Desafío - Preprocesamiento de datos

En este desafío tendrás la oportunidad de poner a prueba los conceptos aprendidos durante la sesión. Los ejercicios están diseñados para reforzar practicar lo explicado en clases y poder implementar un caso real.

Lee todo el documento antes de comenzar el desarrollo individual, para asegurarte de tener el máximo de puntaje y enfocar bien los esfuerzos. Asegúrate de seguir las instrucciones específicas en cada ejercicio y de completar los requerimientos adicionales, si los hubiera. ¡A disfrutar aprendiendo!

Tiempo asociado: 4 horas cronológicas

Descripción

¡Bienvenidos, data scientists y fanáticos del anime, a una tarea de preprocesamiento de datos en el fascinante universo de las series de anime!
En este desafío, trabajarás con un conjunto de datos de episodios y series de anime para prepararlo para análisis y modelamiento predictivo.

El objetivo principal es desarrollar un modelo que pueda predecir la calificación de los usuarios (User Rating) para diferentes animes basándose en las características disponibles. Esta información será utilizada por una plataforma de streaming para mejorar su sistema de recomendaciones y ayudar a los usuarios a descubrir contenido relevante.

La primera etapa de tu tarea implica la evaluación de la calidad de los datos. Deberás identificar valores atípicos, datos faltantes e inconsistencias en el conjunto de datos.
Tu objetivo es depurar los datos, documentando cada paso del proceso para asegurar la transparencia y reproducibilidad de tu análisis.

En la siguiente etapa, realizarás un análisis exploratorio. Utilizando estadísticas descriptivas y visualizaciones, explorarás la distribución de variables clave, identificando patrones y tendencias que puedan informar el preprocesamiento y la posterior modelación.

Continuarás el proceso creando nuevas características.
Deberás concebir al menos dos características derivadas que puedan enriquecer el análisis y mejorar la capacidad predictiva del modelo.
Explica el razonamiento detrás de estas nuevas variables y detalla los métodos utilizados para calcularlas.

Un aspecto crítico será la selección de características.
Utilizarás métodos específicos para identificar las variables más importantes para predecir la calificación de usuarios de anime.

Esta etapa es fundamental, ya que la elección de las variables adecuadas impactará directamente en la precisión y eficiencia de tus modelos predictivos.
Como parte de un proceso iterativo, explorarás cómo la selección de características afecta tu comprensión del problema y el rendimiento potencial del modelo.
La adaptabilidad y el refinamiento continuo son esenciales en este proceso.

Finalmente, compartirás tus hallazgos en forma estructurada. Crearás un informe que resuma tu trabajo de preprocesamiento de datos, incluyendo visualizaciones relevantes, estadísticas importantes y explicaciones claras de tus decisiones metodológicas.

Destacarás los descubrimientos clave de tu análisis exploratorio y cómo pueden informar la construcción de un modelo predictivo efectivo.

Recuerda que este proyecto es un proceso integral que combina análisis técnico con toma de decisiones estratégicas.
Demuestra tu habilidad como data scientist para transformar datos crudos en información valiosa que pueda guiar decisiones basadas en datos.

Para llevar a cabo todo esto, necesitarás el archivo de datos **imdb_anime.csv,** que contiene las siguientes columnas:

- **Title**: Nombre de la animación
- **Genre**: Género(s) bajo el cual cae la animación, por ejemplo, Acción, Aventura, etc.
- **User Rating**: IMDb calificación de usuarios sobre 10.
- **Number of Votes**: Total de usuarios de IMDb que han calificado la animación.
- **Runtime**: Duración de la animación en minutos.
- **Year**: Año en que se estrenó o comenzó a emitirse la animación.
- **Summary**: Un resumen breve o completo de la trama de la animación. Resúmenes completos se obtienen cuando están disponibles.
- **Stars**: Lista de actores principales o actores de voz involucrados en la animación.
- **Certificate**: Certificación de la animación, por ejemplo, PG, PG-13, etc.
- **Metascore**: Calificación de Metascore, si disponible, que es una puntuación agregada de varios críticos.
- **Gross**: Ganancias brutas o recaudación en taquilla de la animación.
- **Episode**: Indicador binario si la lista es para un episodio de una serie (1 para sí, 0 para no).
- **Episode Title**: Título del episodio si la lista es para un episodio; de lo contrario, será None (Ninguno).

Considerando estos datos:

1. Realiza un análisis de calidad de datos, revisando aspectos básicos y selecciona un primer conjunto de variables a eliminar. Luego de ello, realiza un análisis exploratorio inicial considerando gráficos de distribuciones de las diferentes variables, y concluye al respecto. Si observas algo raro respecto a los tipos de variables debes proponer algún tratamiento.

2. **Transformación Inicial de Datos**: Las diferentes columnas que son datos de texto deben ser transformadas a numéricas para poder explorarlas de mejor forma por ejemplo:
    a. **User Rating**: Extraer el número correspondiente al rating
    b. **Number of Votes**: Convertir en número
    c. **Year**: Extraer el año de inicio del anime
    d. Otros. Aplica algún criterio para saber qué variables deben ser transformadas en primera instancia.

3. **Revisión de outliers**: Ahora que tienes variables numéricas revisa la distribución y utiliza algún método para encontrar outliers, por ejemplo IQR o Z-score.

4. **Transformación de variables finales**: Realiza un pequeño análisis de distribuciones y transforma las variables aplicando transformaciones como logaritmo o get_dummies para extraer las diferentes categorías.
   Genera una estrategia para lidiar con los valores nulos y crea las variables que te parezcan necesarias.

5. **Análisis de Correlaciones**: Genera un análisis de correlaciones de las variables. No es necesario que apliques todos los métodos vistos en clases, basta que argumentes bien cuál utilizarás y por qué, y si necesitas algo más. La idea es generar gráficos para entender la relación entre las diferentes variables, poniendo foco en la variable objetivo **User Rating**.

6. Genera una función que resuma todo el procesamiento necesario para el dataset, que lea el dataset original y entregue un dataset ya tratado, con las columnas transformadas y creadas.

7. A partir de las columnas que obtuviste realiza una selección de variables según los siguientes métodos:
    a. Filtros basados en correlaciones
    b. Forward Selection.

   Compara ambos métodos y responde si coincide lo resultante con lo obtenido en el análisis exploratorio.

Requerimientos

1. Analiza datos y los prepara considerando datos nulos, faltantes o outliers, considerando el contexto dado y necesidades de transformación. **(3 puntos)**

2. Analiza correlaciones entre variables, justificando su selección desde el contexto e interpretando los indicadores obtenidos. **(3 puntos)**

3. Selecciona variables para un análisis, considerando diferentes métodos e interpretando sus resultados. **(4 puntos)**

¡Mucho éxito!

Consideraciones y recomendaciones

- Aprovecha las funciones que tiene la librería pandas para el tratamiento de datos.
- Sé ordenado al momento de trabajar y piensa en cómo iterar, es decir la misma función te sirve para revisar cómo queda un dataset antes y después de cierto tratamiento.
- Genera funciones para reutilizar código.
- Para recuperar la categorías de la columna 'Genre' puedes utilizar lo siguiente:
    - **data['Genre'].str.split(', ').str.join('|').str.get_dummies()**