Taller 1 Mineria de Datos

David Morneo

February 08, 2024

Punto 1

Pregunta: Identifique qué procesamiento de datos fue tenido en cuenta. Relacione las funciones de R implementadas y explique su funcionamiento de forma concreta.

Respuesta: Primero observamos que los datos leidos son del tipo .dat. Lo que hace el código en la función **gsub()** es sustituir el string "::" por un espacio. La función **str_split_fixed()** realiza un split de cada elemento del vector anterior, indicando al final con un 3 para que cree 3 columnas. Luego, se asignan nombres a las columnas utilizando **colnames(dataframe)<-c()** para definir el nombre de las columnas.

Posteriormente, realiza un merge de los datos de ratings y movies basándose en el *movieId*, que ambas bases de datos comparten, y une los datos mediante un Left Join, conservando los ratings como la tabla de la izquierda. Depués, divide los datos en conjuntos de Test y Validación utilizando la función **createDataPartition()**, la cual recibe la variable respuesta, y el número de veces que se quiere dividir(este número puede variar si se utiliza validación curzada) y p, el porcentaje de los datos que se utilizarán para validación.

```
validation <- temp %>%
     semi_join(edx, by = "movieId") %>%
     semi_join(edx, by = "userId")
```

La función %>% representa un pipeline, que permite llamar a otras funciones secuencialmente. Esto es útil para realizar varias transformaciones en los datos en un solo paso.

Punto 2

Pregunta: Explique cuáles análisis exploratorios hicieron por cada variable, ¿qué beneficio le trajo al modelo final? ¿Cuáles son algunas funciones clave en ese sentido?

Respuesta:

En el análisis exploratorio, se calculo el número de usuarios, de peliculas, de generos y se análizo el intervalo de fechas de nuestros datos, se hizo esto para los datos de validación y de entrenamiento.

Análisis Exploratorio Rating: Para la variable Rating, se realiza un resumen descriptivo, incluyendo la media y la desviación estándar.

Luego, se crea una gráfica de barras contando el número de ratings por cada 0.5.

Posteriormente, se elabora la gráfica ordenada de mayor a menor. Una de las funciones clave en estos gráficos es **ggplot(data=DataFrame, aes(x, y)) + Geom(additional parameters)**. Esta función recibe el conjunto de datos; la función aes() define el mapeo entre los datos (se elige qué se va a graficar); y el tercer parámetro se añade para especificar qué tipo de gráfico se desea crear, ya sea de línea o de barras, además de los demás parámetros que agregan el título, nombres y demás.

Otra función que me parecio interesante fue **mutate(group=cut(n__rating__of__movie,breaks=c(-Inf,mean(n__rating__of__movie),Inf)** Lo que hace **mutate()** lo que nos permite es agregar variables a un dataframe of modificar algunas que ya sean existentes y la variable **cut()** se utiliza para dividir el rango de una variable continua en intervalos. luego cuando utilizamos **breaks=c(-Inf,mean(),Inf), le definimos los intervalos. Como sabemos en R un número Inf es el equivalente a 2^31.

Análisis Exploratio Movie: Para el análisis de las peliculas una de las funciones importantes es **group__by(movieId)** la cual como su nombre lo indica agrupa filas en este caso por la media y la desviación estandar, puero puede utilizarce para suma, media, y otras funciones más para organizar mejor los datos y hacer gráficos más especificos.

```
movie_sum <- edx %>% group_by(movieId) %>%
  summarize(n_rating_of_movie = n(),
            mu_movie = mean(rating),
            sd_movie = sd(rating))
```

Una función crucial en el análisis de datos con R es **summarize()**, que se emplea para generar resúmenes de los datos. En el contexto mencionado, esta función se utiliza para calcular estadísticas clave como la media y la desviación estándar, organizando los datos por *MovieId*.

Un análisis visual particularmente revelador es la exploración de la densidad del número de calificaciones. Este enfoque permite identificar cómo se distribuyen predominantemente las calificaciones en términos porcentuales.

La visualización gráfica es instrumental en este proceso, ya que facilita la identificación de valores atípicos (outliers) que agregan significado y profundidad a nuestra comprensión de los datos.

Dentro de las herramientas gráficas que encontramos útiles están **geom__hist()**, **geom__point()**, y **geom__vline()**. Este último, por ejemplo, se utiliza para añadir una línea vertical en el valor medio de las calificaciones por película, mediante el uso de **aes(xintercept = mean(n__rating__of__movie))**.

Punto 3

Pregunta: Identifique la estructura de visualización utilizada. ¿Cuáles funciones fueron implementadas y cuál es su funcionalidad de forma concreta?

Respuesta: En el documento, detallo las funciones clave para la creación de gráficos.

Punto 4

Pregunta: ¿Qué le llamó la atención de esta aproximación? ¿Qué hubiera hecho diferente?

Respuesta: En la fase de análisis exploratorio, la estrategia de segmentar las calificaciones en dos grupos, basándose en el promedio, fue fundamental para el reconocimiento de valores atípicos y su posible influencia en el modelo. Esta división revela variaciones en las preferencias o sesgos de los usuarios.

Me resultó particularmente revelador cómo se mejoró la precisión del modelo mediante la incorporación de un efecto individual por usuario (bjbj), ajustando las predicciones conforme a si un usuario tiende a calificar por encima o por debajo del promedio, utilizando la fórmula $b_j = \bar{r}_j - \bar{r}$

Una estrategia alternativa que contemplaría implica analizar las correlaciones entre variables dentro de categorías determinadas para destacar películas altamente calificadas. Luego procedería a realizar un análisis de agrupación (clustering) con dichas variables para identificar posibles categorías distintivas. Si se descubren categorías relevantes, las incorporaría al modelo con el objetivo de mejorar la precisión de las predicciones. Adicionalmente, examinaría los valores atípicos y consideraría tácticas para atenuar su influencia, como la aplicación de técnicas de palanca en la regresión lineal.