Untitled

Geovanny Risco y Robert Novak

3/1/2022

Descripción del Dataset

El dataset que hemos escogido está recogido mediante webscraping en distintas plataformas sobre ofertas de empleo relacionadas con los datos en Estados Unidos. Hemos escogido este dataset principalmente por dos razones.

1. Es un dataset bastante desordenado en el que da lugar a hacer procesos de limpieza de distinto tipo ideal para asentar los conceptos tratados en la asignatura
2. Nos parece interesante conocer el mercado laboral de las distintas profesiones a las que podríamos aspirar tras la finalización del máster y la demanda, aunque sea en un país extranjero.

Nuestro principal objetivo con el dataset es contestar a distintas preguntas relacionadas con el salario y distintas variables que se proporcionan en el dataset:

Aquí concretar un poco algunas de ellas

Integración y selección de los datos de interés a analizar

Limpieza de datos

En primer lugar, hemos realizado un análisis del dominio de las variables a partir de las cuáles hemos hecho hecho las siguientes observaciones:

1. Se utiliza el `-1` para indicar valores faltantes. Adicionalmente, existen columnas que tienen un valor faltante que se representa de forma distinta a `-1` por la forma en la que se han extraído los datos. En la limpieza hemos tratado todos esos casos y representado los valores faltantes de forma homogénea mediante `NA`, que es la forma de representar los valores faltantes en R y gracias al cuál podemos hacer operaciones para algunas de las funciones donde se tienen en cuenta los valores faltantes.
2. La columna `Job title` tiene una gran diversidad de trabajos con una mínima variación en la que es interesante tratarlos como un mismo trabajo. Para ello, hemos definido un subconjunto de trabajos a partir del cuál tratar como iguales las variantes. Ese subconjunto son los que consideramos principales : { data scientist, data engineer, data analyst, machine learning}. Así, por ejemplo, un trabajo de e-commerce data analyst o uno de RFP data analyst será tratado bajo la categoría de data analyst. Aquellos trabajos muy específicos en los cuáles no se engloba bajo ninguna de las categorías anteriores los consideramos muy específicos y, al no ser un número muy elevado hemos decidido no tratarlos.
3. La variable `Company name` tiene la información del rating. Hemos eliminado esa redundancia
4. Hemos añadido una nueva variable binaria a partir de `Location` y `Headquarters` para ver aquellas ofertas de trabajo en la que la cede central de la empresa está en el mismo sitio que la oferta
5. Algunas variables como `Salary Estimate`, `Size` y `Revenue` contienen información que pueden ser aprovechadas mejor separándolas en más columnas a partir de las cuáles sacar más información. Así, las hemos separado en más columnas. Una para los rangos mínimos, otro para los rangos máximos.

6. `Salary Estimate` puede ser considerada una variable cuantitativa ya que, aunque se proporcione un rango variable para todas las ofertas, la realidad es que el salario no es un rango sino un valor concreto dado por un dominio continuo. La decisión que hemos tomado para solucionar esto es considerar el punto medio del rango proporcionado como el salario de la oferta. Esta solución es una aproximación ya que dos ofertas con mismos rangos tendrían el mismo salario y no tendría por qué ser considerados como el mismo. O, incluso, dos salarios con rangos distintos pero con una cierta intersección podrían tener en la realidad el mismo salario pero no tal como lo hemos tratado. Sin embargo, aunque lo ideal sería hacer un estudio externo sobre la distribución del salario dado el rango, la empresa particular, etc. Al no disponer de esa información asumimos esta simplificación.
7. `Size` y `Revenue` deben ser consideradas para análisis posteriores como variables ordinales ya que su dominio corresponde a categorías no solapadas en el que el orden importa.

Análisis de datos

Análisis de la normalidad de la variable salary

Representación de los resultados a partir de tablas y gráficas

Código