PREDICCIÓN DE IMDB PARA PELÍCULAS

Un Enfoque de Regresión Lineal

GRUPO 19

<u>Integrantes</u>:

- *Quipildor, Lorena*
- *Zamora, Rodrigo*
- *Yede, Gisela*

Predicción de la Clasificación de IMDB para Películas: *Un Enfoque de Regresión Lineal*

Introducción

Contexto del Problema

La clasificación de IMDB es una medida de la calidad de una película que se basa en las calificaciones de los usuarios. Es una métrica importante para los consumidores que buscan películas que les puedan gustar.

Importancia y Relevancia

La predicción de la clasificación de IMDB es una tarea importante porque puede ayudar a los consumidores a encontrar películas que les puedan gustar. También puede ser útil para los productores y distribuidores de películas, que pueden utilizar esta información para tomar decisiones sobre el marketing y la distribución de sus películas.

Objetivos del Proyecto

El objetivo de este proyecto es desarrollar un modelo de aprendizaje automático que pueda predecir la clasificación de IMDB de una película.

Metodología

Datos Utilizados

Los datos utilizados en este proyecto provienen del conjunto de datos de metadatos de películas de Emmanuel Iarussi[1]. Este conjunto de datos contiene información sobre más de 50.000 películas, incluidas sus clasificaciones de IMDB.

[1] – Emmanuel Iarussi (Github)
"https://raw.githubusercontent.com/emmanueliarussi/DataScienceCapstone/master/3_MidtermProjects/ProjectIMDB/data/movie_metadata.csv"

Herramientas y Tecnologías

Las herramientas y tecnologías utilizadas en este proyecto son las siguientes:

- Python
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn

Proceso de Análisis/Desarrollo

El proceso de análisis y desarrollo se llevó a cabo de la siguiente manera:

1. **Limpieza de datos:** Se eliminaron los valores nulos y se corrigieron los errores de datos.
2. **Análisis exploratorio:** Se realizaron análisis estadísticos para comprender la distribución de los datos y la relación entre las variables.
3. **Selección de características:** Se utilizó la prueba F para seleccionar las características más relevantes para la predicción de la clasificación de IMDB.
4. **Modelado:** Se entrenó un modelo de regresión lineal con las características seleccionadas.
5. **Evaluación del modelo:** Se evaluó el rendimiento del modelo utilizando el error cuadrático medio.

Resultados

Presentación de Resultados

El modelo de regresión lineal obtuvo un error cuadrático medio de 0,78. Esto significa que, en promedio, las predicciones del modelo están a 0,78 puntos de la clasificación real de IMDB.

Interpretación

Un error cuadrático medio de 0,78 es un resultado de regular a aceptable. Esto significa que el modelo es capaz de predecir la clasificación de IMDB con un grado de precisión algo razonable.

Discusión

Comparación con objetivos

El objetivo de este proyecto era desarrollar un modelo de aprendizaje automático que pudiera predecir la clasificación de IMDB de una película. El modelo desarrollado cumplió a medias con nuestro objetivo ya que obtuvo un Error cuadrático medio (MSE) de 0,78.

Desafíos y Limitaciones

Uno de los desafíos que se enfrentó en este proyecto fue la falta de datos. El conjunto de datos utilizado es relativamente pequeño, lo que podría limitar el rendimiento del modelo.

Otra limitación del proyecto es que solo se utilizó un modelo de aprendizaje automático. Es posible que otros modelos, como los modelos de aprendizaje profundo, puedan obtener resultados aún mejores. Siendo que el que elegimos nos pareció el más adecuado.

Conclusiones

Reflexiones Finales

Los hallazgos principales de este proyecto son los siguientes:

- Es posible predecir la clasificación de IMDB de una película con un grado de precisión razonable utilizando un modelo de aprendizaje automático.
- El conjunto de datos utilizado en este proyecto es relativamente pequeño, lo que podría limitar el rendimiento del modelo.
- Es posible que otros modelos, como los modelos de aprendizaje profundo, puedan obtener resultados aún mejores.

Aplicación de Conocimientos

Los conceptos de big data y machine learning utilizados en este proyecto son los siguientes:

- Limpieza de datos: se utilizaron métodos de limpieza de datos para eliminar los valores nulos y corregir los errores de datos.
- Análisis exploratorio: se realizaron análisis estadísticos para comprender la distribución de los datos y la relación entre las variables.
- Selección de características: se utilizó la prueba F para seleccionar las características más relevantes para la predicción de la clasificación de IMDB.

- Modelado: se entrenó un modelo de regresión lineal con las características seleccionadas.
- Evaluación del modelo: se evaluó el rendimiento del modelo utilizando el error cuadrático medio.

Sugerencias para Futuras Investigaciones

Se podrían realizar las siguientes investigaciones futuras:

- Utilizar un conjunto de datos más grande para mejorar el rendimiento del modelo.
- Utilizar otros modelos de aprendizaje automático, como los modelos de aprendizaje profundo.
- Investigar la relación entre las variables del conjunto de datos y la clasificación de IMDB.

Entrega de Código y Documentación

Enlace a Google Colab: https://github.com/Grupo19MLBD/ML-BD_TrabajoFinal.git

Documentación del Código: Se proporcionará una documentación clara y comentada del código para facilitar su comprensión y reproducción.

Referencias

Emmanuel Iarussi (archivo .csv en Github)
https://raw.githubusercontent.com/emmanueliarussi/DataScienceCapstone/master/3_MidtermProjects/ProjectIMDB/data/movie_metadata.csv

https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset