EDA con Python

En esta entrega se resuelve el enunciado propuesto. Se han cargado las fuente de datos propuestas y limpiado para conseguir una base de datos con la que trabajar de forma eficiente y clara. Para explicar los pasos y razonamientos seguidos se han hecho anotaciones recurrentemente antes y después del código en cuestión. En algunos casos, para poder seguir también el análisis efectuado se han añadido algunas observaciones. Finalmente también se incluyen representaciones gráficas para que leer y entender los cruces y cálculos sean más sencillos.

Composición del archivo

- MT_Entrega_Análisis
- M7_Entrega_Preparación
- M7_Readme
- Mi_dataframe
- Tabla_final_definitiva
- Tabla_final_limpia.pkl
- Anexo gráficos y tablas

Instalación y requisitos

Para poder ejecutar y leer correctamente el proyecto hay que tener instalado “Python”, además de algunas extensiones como “Jupyter Notebook”, herramientas como “Pandas”, “NumPy”, “ e importar librerías como “Math”, “functools”, “Matplotlib” y “Seaborn”.

Al inicio de los archivos Python sea especifica las librerías o herramientas necesarios para el correcto funcionamiento, ya que si no se contara con algunas de las señaladas, las posibilidades de interactuar con el código se reducirían bastante. Adicionalmente, es probable que antes de algún paso se nombre la herramienta usada.

Resultados y conclusiones

INTRODUCCIÓN

Antes de empezar a analizar los datos obtenidos, quiero plantear la misma “primera fase” con la que empecé la tarea. Mi primera misión fue intentar entender las tablas y los datos que contenían, atendiendo especialmente a los datos socio demográficos, que suelen arrojar conclusiones interesantes cuando las empresas intentan entender la población o su demanda potencial. Es por eso por lo que mis cruces de datos y valoraciones se centran en combinar de las formas que he creído mas interesantes los datos generales de la fuente original, con el número de hijos, el nivel de estudios o la edad. Después de eso, me centré en conocer las dummies con las que contaba, viendo que serían útiles para enriquecer este análisis que inicialmente había planteado.

He de reconocer que me perturba no haber sido capaz de convertir las coordenadas en un dato geográfico aplicable en el análisis, ya que considero que podría haber creado mapas de calor y demás gráficos que contribuyeran a hacer un análisis mucho más profundo.

Después del primer vistazo inicial, preparé el DataSet para que el trabajo con el fuera más amigable, normalizando las variables binarias, adecentando el nombre de las columnas, colocando el id como índice o creando diccionarios para que el dato en sí se adaptara a

cómo en España tenemos esos conceptos como nivel educativo o trabajo, intentando arrastrarlos a nuestro vocabulario de la forma más fiel posible.

PRIMERAS OBSERVACIONES

- Análisis preliminar

En primera instancia, vemos como están distribuidos estadísticamente algunas de las variables más descriptivas del conjunto de datos. La edad media ronda los cuarenta años aunque la edad más frecuente de los individuos de nuestro universo es 31 años. Vemos que los ingresos medios superan por poco los 90.000 u.m. al año, secundado por la mediana, pero podemos inferir que hay una gran cantidad de individuos con una renta sensiblemente superior porque la moda de la renta anual es notablemente menor (70.000 u.m./año). La cantidad de hijos por individuo es consecuente con la edad media por observación, y su distribución es simétrica entre hijos pequeños y adolescentes, aunque el número de hijos pequeños más frecuente (2), también es coherente con la edad media de los individuos de nuestros datos (adulto joven).

Continuamos con el análisis preliminar, y vemos que los datos parecen tener sentido también en cuanto a nivel educativo se refiere. A medida que aumenta el nivel educativo aumenta, la cantidad de individuos que se ven recogidos por ese rango disminuye, siendo la educación primaria el que cuenta con mayor número de individuos. (He estado investigando y gado universitario puede contener observaciones de otros tipos de ciclos formativos por la pérdida de contexto por la traducción, es por eso que cuenta con más individuos que el instituto). Vemos también que la ampla mayoría de la muestra se casó y sigue en régimen de matrimonio en el momento de extracción de datos.

