Examen 2

Diplomado en Ciencia de Datos — Módulo 1

September 2019

Conjunto de datos

Considere el siguiente conjunto de datos.

	movie_title	movie_imdb_link	color	genre_4	duration	gross	genre_1	genre_2	genre_3	num_voted_users
0	Avatar	http://www.imdb.com/title/tt0499549/?ref_=fn_t...	Color	Sci-Fi	178.0	760505847.0	Action	Adventure	Fantasy	886204
1	Pirates of the Caribbean: At World's End	http://www.imdb.com/title/tt0449088/?ref_=fn_t...	Color	NaN	169.0	309404152.0	Action	Adventure	Fantasy	471220
2	Spectre	http://www.imdb.com/title/tt2379713/?ref_=fn_t...	Color	NaN	148.0	NaN	Action	Adventure	Thriller	275868
3	The Dark Knight Rises	http://www.imdb.com/title/tt1345836/?ref_=fn_t...	Color	NaN	164.0	448130642.0	Action	Thriller	NaN	1144337
4	Star Wars: Episode VII - The Force Awakens ...	http://www.imdb.com/title/tt5289954/?ref_=fn_t...	NaN	NaN	NaN	NaN	Documentary	NaN	NaN	8

La tabla proporcionada registra información del sitio IMDb. Eb los registros se proporciona información sobre peliculas.

Diccionario de datos

- **movie_title:** Título de la película
- **movie_imdb_link:** Hipervínculo hacia la película en IMDb
- **color:** Colorización de la película
- **duration:** Duración en minutos
- **gross:** Ganancias brutas de la película en dólares
- **genre_1:** Primer género de la película
- **genre_2:** Segundo género de la película, si aplica
- **genre_3:** Tercer género de la película, si aplica
- **genre_4:** Cuarto género de la película, si aplica
- **num_voted_users:** Número de personas que votaron por la película
- **facenumber_in_poster:** Número de rostros en el poster de la película
- **language:** Lenguaje original de filmación de la película
- **country:** País de producción de la película

- **content_rating:** Calificación del contenidoo de la película
- **title_year:** Año de lanzamiento de la película
- **imdb_score:** Score de la pelicula en IMDb

Análisis exploratorio de datos

- (1 puntos) Indique cuales de las variables presentadas son discretas y cuales continuas.
- (1.5 punto) Realice una exploración visual ligera de los datos. Independientemente de la herramienta (Tableau, Pygal), agregue imágenes que respalden la exploración.
- (1.5 puntos) Remueva outliers de aquellas variables que los presenten. Utilice cualquier método para ello. Muestre un cuadro con el número de registros antes y después de los tratamientos.
- (1.5 punto) Normalice las variables discretas que lo requieran. Muestre un cuadro con las categorías resultantes por cada variable.
- (1 punto) Realice una nueva exploración visual, esta vez con el objetivo de ver los efectos de los tratamientos realizados. De igual modo, agregue las imágenes correspondientes.

Ingeniería de datos

- (1 punto) Cree una variable binaria que indique sí el score de la película es mayor al promedio. Analice gráficamente la frecuencia de los valores generados.

Tratamiento de valores ausentes

- (1 punto) Elimine aquellas columnas que superen el umbral de 70% o más de presencia de valores ausentes. Indique que columnas fueron eliminadas.
- (1 punto) Impute las variables discretas que lo requieran mediante el uso de la moda. La variable imputada debe almacenarse en una variable nueva, dejando a la original intacta.
- (1 punto) Impute las variables continuas que lo requieran mediante el uso de la mediana. La variable imputada debe almacenarse en una variable nueva, dejando a la original intacta.
- (1 punto) Impute las variables continuas que lo requieran mediante el uso de la media. La variable imputada debe almacenarse en una variable nueva, dejando a la original intacta.

Reducción de dimensiones

- (1 puntos) A partir de las variables continuas, reduzca el dimensiones y visualice los datos en un gráfico de 2D.
- (1 punto) Mediante el uso de pruebas de poder predictivo, ordene las variables continuas contra el score de la película.
- (1.5 puntos) Mediante el uso de transformación entrópica, seleccione las 5 variables más potentes contra la variable binaria creada previamente. Considere las variables con el siguiente tratamiento: sin valores atípicos, normalizadas y con missings. Muestre un cuadro con el poder predictivo de las variables.

Feedback

Por favor, aporte comentarios sobre el curso, el ponente y las clases. El objetivo es poder mejorar los contenidos y el desarrollo del módulo. Este punto es obligatorio para la calificación del examen.

Entrega

La entrega debe cumplir con los siguientes requisitos:

- Debe enviarse vía e-mail a: oscar.acosta.mac@gmail.com
- La fecha límite para hacerlo es el dá jueves 3 de octubre a las 5 PM, hora Ciudad de México.
- Enviar un archivo comprimido (ejemplo: nombre_apellido.zip) con las evidencias del examen: gráficas, códigos, imágenes, feedback, etc.
- El envío no será válido si no contiene el feedback.
- Indicar claramente la respuesta asociada a cada pregunta.
- Incluir en el cuerpo del correo el nombre completo del alumno.