
Machine Learning
aplicado al mundo
del cine
¿Puede un algoritmo predecir la nota de una película?
Rubén Peinado

Objetivos del trabajo

- Lanzamiento de nuevas películas.
- Programación de la cartelera de cine en una sala de proyección de películas:
  - Queremos proyectar películas que dispongan de una determinada nota mínima.
  - Predecir la nota que puede tener una película.
  - Fidelizar a un público en concreto para nuestra sala de cine.
  - Destacarnos de otras salas de cine convencional.

Estructura del trabajo

01

Análisis de datos

Fuente de datos
Limpieza y visualización

02

Selección de algoritmos

Preprocesamiento de datos
Ajuste hiperparámetros

03

Evaluación del modelo

Métricas de evaluación
Interpretación resultados

04

Conclusiones

Limitaciones del modelo
Posibles mejoras

Fuente de datos

Filmaffinity

Sistema recomendador de cine con una base de datos en la que se encuentra la ficha completa (técnica y artística) de gran cantidad de películas.

Listado de películas

Más de 10.000 películas

Ficha técnica y artística

10 categorías

Premios y Festivales

47 festivales y
278 categorías

Limpieza de datos - features

Numéricas

★★★★★★★★★★

2/10

Año y duración.

Categóricas

★★★★★★★★★★

8/10

Título, pais, director, guión, música, fotografía, género y actores.

Preprocesamiento de datos

Categorias

Año
Discretización (cine mudo, sonoro, technicolor y digital)

Film
Total premios y premios Razzie

Géneros
get_dummies

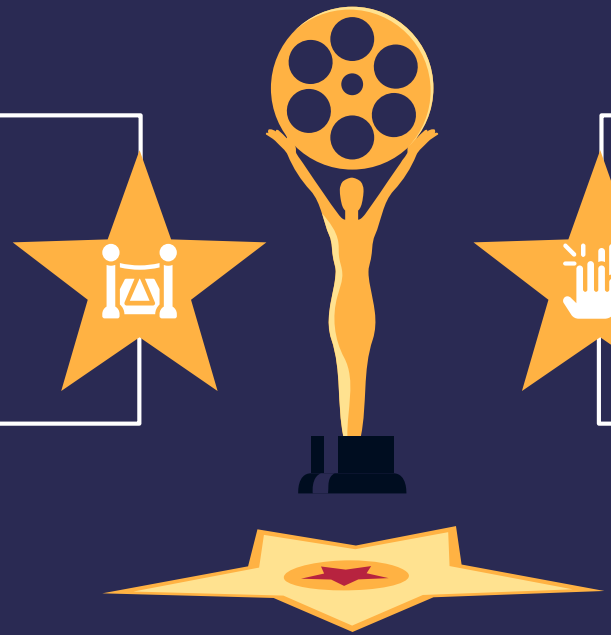
Duración
Discretización y eliminación outliers

País
Discretización por frecuencia

Película: The Artist

2 Oscars

1 Globos de Oro



César 2

BAFTA 1

Actriz: Meryl Streep

3 Oscars

8 Globos de Oro

Satellite Awards 3

BAFTA 2

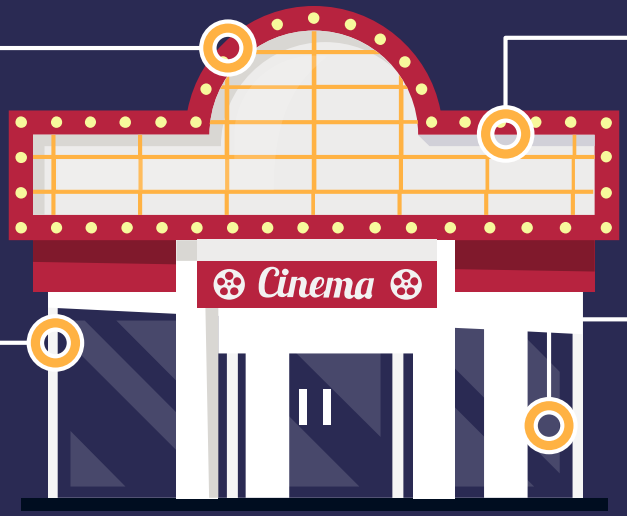
Selección de algoritmos

Regresión

LinearRegression, Lars y Huber

Regularización

Ridge, Lasso, ElasticNet y BayesianRidge



Basados en árboles

RandomForest, DecisionTree, CatBoost, LightGBM, GradientBoosting, XGB y AdaBoost

Otros

Kneighbors, Support Vector Machine, MLPRegressor

Elección del modelo

Paso 1

Comparativa de resultados (métricas)