- Edad y duración de las llamadas

Con el diagrama de dispersión “Edad vs Duración de la última llamada” podemos ver como las llamadas se concentran en el rango de edad 20-60, y que es además el rango de edad en el que es más probable conseguir que los teleoperadores tengan llamadas más largas con el potencial cliente.

- Correlación entre variables

En la correlación entre variables reconozco que no me interesan demasiado, porque con este primer vistazo se puede ver claramente con la correlación significativa entre variables queda reservada a aquellas en las que el sentido común ya establece esa relación causa-efecto. Como por ejemplo, existiendo correlación significativa entre todas las variables que contienen información sobre los infantes de los individuos. No obstante, esa “última llamada” sí que parece haber surtido buen efecto en la respuesta a la campaña por parte de los impactados.

Antes de seguir, vemos si el impacto de la campaña ha sido cuantioso o no. Cuando se lanzan campañas masivas, como por ejemplo de telefonía móvil o de mejora de planes de energía domésticos, la población suele ser inmensa, pero los individuos verdaderamente impactados, representan un bajo porcentaje del total, aunque en valores absolutos, suelen sumar un número bastante elevado. Probablemente sea lo que ocurre en esta campaña, ya que el impacto de la campaña representa en torno al 6%, pero esto representa algo menos de 5.000 usuarios, que son muchos usuarios con los que has hablado de tu producto. En nuestro caso, vemos que los individuos no impactados tienen

un rango de edad más concentrado con más outliers, mientras que los que si son considerados por la firma como impactados, se extienden desde los 30 hasta casi los 50 años, con menos valores externos al rango y también en un rango de edad menor.

COMPARACIONES DE SUBGRUPOS: TOP 1000

- Usuarios que más visitan la web

Si comparamos los estadísticos generales con los clientes más fieles (top 1000 de usuarios que mas frecuentan la web de la empresa), vemos que todo lo analizado está en un rango bastante estrecho salvo por el tiempo que los usuarios llevan siendo clientes de la entidad, habiendo un gap de casi tres años. Esto es más fácilmente visible en la tabla creada justo después llamada " Top1000 fieles". Adicionalmente se crea un diagrama de barras para que sea más fácilmente que solo el tiempo que llevan siendo clientes los individuos de ese grupo es el aspecto más diferenciador. (De este análisis con diagrama de barras y de otros similares se excluye inicialmente los ingresos por la diferencia en escala, aunque al final de este bloque se crea un gráfico solo para ver las diferencias en los ingresos medios de los individuos).

- Usuarios en función del nivel educativo

Con el nivel educativo, vemos algo curioso es que se muestra claramente como haber alcanzado cotas de estudio mas altas académicamente hablando, hace más probable que seas cliente durante más tiempo de esta entidad. Esto puede significar que el target de este banco sean personas con cierto nivel académico, aunque no estoy presuponiendo un grado universitario como el tope de los estamentos de estudios, en este conjunto de datos si lo es.

- Usuarios en función de sus ingresos anuales

Si ahora aislamos a los 1000 individuos que más ingresos al año perciben, vemos como en algunas de las variables explicativas del contexto socio demográfico no representan una gran diferencia con respecto al total del universo, pero es una prueba clara de como los individuos que más renta perciben, ganan cantidades muy notablemente superiores a los valores medios observados previamente. Obviamente sus registros generan una brecha mayor con respecto a los valores moda.

- Comparativa de ingresos percibidos

Posteriormente, analizamos las diferencias en ingresos entre los diferentes subgrupos que habíamos ido generando. Vemos como el universo total percibe menos ingresos medios que los individuos que alcanzan estudios universitarios, los individuos que llevan más tiempo siendo clientes del servicio/ entidad y que más visitan la web, además de obviamente el subgrupo de "los más ricos económicamente". Esto tiene una representación gráfica muy explicativa en el diagrama de barras "Comparación de diferencias en INGRESOS entre grupos vs Todos"

COMPARACIONES ENTRE SUBGRUPOS II

- **DataFrame Financiación**

Ahora vamos a ver como afectan a los individuos las variables: "tiene_hipoteca", "tiene_prestamo", "incumplimiento_pagos", "edad".

