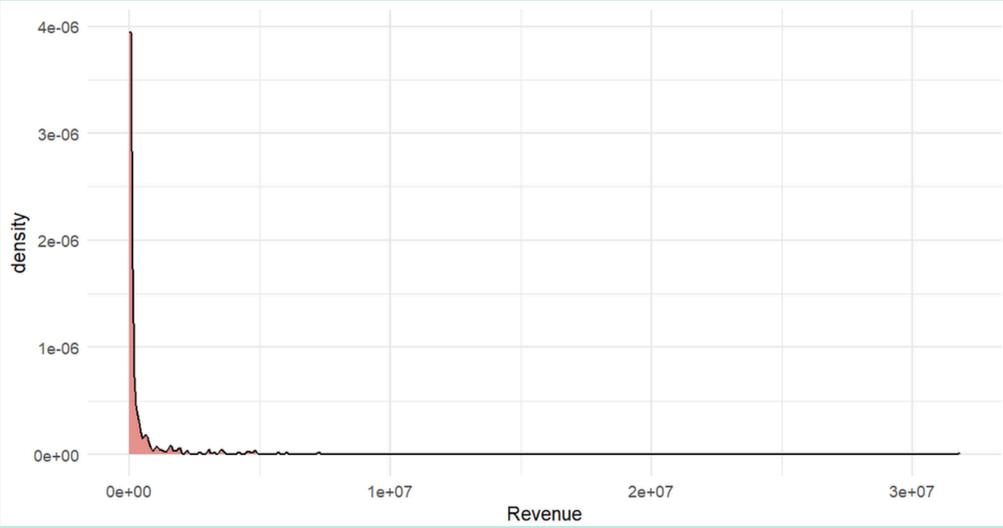






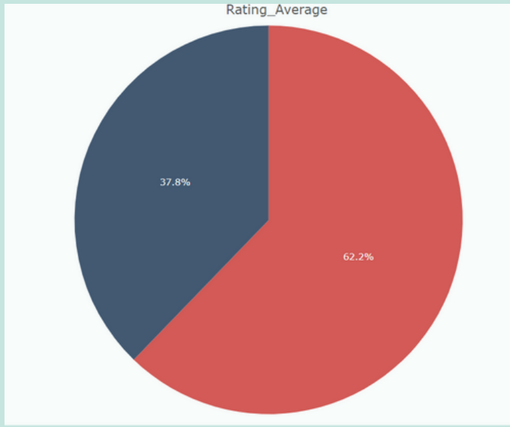




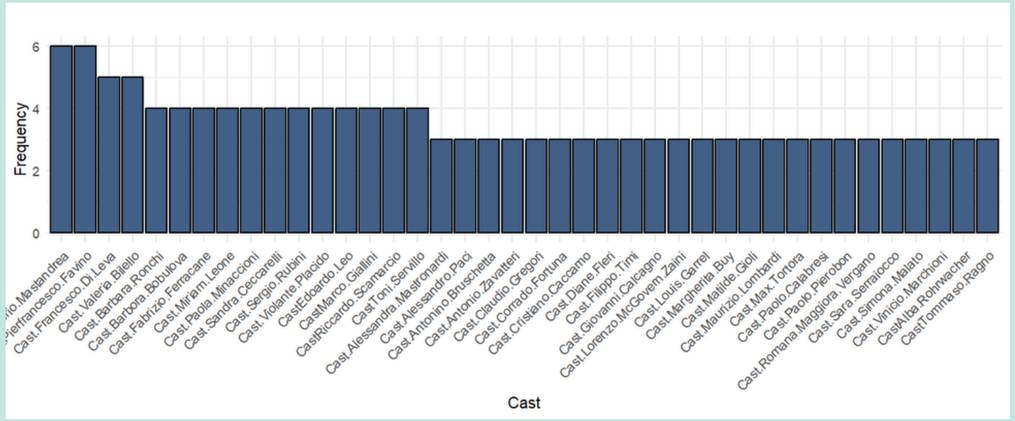


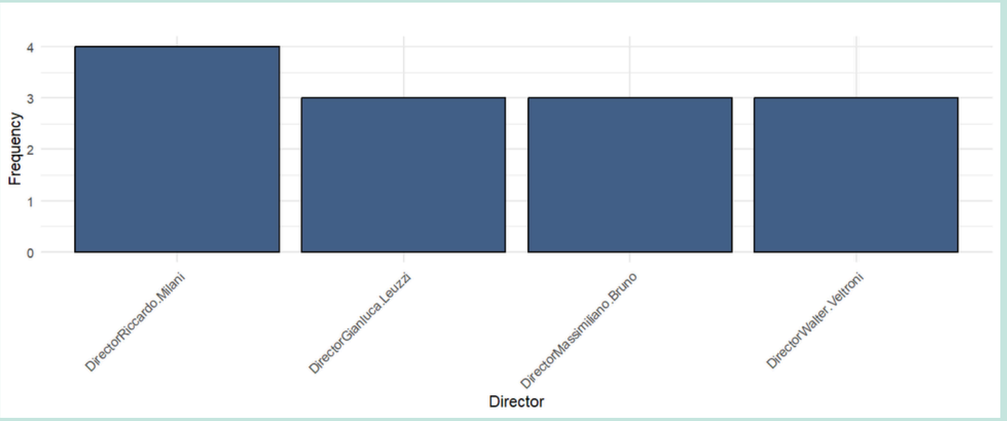
Success:

