



GOOGLE PLAY STORE SUCCESS PREDICTION

MODELO DE MACHINE LEARNING PARA PREDECIR EL ÉXITO DE APLICACIONES MÓVILES.

ALEJANDRO PACHECO BASTIDAS – PROYECTO FINAL BOOTCAMP DATASCIENCE UDD

PROBLEMA

Pregunta del negocio:

¿Podemos predecir si una app será exitosa antes de lanzarla?

Definición de éxito:

Rating ≥ 4

Importancia:

- Reduce riesgo de inversión
- Apoya decisiones estratégicas
- Permite priorizar desarrollo

DATASET

Fuente: Kaggle – Google Play Store Apps

Variables utilizadas:

- Installs
- Price
- Size
- Category
- Content Rating
- Type

Cantidad de registros: ~9.300 apps

PROBLEMA TÉCNICO

El dataset presenta:

- Distribuciones altamente asimétricas
- Desbalance moderado entre clases

Decisiones tomadas:

- Transformaciones logarítmicas
- Uso de F1 Score como métrica principal

MODELOS ENTRENADOS

1. Logistic Regression
2. Random Forest (ensamble)
3. Random Forest + Tuning (GridSearchCV)

Modelo	Accuracy	F1
LogReg	0.64	0.73
RF	0.74	0.84
RF Tuned	0.75	0.84

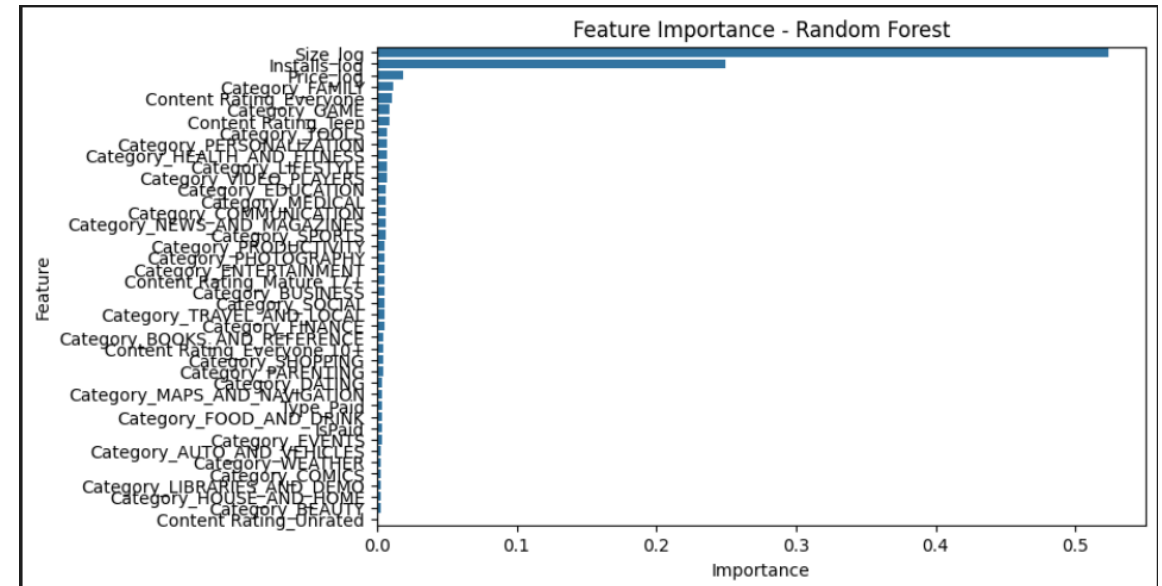
Random Forest tuneado fue el modelo seleccionado.

VARIABLES MÁS IMPORTANTES

Las variables más influyentes fueron:

- Size_log
- Installs_log

El tamaño y la popularidad tienen fuerte impacto en la probabilidad de éxito.



API y DESPLIEGUE

Arquitectura: El modelo fue desplegado en una API REST funcional

Usuario → API Flask → Modelo → Respuesta

Entrada:

```
{
  "Installs": 5000000,
  "Price": 0,
  "Size_KB": 20000,
  "Category": "GAME",
  "Content_Rating": "Teen",
  "Type": "Free"
}
```

Salida:

```
{
  "prediction": 1,
  "probability_success": 0.82
}
```

CONCLUSIÓN FINAL

- El modelo de ensamble mejora significativamente al modelo base.
- El tuning optimiza desempeño
- Se logró un pipeline completo:
 - EDA
 - Modelado
 - Evaluación
 - Ajuste
 - Despliegue

El proyecto demuestra un flujo completo de Machine Learning desde datos crudos hasta producción.



GRACIAS

LA TECNOLOGÍA ES
IMPORTANTE, PERO LO
ÚNICO QUE REALMENTE
IMPORTA ES QUÉ
HACEMOS CON ELLA.
(MUHAMMAD YUNUS)