Inicialmente, observamos que la edad media de los individuos que cuentan con una hipoteca y que tienen al menos un préstamo en vigor es casi la misma (39 y 38,8 respectivamente) y la edad de los individuos que al incumplido alguno de los pagos de alguno de sus productos es algo mayor, de casi 50 años.

Si queremos ver cómo se distribuyen en función de la edad estas variables, vemos como el mayor número de individuos con hipoteca se concentra entre los 25 y los 44 años, donde se embarcan en la compra de una vivienda y aún, en líneas generales, n han tenido tiempo de completar el pago de esta. A partir de esa edad, el número de hipotecados va disminuyendo sostenidamente, posiblemente porque hayan podido completar la hipoteca en años anteriores. Las edades en las que los individuos tienen préstamos también se concentran en el mismo rango, porque es idealmente cuando los individuos y las familias han de hacer desembolsos más grandes, como el inicio de vida y estudios de un niño, y vacaciones más caras, por ejemplo, ya que en los últimos años vemos como cada vez más familias han de solicitar préstamos para costearse las vacaciones (En general, de época estival).

Si usamos probabilidad condicionada, vemos como el hecho de haber incumplido en obligaciones de pago, elimina a esos individuos la posibilidad (por los datos observados) la posibilidad de obtener un préstamo. Algo bastante consecuente con la realidad, ya que es normal que los bancos muestren reticencia a prestar capital a usuarios que hayan incumplido con pagos previamente. Vemos que la probabilidad de que un individuo que tiene hipoteca solicite un préstamo es menor al 20%, pero la probabilidad de que un individuo que tenga un préstamo pida una hipoteca, es del 50%, es curioso, tiene sentido y además sirve como ejemplo de la falacia “Aristóteles es humano pero no todos los humanos son Aristóteles”.

- **DataFrame Familia**

Ahora, de manera simétrica a lo visto en el apartado anterior, vamos a ver como se relacionan entre sí las variables: "edad", "total_hijos", "hijos_pequeños", "hijos_adolescentes", "nivel_educativo"

El nivel educativo no parece ejercer un efecto claro en la decisión de las familias de tener hijos., pero, como sí era de esperar, vemos que el rango de edad más normalmente distribuido se encuentra en la educación primaria.

En el archivo se pueden ver tablas con grupos de edad más y menos granulados para observar los datos de distintas formas.

- **DataFrame Conjunto: Familia + Financiación**

Si combinamos los dos últimos DataFrames y usamos probabilidad condicionada, podemos justificar muchas de las asunciones que probablemente podamos hacer si se nos plantaran las variables del conjunto de datos.

La probabilidad de que un individuo que tiene hijos tenga una hipoteca es de casi el 50% y además vemos como el hecho de tener hijos, hace que las probabilidades de que ese individuo tenga hipoteca tengan un préstamo o haya sido estudiante universitario crezcan significativamente. La visualización gráfica de esta parte se puede apreciar en el diagrama de barras horizontal “Probabilidades condicionales vs proporciones generales” y en “Heatmap de Probabilidades Condicionales” además de más gráficos.

CONCLUSIONES

A lo largo de este proyecto se ha hecho una simulación de lo que a una empresa le interesaría conocer sobre su publico objetico, para poder plantear futuras campañas publicitarias, productos o comunicaciones e imagen de marca. Bien es cierto, que por mi corta experiencia en trabajos o proyectos similares, se suele acotar por parte del cliente lo que necesita o busca, permitiéndote afinar la búsqueda y tratar los datos de una forma más específica para orientar el desarrollo y presentación de los datos a lo que la entidad realmente necesita.

No obstante, he disfrutado analizando, preparando y tratando los datos con el fin de conseguir explicar a los individuos que componen el universo muestral aportado.

🖌 **Autor**

- David Fernández
- Davidflopez21