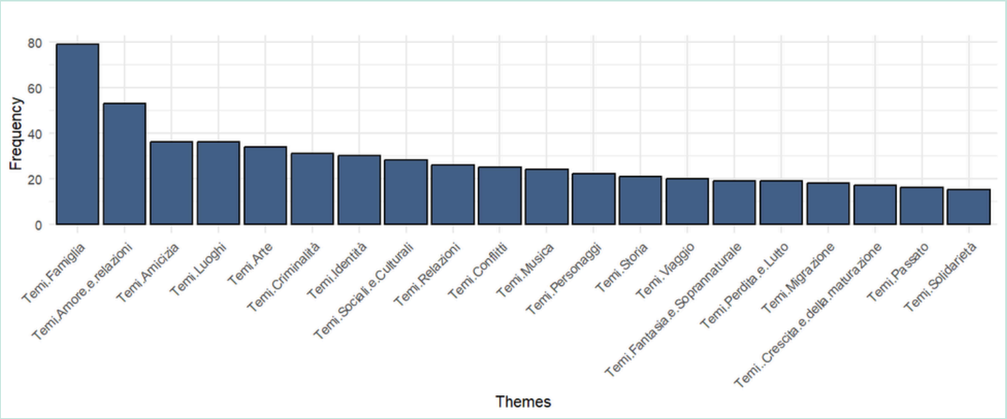
1 if Rating_Average ≥ 3
0 otherwise

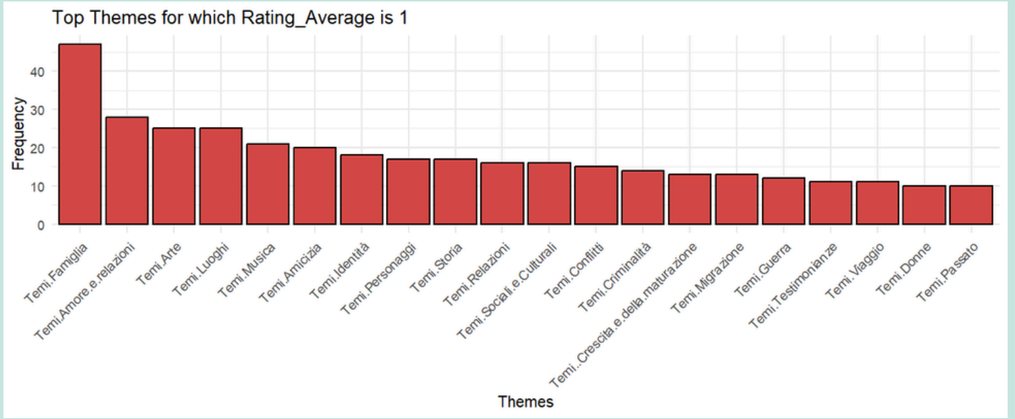


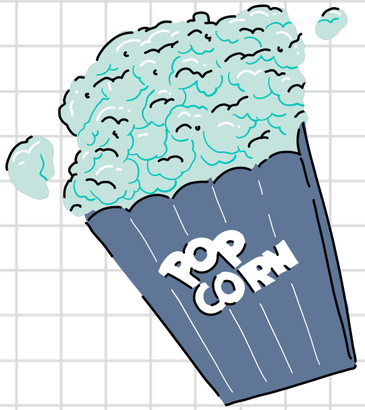
EDA


Frequency
0
2
4
6
Cast
Cast.Fabrizio.Ferracane
Cast.Paola.Minaccioni
Cast.Sandra.Ceccarelli
Cast.Sergio.Rubini
Cast.Violante.Placido
CastEdoardo.Leo
CastMarco.Giallini
CastRiccardo.Scamarcio
CastToni.Servillo
Cast.Alessandra.Mastronardi
Cast.Alessandro.Haber
Cast.Alessandro.Preziosi
Cast.Antonio.Albanese
Cast.Antonio.Bruschetta
Cast.Antonio.Zavattieri
Cast.Claudio.Gregori
Cast.Corrado.Fortuna
Cast.Cristiano.Caccamo
Cast.Diane.Fleri
Cast.Filippo.Timi
Cast.Giovanni.Calcagno
Cast.Lorenzo.McGovern.Zaini
Cast.Louis.Garrel
Cast.Margherita.Buy
Cast.Matilde.Gioli
Cast.Maurizio.Lombardi
Cast.Max.Tortora
Cast.Paolo.Calabresi
Cast.Paolo.Pierobon
Cast.Romana.Maggiora..Vergano
Cast.Sara.Serraiocco
Cast.Simona.Malato
Cast.Vinicio.Marchioni
CastAlba.Rohrwacher
CastTommaso.Ragno


Frequency
0
1
2
3
4
Director
DirectorRiccardo.Milani
DirectorGianluca.Leuzzi
DirectorMassimiliano.Bruno
DirectorWalter.Veltroni


Frequency
0
20
40
60
80
Themes
Temi.Famiglia
Temi.Amore.e.relazioni
Temi.Amicizia
Temi.Luoghi
Temi.Arte
Temi.Criminalità
Temi.Identità
Temi.Sociali.e.Culturali
Temi.Relazioni
Temi.Conflitti
Temi.Musica
Temi.Personaggi
Temi.Storia
Temi.Viaggio
Temi.Fantasia.e.Soprannaturale
Temi.Perdita.e.Lutto
Temi.Migrazione
Temi..Crescita.e.della.maturazione
Temi.Passato
Temi.Solidarietà


Top Themes for which Rating_Average is 1
Frequency
0
10
20
30
40
Themes
Temi.Famiglia
Temi.Amore.e.relazioni
Temi.Arte
Temi.Luoghi
Temi.Musica
Temi.Amicizia
Temi.Identità
Temi.Personaggi
Temi.Storia
Temi.Relazioni
Temi.Sociali.e.Culturali
Temi.Conflitti
Temi.Criminalità
Temi..Crescita.e.della.maturazione
Temi.Migrazione
Temi.Guerra
Temi.Tesimonianze
Temi.Viaggio
Temi.Donne
Temi.Passato



Goals

1

Identify and select **pertinent variables** to incorporate into our dataset, evaluating their potential for meaningful predictive capabilities

2

Try to estimate if a movie had a **positive or negative rating** in order to discern if the audience actually enjoyed the movie

Rating Classification

3

Make predictions about the **revenue** of a movie to better understend how much interest it was able to capture

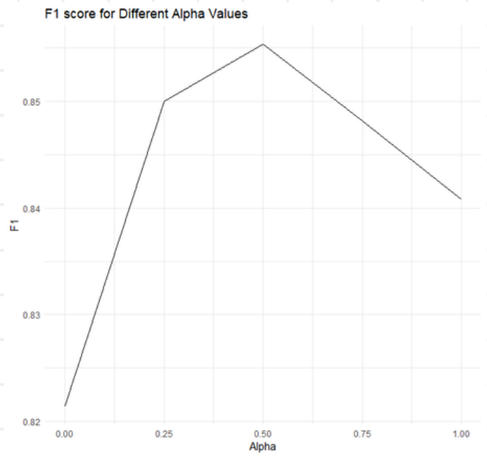
Revenue Regression

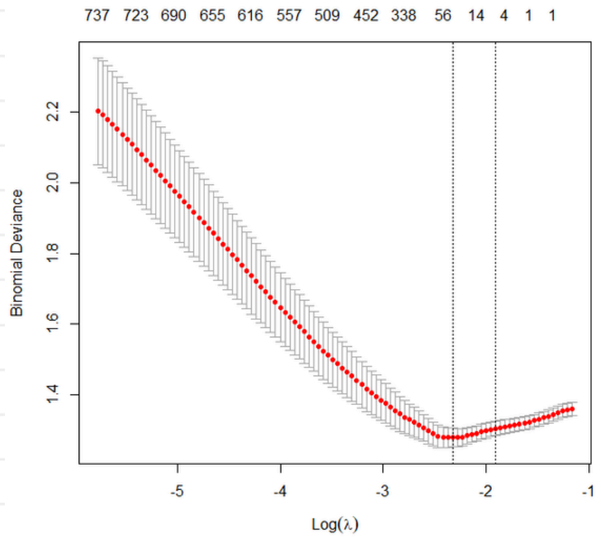
Classification Models

- LASSO
- Ridge
- Elastic-net
- Group LASSO
- LASSO SVM
- KNN
- Random forest

Ridge, Lasso, Elastic-Net

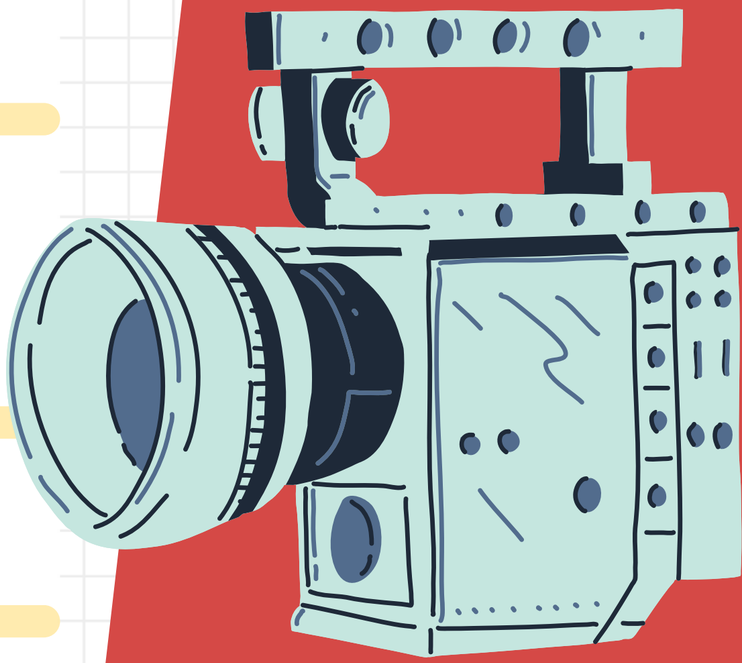
- Various alpha regularization values are employed
- A **20-fold cross-validation** approach is used to select the best lambda
- **Elastic-Net** (with alpha=**0.5**) emerges as the top performer based on the F1-score