14 modelos diferentes

Paso 2

Voting con el mejor de cada grupo

Ridge, LinearRegressor, Huber, CatBoost, SVM y MLP

Paso 2

Combinación y Ponderación

Combinar/ ponderar varios modelos

Evaluación del modelo

Métricas

Utilizaremos las métricas más apropiadas para la decisión de eficacia del modelo según la fase del trabajo.

R2 Score

Medida de la performance del modelo.

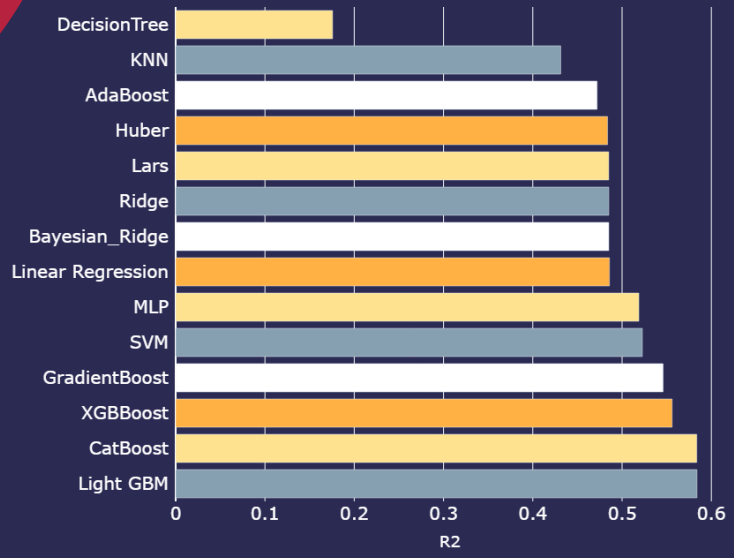
MAE

Medida del error absoluto en las predicciones.

RMSE

Medida que señala un menor error promedio.

Comparativa de modelos -Inicio



R2 Score

CatBoost
0.58
Mejor MAE

SVM
0.52

Bayesian Ridge
0.48

Linear Regressor
0.48

Mejores modelos

CatBoost
Basados en árboles
★★★★★★★★★★ 9/10

Support Vector Machine
Otros
★★★★★★★★★★ 7/10

Bayesian Ridge
Regularización
★★★★★★★★★★ 6/10

Linear Regressor
Regresión
★★★★★★★★★★ 6/10

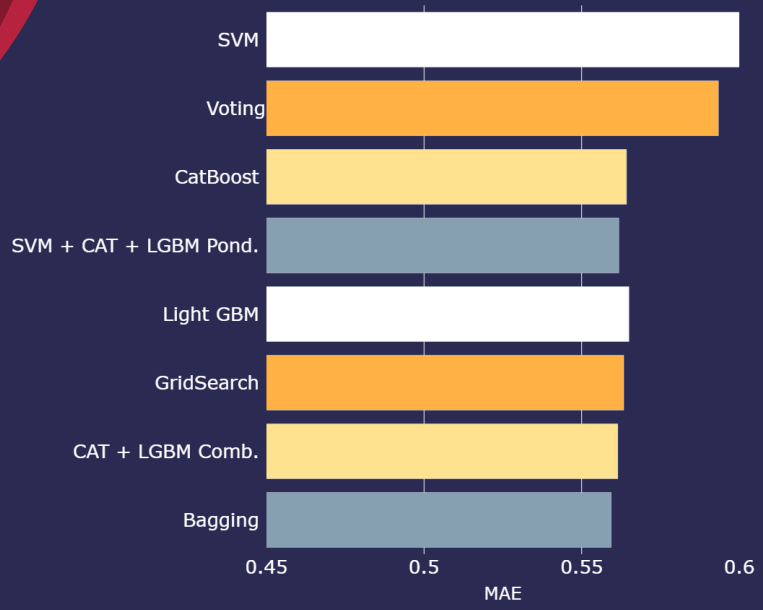
Optimización del modelo

CatBoost

Potente y flexible, procesamiento eficiente de grandes volúmenes de datos, regularización incorporada y herramientas de interpretación.

