Memoria

En este documento, proporcionaremos explicaciones detalladas con el objetivo de ofrecer una comprensión más completa de los análisis llevados a cabo y de los códigos aplicados.

Contexto, Telecom brinda una BBDD de las interacciones con sus clientes la cual analizaremos la información es publica y esta disponible en la siguiente web:

<https://www.kaggle.com/datasets/datazng/telecom-company-churn-rate-call-center-data>

1. Análisis Exploratorio:

* **Importamos librerías** matplotlib, pandas, numpy y seaborn si bien en el documento final no hay grafica se importan para simplificar su uso en el futuro.
* **Lectura de documento** copiamos el path y utilizamos r’’ dado que mi sistema operativo es Windows y con esto no aplica las barras invertidas, también la lectura en una variable. Luego lo transformamos en en dataframe mediante mandas. Esto lo realizamos en 2 partes a modo de practica dado que en el futuro aplicaremos una sola línea de código.
* **Entendimiento,** Visualizamos:

Las de columnas y observaciones lo realizamos mediante Shape()

Tipo columna, de datos y nulos lo realizamos mediante info()

Realizamos un describe para ver el resumen estadístico de los datos.

Cantidad de agentes, lo realizamos con unique()

Cantidad de Motivos de llamados lo realizamos con unique()

Horario de atención lo realizamos mediante sort value()

1. Limpieza & Set Up de la base de datos:

* **Lectura de documento,** Volvemos a leer el documento y lo realizamos en donde incluyendo "parse\_dates" para que la columna 'Date' se lea en el formato más adecuado y hacemos algo similar con otras 2 columnas del tipo Float que serían Speed of answer in seconds' y 'Satisfaction rating'
* **Cambiar los títulos de las columnas a mayúsculas y agregar la leyenda del índice, p**ara trabajar de una manera más ordenada se decide insertar un índice al cual se puede llamar en el futuro
* **Conversión de columnas, Time & AvgTalkDuration,** la columna 'Time' se modificara como tipo de dato a delta time

La columna 'AvgTalkDuration' se modificará como tipo de dato datetime con el formato días, horas, minutos y segundos.

Unicamente se completan los Nan de la columna con el fin de convetir el tipo de dato proximamente se analizaran por partes el Dt con el fin de analizarlo en 2 etapas del "embudo de análisis

* **Creación de columna "DAY\_OF\_THE\_WEEK**, se decide crear esta columna para visualizar de una manera más simple los días en que la empresa brinda servicio.

Una vez creada realizamos una pivot table con el índice de los días de la semana para visualizar que días se brinda servicio.

* **Limpieza & generación de embudo para análisis de datos**

En primer lugar, verificamos que no figuren datos existentes en las llamadas concretadas, Una vez corroborado, procederemos separar los dt con el fin de analizaros por 2 partes distintas del embudo:

interacciones concretadas = bbdd\_Y

interacciones no concretadas = bbdd\_N

Realizamos la verificación de nulos sobre la base de interacciones concretadas.

Realizamos la verificación de duplicados sobre toda la base y sobre las columnas clave por ejemplo CALL ID.

Realizamos la verificación de absurdos, recurrimos a una función que nos brinde todos los datos únicos de cada columna.

En una segunda verificación realizamos una búsqueda de llamados fuera del horario de atención.

* **Se reordenan las columnas**, con el fin de agilizar la visualización
* **Se procede a guardar las bases de datos depuradas** en nuevos CSV con el fin de poder llamarla en nuestros futuros análisis y no tener que correr todas las customizaciones

1. Estadística descriptiva & as-is de los datos:

* **Importamos librerías** matplotlib, pandas, numpy, scipy.stats y seaborn
* **Lectura de mi csv** y se guarda en 3 variables, lectura de la base en bruto, lectura de las interacciones concretadas y lectura de las interacciones no concretadas, si bien trabajaremos principalmente con una, se realiza la lectura de las 3 con el fin de que en el caso de precisarlas ya estén cargadas y seteadas con igual formato.
* **Conversión de columnas**, Time & AvgTalkDuration, se realiza nuevamente dado que al guardar la base en un csv se vuelve a cambiar al tipo de formato más simple.
* **Detalle & Entendimiento de las columnas,** disponibilizamos una tabla con el detalle de las columnas y a que hace referencia cada una.
* **Análisis Descriptivo & Insights preliminares**, calculamos la moda, la media y la mediana de los datos disponibles en Telecom
* **Realizamos el análisis, de** cuantas llamadas fueron atendidas y cuantas fueron atendidas con un tiempo de espera menor a un minuto y cuantas fueron resueltas, establecemos variables tipo métricas para cada uno de estos análisis.
* **Cardinalidad de los datos,** realizamos el análisis de cardinalidad con el fin de ver cuántos únicos hay en cada una de las columnas cuales tienen cardinalidad alta, media y baja y proceder con el análisis univariante según el resultado, esto lo obtenemos mediante una función.
* **Correlación de los datos,** utilizamos un gráfico de correlación para analizar si las variables numéricas se correlacionan entre si, según la información no una relación clara entre las variables en este primer análisis.
* **Análisis Univariante & distribución de las llamadas**, en base a nuestro análisis de cardinalidad realizamos un análisis univariante de la distribución de llamadas, por agente, por tiempo de respuesta, por fecha, por día de la semana, llamados atendidos diariamente, semanalmente, llamados no atendidos diariamente y semanalmente, cantidad de llamados por franjas horarias de una hora y procedemos a graficarlos.
* Análisis Bivariante, En primer lugar aplicamos un pairplot para visualizar una matriz de gráficos de dispersión para visualizar las relaciones bivariadas entre diferentes variables numericas
* Análisis multivariante: analizamos 3 variables al mismo tiempo Agente, consultas resueltas o no y nivel de satisfacción con el fin de identificar si las mismas se correlacionan o no, en este análisis no logramos visualizar una correlación directa y procedemos a graficarlo.

1. Hipótesis

* Realizamos el planteo de 5 hipótesis
* 1) El principal motivo de contactos representa al menos un 25% del volumen total y además este representa más del 35% del tiempo operativo insumido. Definimos las variables mayor motivo de contacto y cantidad de contacto los porcentajes que representaban y el tiempo operativo estimado, para esta hipótesis realizamos una prueba Z de proporciones.
* 2) La distribución de llamados siguen una distribución normal para esta hipótesis realizamos un test de shapiro para analizar la normalidad de la distribución
* 3) La velocidad de respuesta un 15 % más lento que el promedio y la satisfacción del usuario afectan la probabilidad de resolución, para este test utilizamos un análisis de regresión lineal para examinar la relación entre los niveles de satisfacción y dos variables independientes: la resolución del caso (binaria: resuelto o no resuelto) y el tiempo de espera.
* 4) El 70% de los casos relacionados a los pagos se dan del 25 al 05 de cada mes y son resueltos, nuevamente utilizamos una prueba Z de proporciones para realizar el test de hipótesis, filtramos la variable que buscamos por fecha y luego aplicamos la prueba de proporciones
* 5) De reducir un 10% de los contactos del principal motivo se generaría una eficiencia en tiempos operativos de más de un 4%, en primer lugar definimos las variables con reducción, sin reducción y lo mismo con los tiempos operativos, luego aplicamos el test t de Student independiente (asumiendo varianzas iguales, se realiza de tal manera dado que es el mismo grupo a analizar y ya se encuentra filtrado por el motivo más frecuente de llamado)

1. Visualizaciones.

Si bien se realizaron visualizaciones de los distintos análisis e hipótesis se visualizan todas juntas la ultima parte del EDA Visualizaciones