- Impact of **selected variables** in the prediction

```
> coefficients_list[["coefficients_alpha_0.5"]]
                                          s0
(Intercept)                      0.787978987
Awards.Nastri                    0.977196367
Cast.Corrado.Fortuna            -0.198072642
Cast.Miriam.Leone               -0.362478240
Cast.Nina.Torresi               -0.201890774
Cast.Paz.Vega                   -0.045515554
CastEdoardo.Leo                 -0.181887996
CastStefania.Sandrelli           0.063004284
GenreAzione                     -0.094934140
GenreCommedia                   -0.707959658
GenreDocumentario                0.195363894
Nominations.Cannes               0.003157114
Nominations.EU                   0.042374545
Nominations.Locarno              0.120240171
Place_setting.Lombardia          0.082589901
Place_setting.Roma              -0.466102750
Place_settingBari               -0.609721098
ProductionCinecittà..Luce        0.087083854
ProductionI.Wonder.Pictures      0.273729268
ProductionNexo.Digital           0.149799904
ProductionNotorious.Pictures    -0.331945150
ProductionWarner.Bros.Italia    -0.793713496
Temi.Alcol.e.Droga              -0.042904892
Temi.Celebrità                  -0.281310132
Temi.Commedia.e.Umorismo        -0.359335369
Temi.Emotivi.e.Psicologici      -0.354221615
Temi.Frammentazione             -0.362559119
Temi.Fuga                       -0.198086945
Temi.Indagine.personale         -0.043601240
Temi.Libertà                     0.282613561
Temi.Musica                      0.270853020
Temi.Segreti                    -0.263051399
TemiOlocausto                   -0.450215282
Time_setting.Contemporaneo      -0.303559294
```



Group Lasso

- **Groups** created with **features** that had been **separed** during the **one hot encoding** to see the importance of the whole category
- Used a **5-fold cross validation** to select the best lambda
- Performance metrics closely aligned with other models

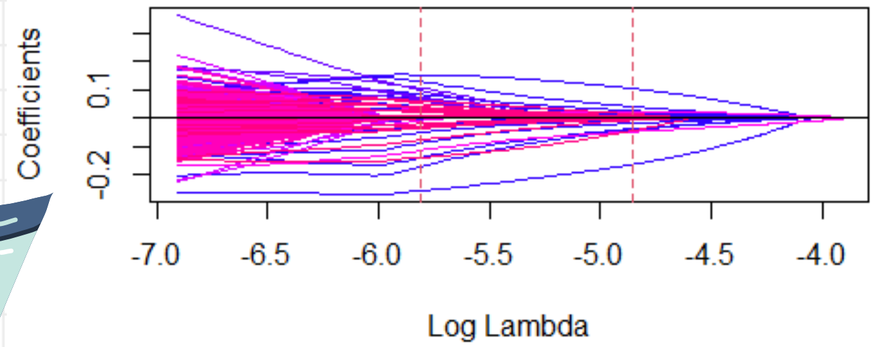
Significant variables

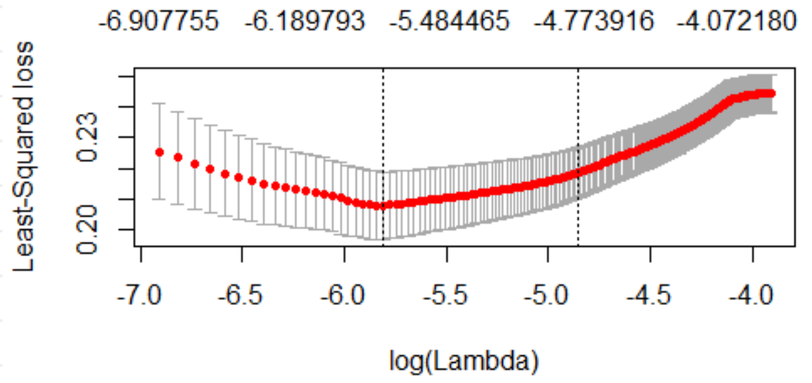
Group Lasso model at **one-standard-error lambda**:

- Genre
- Time Setting
- Suggested Audience

Group Lasso model at **minimum lambda**:

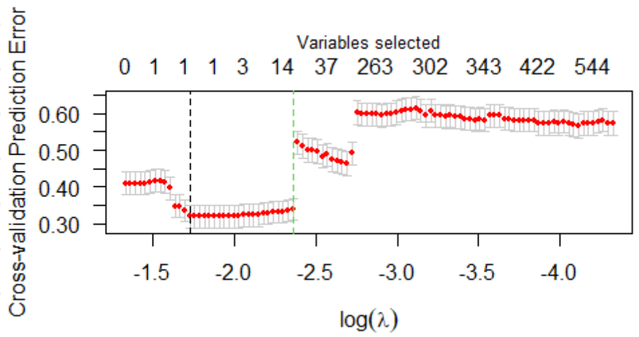
- Genre
- Time Setting
- Nomination
- Suggested Audience