- Bagging
  - CatBoost model
- Features Selection
  - Select Kbest, features importances, RFE
- GridSearch
  - Best_params_

Comparativa de modelos - Optimizados



MAE

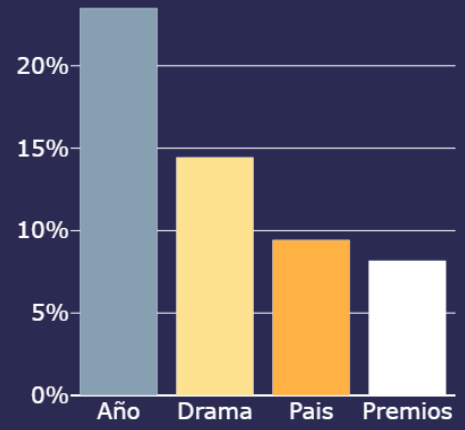
Bagging
0.559

CatBoost + LGBM
0.561

GridSearch
0.563

Light GBM
0.565

Features Importances



23%

Año

Preferencia de cine clásico

14%

Drama

Género cinematográfico predominante

9%

Pais

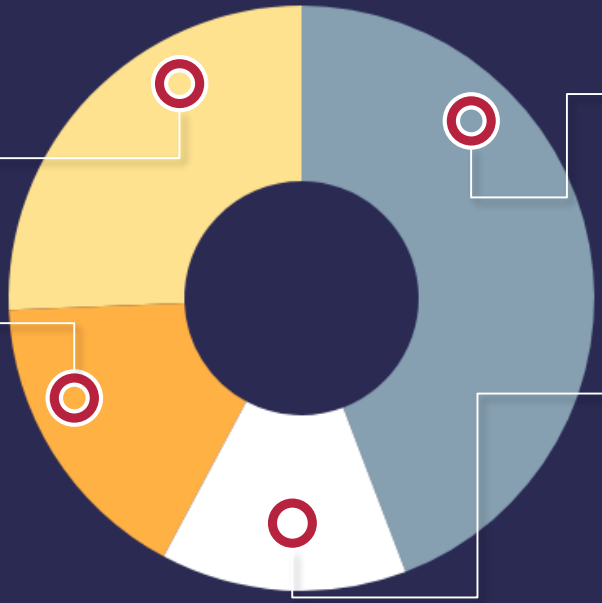
Feature desbalanceada

8%

Total premios

Reconocimiento de la película

Select KBest



14%
Drama
Género cinematográfico predominante

9%
Director
Figura técnica más importante

25%
Año
Preferencia cine clásico

8%
Total premios
Reconocimiento de la película

Recursive Feature Elimination

Features

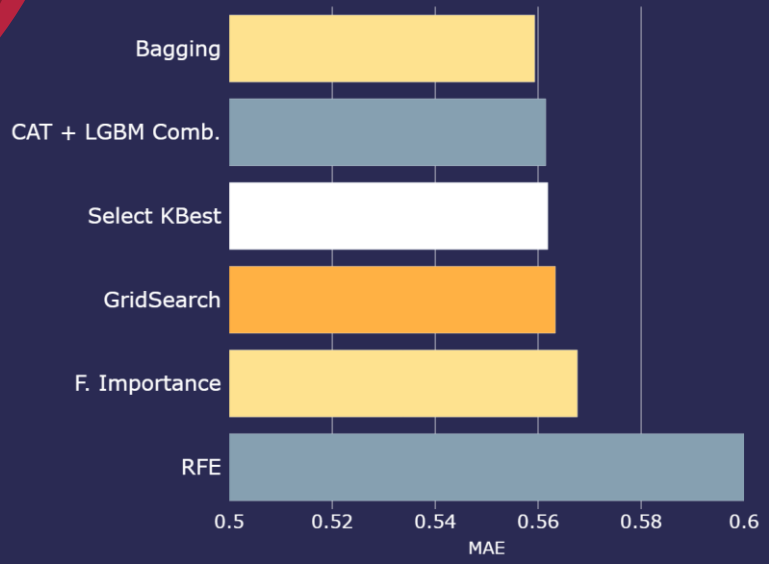
Total premios
Reconocimiento de la película

Año
Preferencia cine clásico

Drama
Género cinematográfico predominante

Pais
Feature desbalanceada

Comparativa de modelos – Features Select.



