Recomendación de películas

Para nuestro trabajo seleccionamos un dataset con información de 10 002 películas. A partir dicha elección formulamos como objetivos caracterizar las variables según la calificación promedio para armar un sistema de recomendación a partir del modelo de aprendizaje supervisado y no supervisado. Planteamos como objetivos:

Objetivo general: elaborar un sistema de recomendación tomando la calificación final de una película como indicador de aceptación por el público.

Objetivo específico 1: analizar la aceptación de películas por protagonista, director, género y país de origen según calificación promedio y una nueva variable, Porcentaje_likes.

Objetivo específico 2: determinar si existe relación, y cómo se comporta, entre las variables vistas, duración y likes con la calificación final.

Para la aplicación de ambos modelos, previamente realizamos análisis exploratorio de datos, selección de variables, elaboración de una nueva variable, análisis y manejo de datos faltantes y outliers, y transformación de datos con One-hot encoding.

Modelo de aprendizaje supervisado

Realizamos estandarización (distribución normal) y normalización (sesgo) de las variables.

Para seleccionar el mejor modelo analizamos Árbol de decisión, Random Forest y regresión lineal. En los primeros dos realizamos ajuste de parámetros y el modelo que mejor predijo nuestros datos fue Random Forest.

Este modelo presentó capacidad de manejar muchas variables y evitar overfitting, maneja variables categóricas, numéricas y de relación no lineales, explicó el 75% de lo que influye en la calificación ($R^2 = 0,75$) y mostró poco margen de error (RMSE = 0,50)

Sería de utilidad para que una app o plataforma recomiende películas según gustos, una empresa sepa qué tipo de películas tienen más posibilidades de gustar y para predecir el éxito de una película antes de su estreno.

Modelo de aprendizaje no supervisado

Realizamos escalado de los datos y eliminamos aquellas variables que tenían alta correlación.

Para seleccionar el mejor modelo analizamos K-Means, DBSCAN y Gaussian Mixture Models. En los tres casos realizamos ajuste de parámetros, solo el último modelo se ajustó a nuestros datos y permitió elaborar un modelo de predicción. Dicho modelo podría realizar:

- *Recomendación basada en una película*

El usuario selecciona una película que pertenece a un cluster específico, se le pueden recomendar otras del mismo cluster.

- *Recomendación basada en perfil de usuario*

Si el usuario ha visto varias películas, visualizamos a que clusters pertenecen, determinamos el cluster dominante en su historial y por último recomendamos películas de ese cluster que aún no haya visto.

Ambos sistemas de recomendación se complementan con ponderar calificación dentro del cluster, utilizar más de un cluster y el usuario elige películas y recibe recomendaciones dinámicas.

Conclusión

A partir de los contenidos aprendidos en clases pudimos cumplir con nuestros objetivos de manera satisfactoria. Como nota personal agregaría que el trabajo final es solamente un pequeño resumen de todo el proceso de elaboración. Todos los resultados son el resumen de largos procesos que incluyeron todo lo aprendido en clases, investigación complementaria e innumerables intentos de prueba y error.