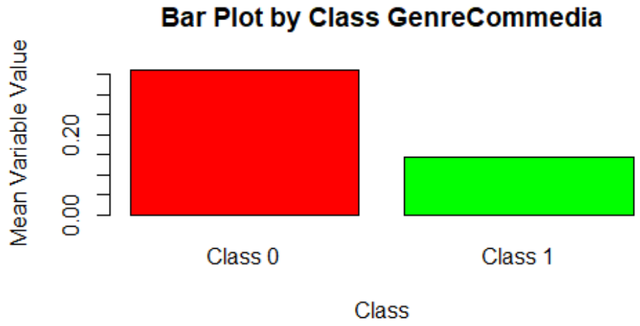




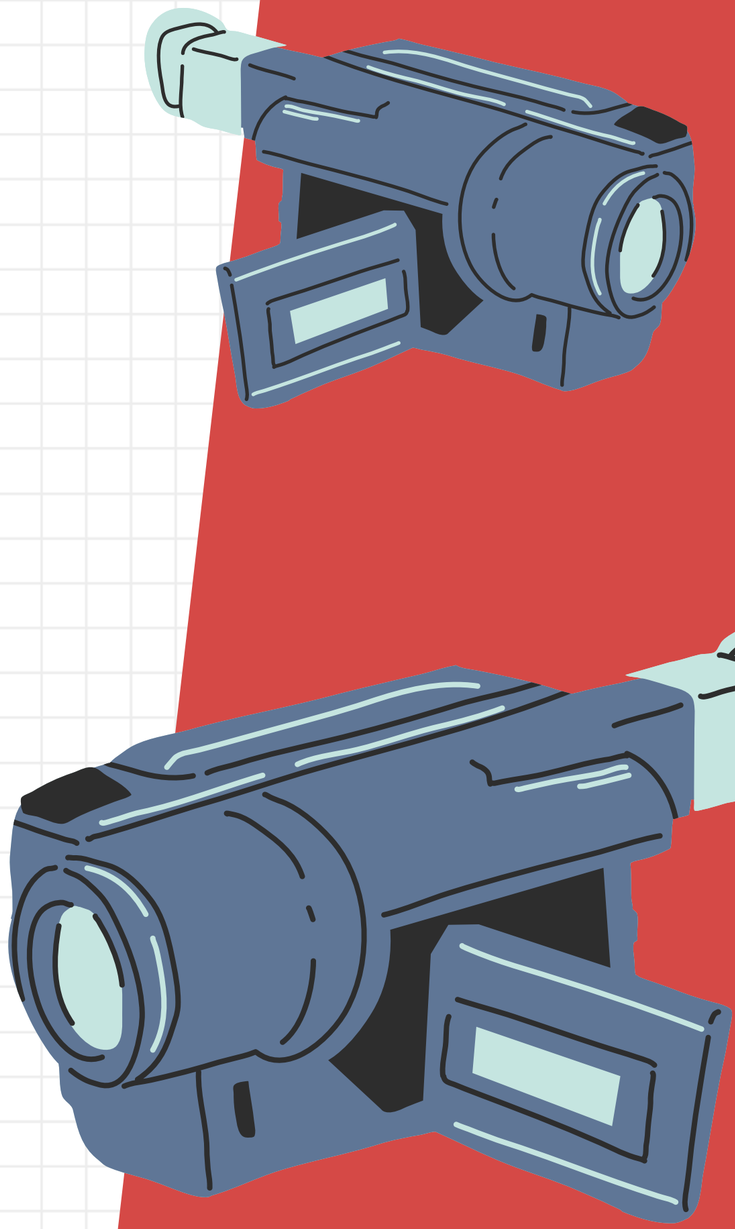
Lasso SVM

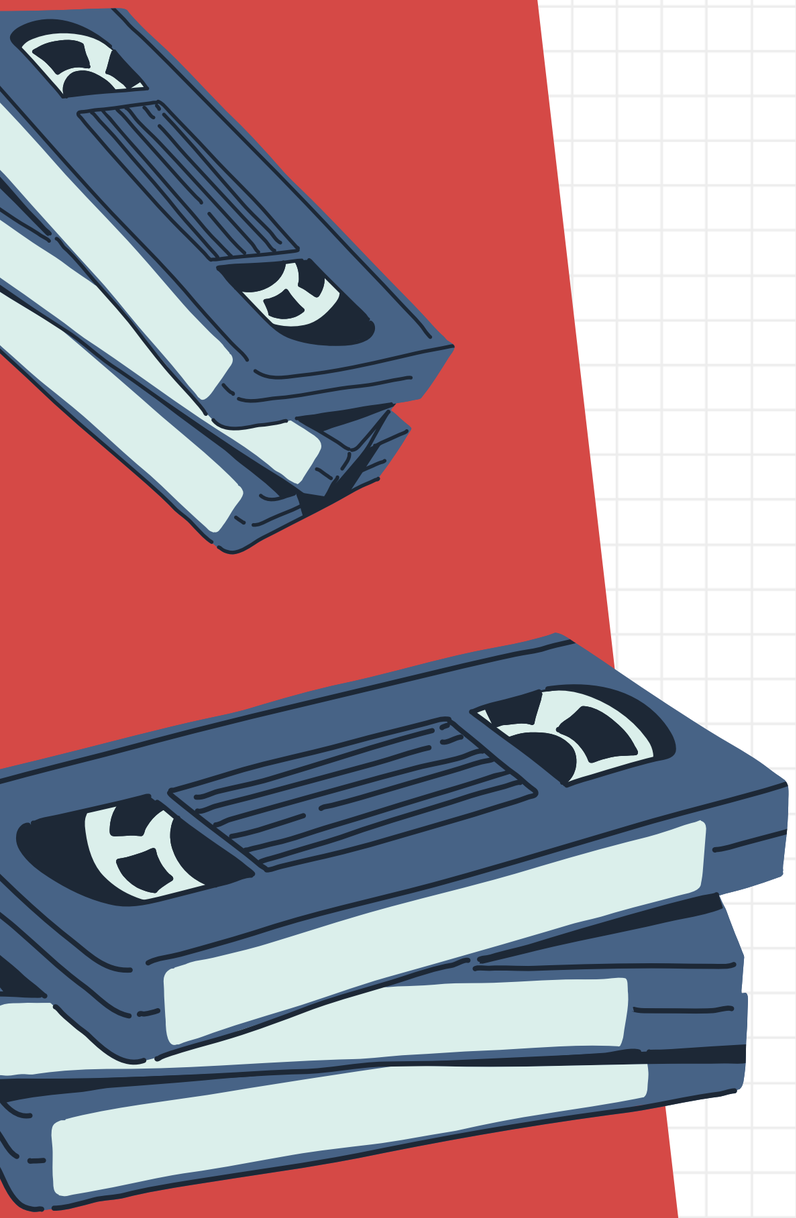
- We used a **20-fold cross validation** to select lambda
- By selecting the **lambda** using the **one-standard-error rule**, the following **correlations** were observed with the **class 0**:
- With the **minimum lambda** it identified just **one variable** as positive correlated with class 0:**GenreCommedia**





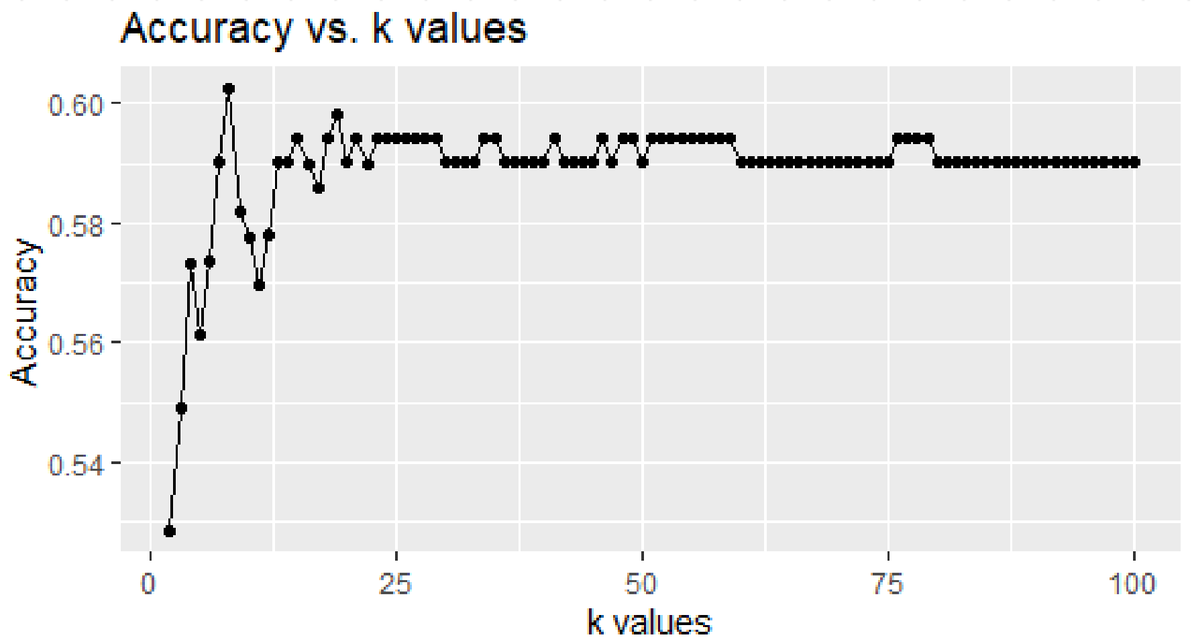
(Intercept)	Awards.Nastri	Cast.Miriam.Leone
-1.001917416	-0.047553461	0.046248383
GenreCommedia	GenreDocumentario	Place_setting.Roma
1.929573630	-0.011954935	0.073055614
Place_settingBari	ProductionI.Wonder.Pictures	ProductionNotorious.Pictures
1.920169335	-0.012841373	0.006478928
ProductionWarner.Bros.Italia	Temi.Commedia.e.Umorismo	Temi.Emotivi.e.Psicologici
0.076575761	0.030919223	0.034924408
Temi.Libertà	Temi.Musica	Temi.Tradimento
-0.022853709	-0.023460701	0.001316956
Time_setting.Contemporaneo		
0.021949177		