MAE

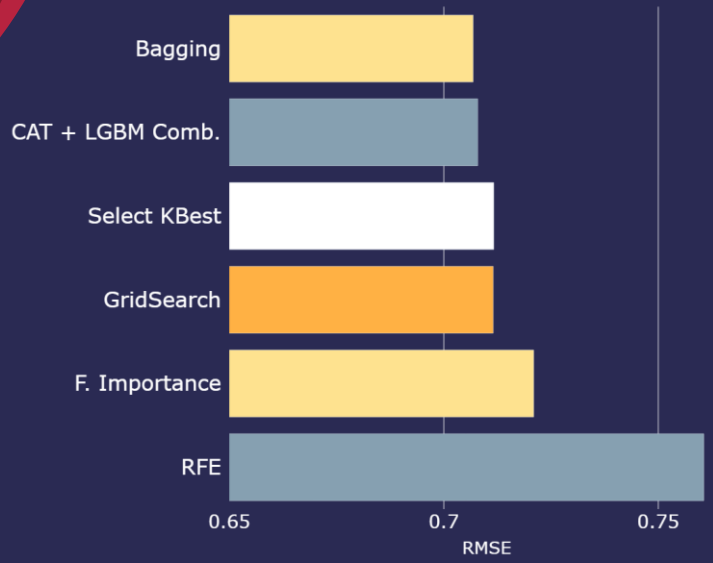
Bagging
0.5593

CatBoost + LGBM
0.5614

Select KBest
0.5618

Grid Search
0.5633

Comparativa de modelos – Features Select.



RMSE

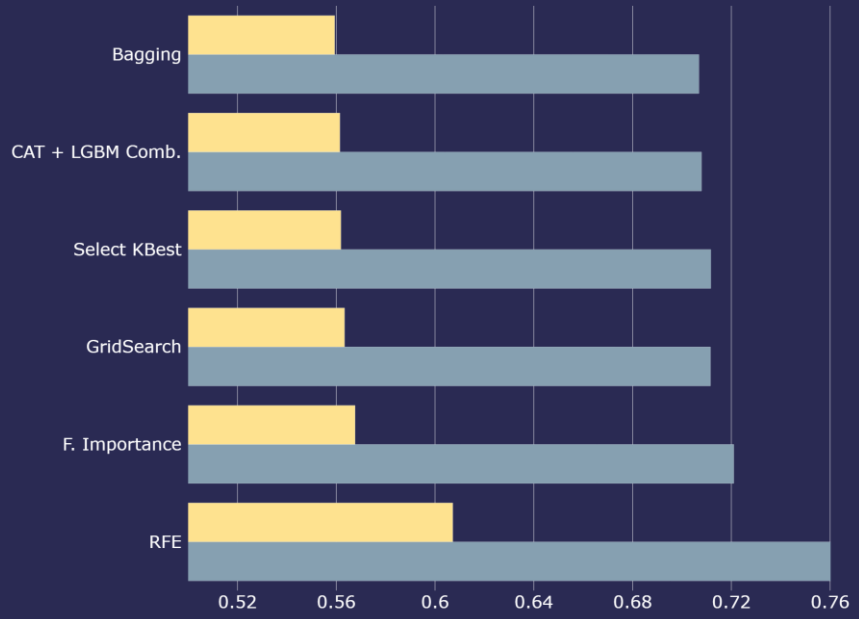
Bagging
0.7068

CatBoost + LGBM
0.7079

Select KBest
0.7116

GridSearch
0.7115

Comparativa de modelos – Features Select.