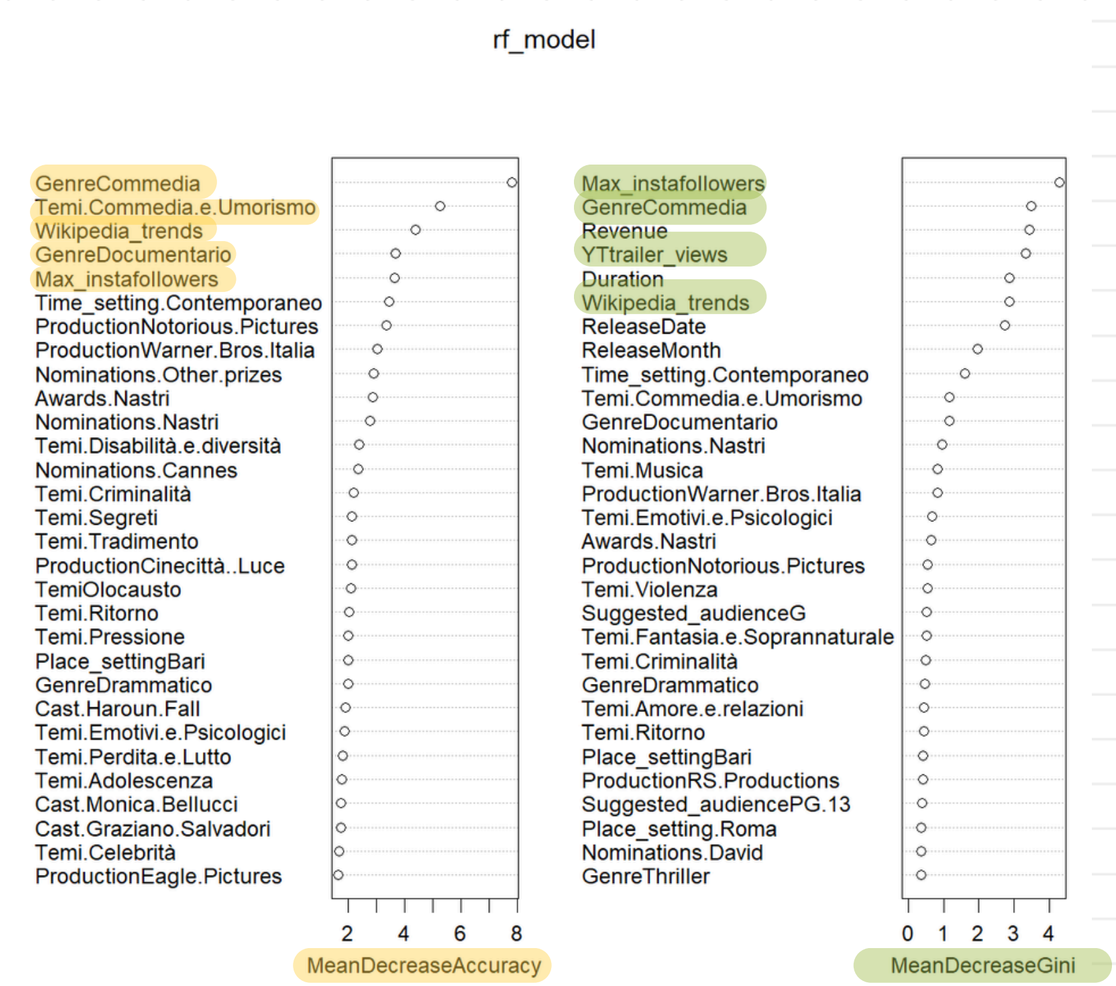


KNN

- We used a **5-fold cross-validation** technique
- It identified **7 neighbours** as the **optimal parameter** choice to **maximize accuracy**.
- Provided acceptable results but showed **slightly inferior performance** compared to other models.
- Unfortunately, KNN does **not offer insights** into **variable selection**, limiting its interpretability.



Random forest



- 200 decision tree
- mtry = 96 (tuned)
- The variable GenreCommedia has the most significant impact on the accuracy of the model
- Max_instafollowers is the variable that better ensures purity

Classification Results

	Accuracy Precision	F1 Recall
Ridge $(\alpha = 0; \lambda = 1.562)$	0.7115385 0.7187500	0.8214286 0.9166667
Elastic Net $(\alpha = 0.25; \lambda = 0.186)$	0.7692308 0.7727273	0.8500000 0.9444444
Elastic Net $(\alpha = 0.5; \lambda = 0.098)$	0.7788462 0.7816092	0.8553459 0.9444444
Elastic Net $(\alpha = 0.75; \lambda = 0.068)$	0.7692308 0.7790698	0.8481013 0.9305556
Lasso $(\alpha = 1; \lambda = 0.051)$	0.7596154 0.7764706	0.8407643 0.9166667
Grouped Lasso	0.7980769 0.9305556	0.8645161 0.8072289
Lasso SVM	0.7692308 0.8888889	0.8421053 0.800000
KNN	0.7211538 0.7361111	0.7851852 0.8412698
Random Forest	0.7884615 0.7906977	0.8607595 0.9444444

- The analysis highlighted the efficacy of generalized linear models
- the Group Lasso regularization technique outperformed others in classification tasks

Regression Models

- LASSO
- Ridge
- Elastic-net
- Adaptive LASSO
- GAM Stepwise
- Sparse GAM

LASSO, Ridge, Elastic-Net

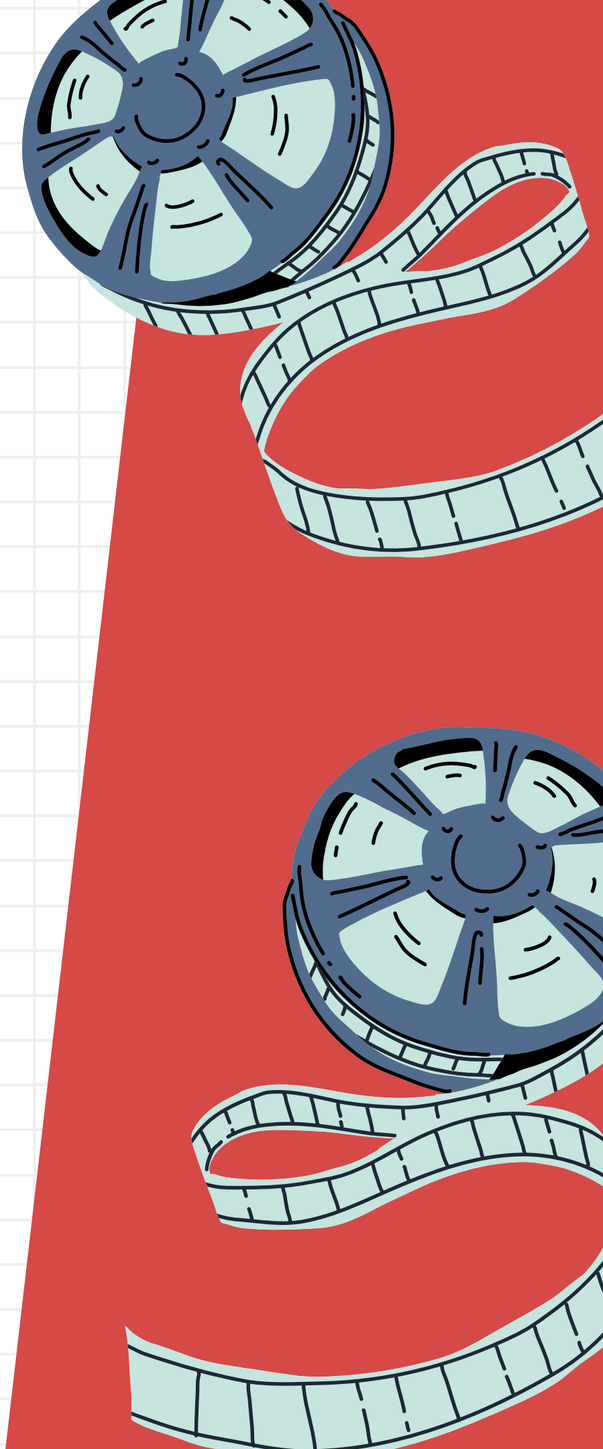
- We used the package '**cv.glmnet**'to **optimize hyperparameters**.
- The different models demonstrated **similar performance** (around **1.2 RMES, 1 MAE**) except for **Ridge** that had **slightly lower** metrics (**1.7 RMES, 1.4 MAE**)

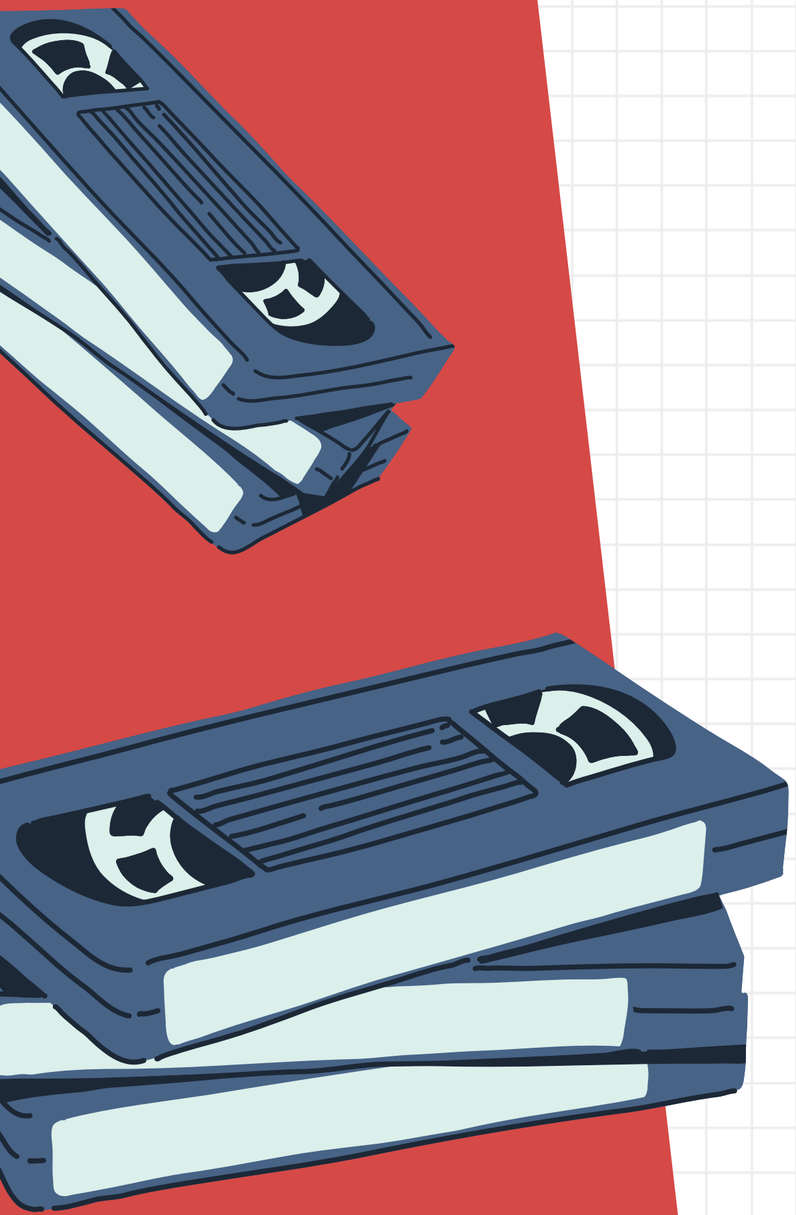
LASSO

	s1
Cast.Amedeo.Grieco	2.4969315
Cast.Antonella.Carone	1.0397970
Cast.Doodou.Sagna	1.9222666
Cast.Elena.Lietti	1.2771620
Cast.Elena.Pozzallo	2.1926969
Cast.Lorenzo.McGovern.Zaini	0.5878703
Cast.Lydia.Page	0.8821355
CastAntonio.Albanese	0.9196683
CastVasco.Rossi	1.0104920
Director.Giulio.Boato	0.5327103
Production01.Distribution	0.5606721
ProductionBim.Distribuzione	0.8063504
ProductionMedusa	1.4707211
ProductionVision.Distribution	0.7092657
ReleaseDate	6.3075404
Temi.Frammentazione	0.8734209
Temi.Maschio	0.4366516
Temi.Redenzione	0.5572576
Time_settingdiciannovesimo.secolo	0.4072725
YTtrailer_views	0.4249537

Elastic-net

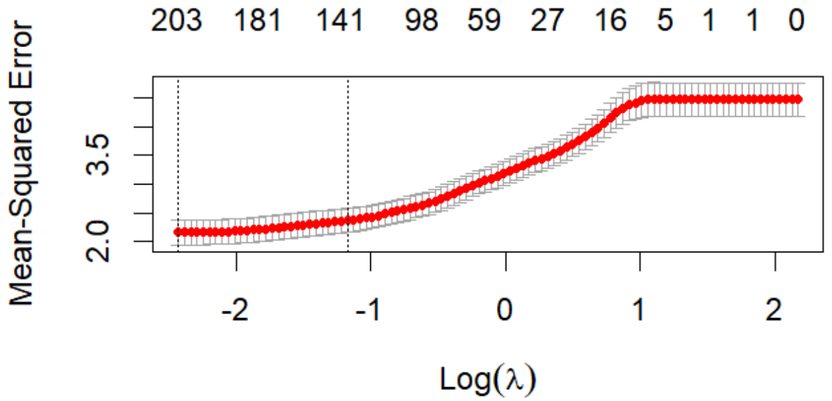
	s1
Cast.Amedeo.Grieco	2.9768348
Cast.Doodou.Sagna	2.3445318
Cast.Elena.Lietti	2.2700457
Cast.Elena.Pozzallo	2.4867979
Cast.Emanuela.Fanelli	1.9418172
Cast.Ettore.Scarpa	0.6806181
Cast.Fabio.Balsamo	2.2815017
Cast.Luigi.Calagna	1.3941504
Cast.Margherita.Buy	0.7545145
Cast.Max.Tortora	0.9864009
CastAntonio.Albanese	0.6732092
CastClaudio.Bisio	0.5851859
CastPierfrancesco.Favino	0.6788790
CastStefania.Sandrelli	0.5079706
Director.Giulio.Boato	0.8227161
DirectorFrancesco.Patierno	0.8252921
DirectorMarco.Pianigiani	0.7157033
DirectorMichele.Mally	0.5986423
DirectorRiccardo.Milani	1.5164882
Place_setting.Toscana	0.5301439
ProductionBim.Distribuzione	0.7240892
ProductionMedusa	0.8776504
ReleaseDate	13.1178437
Temi.Maschio	0.7035979
Temi.Redenzione	0.8186232