RMSE/ MAE

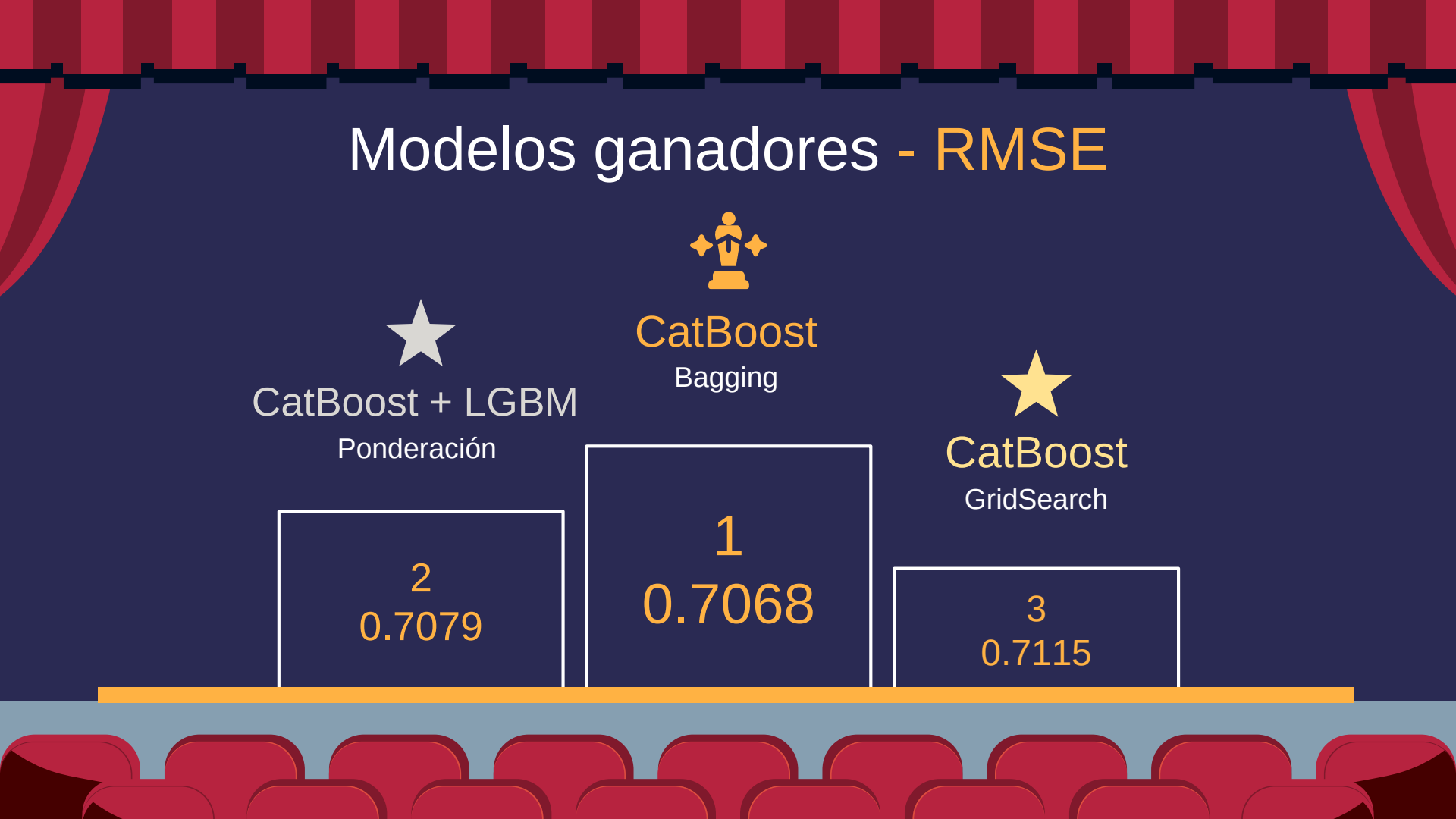
Bagging

CatBoost + LGBM

Select KBest

GridSearch

Modelos ganadores - RMSE



Conclusiones del trabajo

	Resumen	Mejoras/ próximos pasos
Data Analysis	Una única fuente facilita el proceso pero limita la obtención de nuevas features.	Ampliar fuentes de datos para mayor información. Enriquecer el dataframe con nuevas fuentes.
Feature Engineering	Preprocesado mixto (automático/ manual). Favorece el trabajo del modelo.	Establecer nuevas estrategias para sacar partido a las features iniciales.
Machine Learning	Se mejora la performance con la optimización. Ponderación de modelos, bagging o feature selection.	Establecer nuevo punto de vista de resolución del problema. Clasificación multiclase.

Preguntas y sugerencias

Gracias por vuestra atención!!!