Adaptive LASSO

- Adaptive LASSO dynamically adjusts penalties on predictors, prioritizing significant variables while reducing the impact of less influential ones.
- Penalization parameters for Adaptive LASSO were derived through a comprehensive grid search of 1000 lambdas using ridge regression
- Effectively identified and emphasized the most impactful predictors crucial for predicting movie revenue



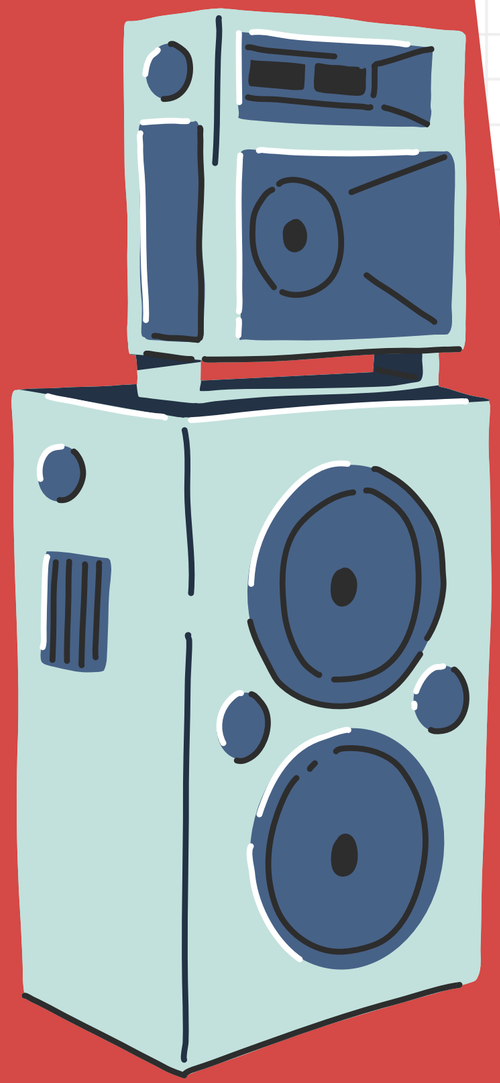
	s1
Cast.Amedeo.Grieco	2.9768348
Cast.Doodou.Sagna	2.3445318
Cast.Elena.Lietti	2.2700457
Cast.Elena.Pozzallo	2.4867979
Cast.Emanuela.Fanelli	1.9418172
Cast.Ettore.Scarpa	0.6806181
Cast.Fabio.Balsamo	2.2815017
Cast.Luigi.Calagna	1.3941504
Cast.Margherita.Buy	0.7545145
Cast.Max.Tortora	0.9864009
CastAntonio.Albanese	0.6732092
CastClaudio.Bisio	0.5851859
CastPierfrancesco.Favino	0.6788790
CastStefania.Sandrelli	0.5079706
Director.Giulio.Boato	0.8227161
DirectorFrancesco.Patierno	0.8252921
DirectorMarco.Pianigiani	0.7157033
DirectorMichele.Mally	0.5986423
DirectorRiccardo.Milani	1.5164882
Place_setting.Toscana	0.5301439
ProductionBim.Distribuzione	0.7240892
ProductionMedusa	0.8776504
ReleaseDate	13.1178437
Temi.Maschio	0.7035979
Temi.Redenzione	0.8186232

Gam stepwise

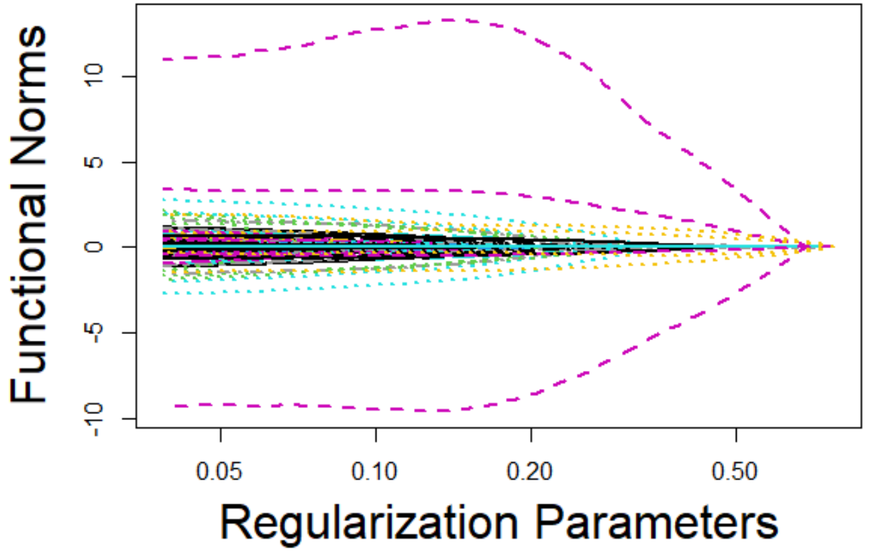
- Utilizing stepwise variable selection via the 'step.Gam' function, we identified the most relevant predictors significantly contributing to revenue variability.
- Due to our dataset's high dimensionality, the stepwise variable selection process faced challenges. Multiple trials with subsets were conducted to strike a balance between AIC and predictive power.
- By employing stepwise variable selection in GAM, we extracted key predictors, shedding light on the factors crucial for success and profitability in the film industry.

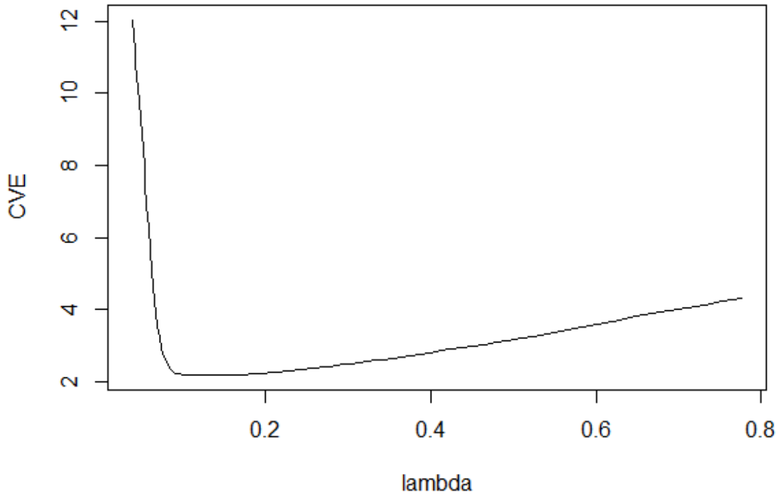
```
Anova for Parametric Effects
                             Df Sum Sq Mean Sq  F value    Pr(>F)
Duration                      1  31.12  31.120  19.2184 1.745e-05 ***
ReleaseDate                   1   0.29   0.287   0.1772 0.6742007
s(Rating_Average, df = 2)     1   0.20   0.201   0.1243 0.7247017
s(Wikipedia_trends, df = 5)   1 131.85 131.854  81.4266 < 2.2e-16 ***
s(YTtrailer_views, df = 5)    1 204.16 204.162 126.0799 < 2.2e-16 ***
s(Nominations.Nastri, df = 5) 1   5.41   5.409   3.3404 0.0688398 .
Nominations.Venezia           1  24.08  24.083  14.8726 0.0001478 ***
Residuals                   240 388.63   1.619
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects
                             Npar Df  Npar F     Pr(F)
(Intercept)
Duration
ReleaseDate
s(Rating_Average, df = 2)          1  5.6456  0.018285 *
s(Wikipedia_trends, df = 5)        4 11.2584 2.177e-08 ***
s(YTtrailer_views, df = 5)         4  4.5006  0.001594 **
s(Nominations.Nastri, df = 5)      4  4.0206  0.003557 **
Nominations.Venezia
```

Sparse GAM



- Sparse GAM, leveraging the sparseGAM package, combines GAM's flexibility with penalties that induce sparsity in variable selection.
- Prioritizes essential predictors, contributing to a more efficient model





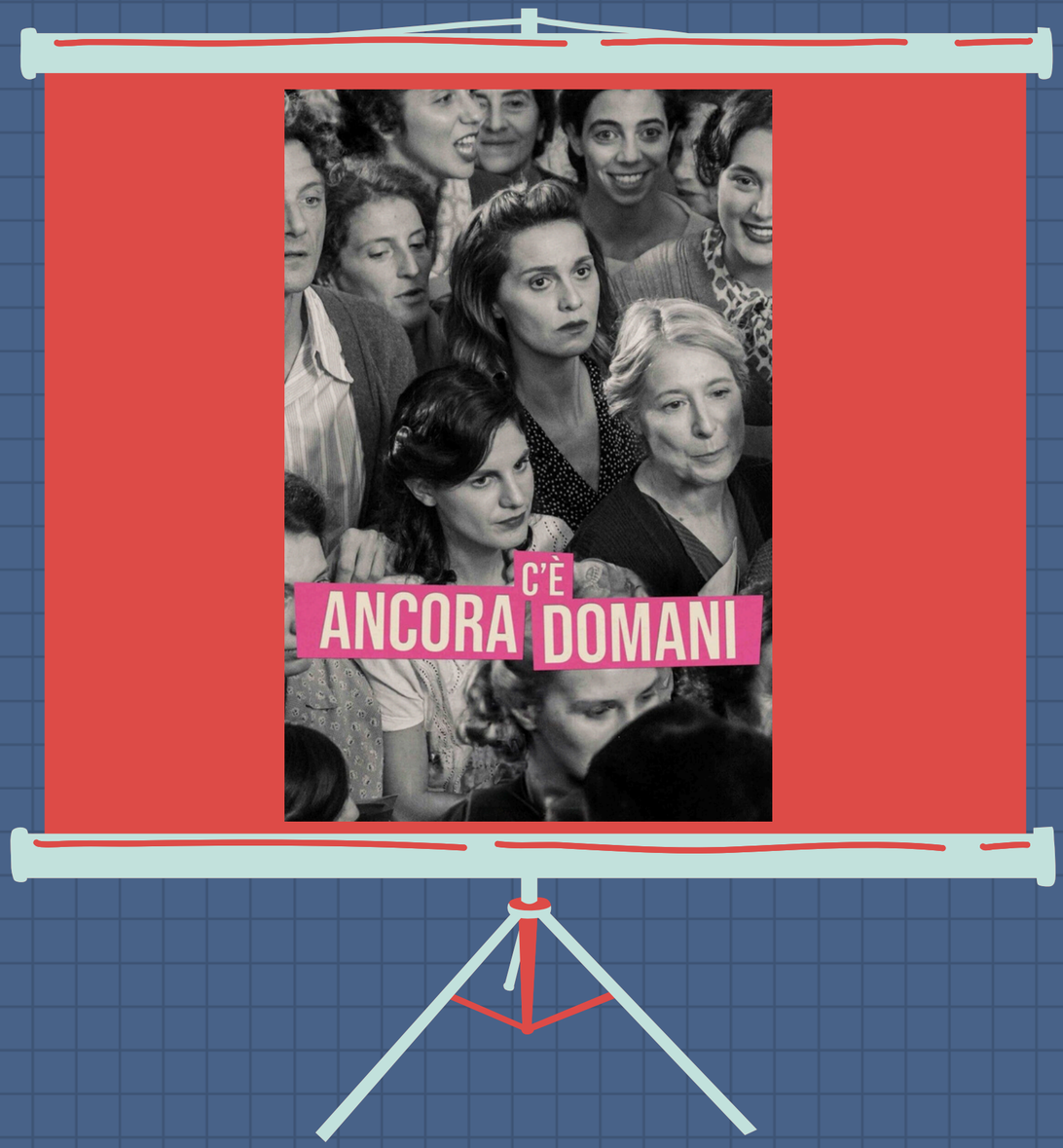
Relevant features:

- **Wikipedia_trends**
- **YTubeTrailer_views**
- **Awards.Venezia**

Regression results

	RMSE	MAE
Ridge ($\alpha = 0; \lambda = 0.01$)	1.747079	1.426744
Elastic Net ($\alpha = 0.7; \lambda = 0.1066203$)	1.259231	1.006204
Lasso ($\alpha = 1; \lambda = 0.0858112$)	1.263311	0.999402
Adaptive Lasso ($\lambda = 0.2351088$)	1.300599	1.053719
GAM Stepwise	1.458371	1.05606
Sparse GAM ($\lambda = 0.14$)	1.144435	0.9373917

- Elastic Net and Sparse GAM stand out as the most notable models, providing distinct and valuable insights into the underlying features



C'è ancora domani

Rating

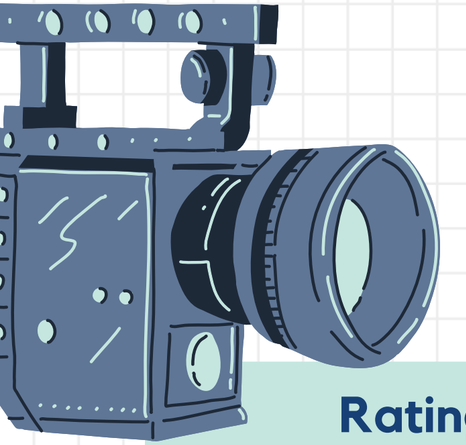
Revenue : 32.249.958€

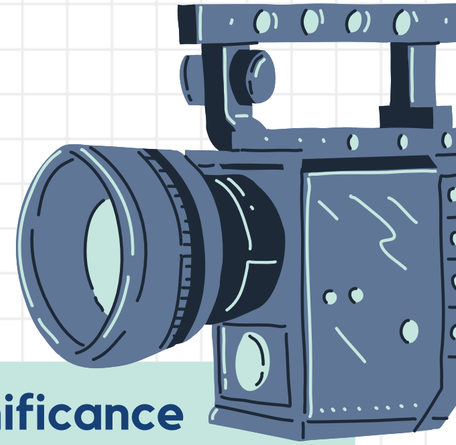
Director : Paola Cortellesi

Themes :

- Postwar
- Role of women
- Domestic violence

Conclusions





Ratings

Group Lasso regularization emerged as the most effective technique for predicting movie success, surpassing other methods in classification tasks.

Revenue

Elastic Net and GAM offered nuanced perspectives on features, slightly distinguishing themselves from each other, enriching the understanding of movie success factors.

Limitation and challenges

- Missing movie budget data
- Multiclass classification faced challenges due to prevalent average ratings,
- Outliers, like *C'è ancora domani*, posed dilemma in considering exemplary dataset samples against privileging performance.

Significance of Curated Variables

Variables providing additional information that we decided to add in our dataset displayed significance across various models.

Our primary aim was to evaluate the utility of specific added variables for prediction. We are satisfied with the outcome as these variables were consistently selected by various models, affirming their significance in predicting movie success.

Thank you for the attention