Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Querétaro

TC3006C. Inteligencia artificial avanzada para la ciencia de datos I
Grupo 101

Momento de Retroalimentación :

"Reto Limpieza del Conjunto de Datos"

Evidencia presentada por los estudiantes :

Diego Alfonso Ramírez Montes	A01707596
Javier Suarez Duran	A01707380
José Ángel García López	A01275108
Emiliano Mendoza Nieto	A01706083

Profesores :

Benjamín Valdés Aguirre
José Antonio Cantoral Ceballos
Carlos Alberto Dorantes Dosamantes
Denisse L. Maldonado Flores
Alejandro Fernández Vilchis

Fecha de entrega :

Domingo 27 de Agosto de 2023

Resumen — En este reporte se explicara y documentaran las decisiones tomadas en equipo para limpiar los atributos y valores. Se trabajó sobre un dataset de Kaggle, "Store Sales - Time Series Forecasting". Nos apoyamos en un Repositorio de GitHub, y en Notebooks de Cocalc. Estos recursos se anexan al final del reporte para su acceso.
Palabras clave — *limpia, atributos, dataset, Kaggle, GitHub, Cocalc.*

I. Introducción

En el corazón de la gestión de datos radica un proceso fundamental: ETL, que significa Extracción, Transformación y Carga. Aunque no sea un software en sí mismo, ETL es un proceso crítico que da forma a los datos en información valiosa. Enfocándonos en la etapa de Transformación y Limpieza de Datos, exploraremos cómo este proceso esencial garantiza que los datos sean confiables y útiles para la toma de decisiones. Desde la extracción de datos hasta su refinamiento, cada paso en esta travesía contribuye a transformar datos en conocimientos significativos para las organizaciones.

II. Descripción del Dataset

La información a analizar pertenece a las tiendas "Favorita" ubicadas en Ecuador. Los datos entre los archivos incluyen fechas, información de la tienda y de los productos, si los artículos se estaban promocionando, los números de ventas, entre otras cosas que los archivos adicionales incluyen complementando la información. Todos los archivos proveídos se encuentran en formato ".csv" .

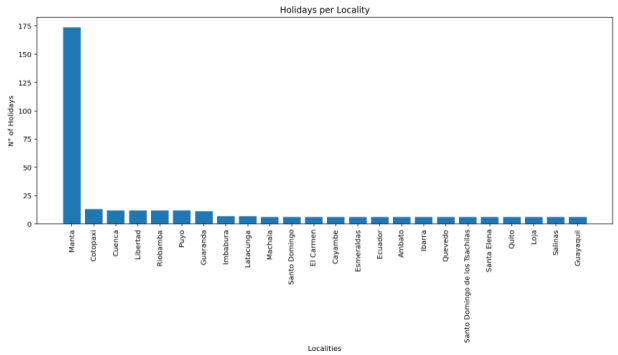
Descripción de los archivos:

- ➔ Sample_submission.csv : Un archivo muestra el formato final que deben tener las predicciones.
- ➔ Holidays_events.csv : Muestra todos los días festivos celebrados a través del tiempo, en este archivo hay información importante como el tipo de celebración realizada, su alcance y si la fecha fue celebrada el día mismo que se indica o si fue transferida a una distinta.
- ➔ Oil.csv : Aquí se muestra el precio que tuvo el barril de petróleo crudo ese día; este archivo es relevante ya que Ecuador es un país dependiente del petróleo por lo que el precio de este tiene un alto impacto en su economía
- ➔ Stores.csv : La información del archivo muestra donde se ubica cada tienda, el número de tiendas similares y el tipo de tienda
- ➔ Transactions.csv : El archivo es un registro del número de ventas totales que tuvo la tienda en el día, es decir, el número de cuentas o clientes que hubo
- ➔ Train.csv : Dentro de este archivo se encuentran las ventas por familia de producto que tuvo cada una de las tiendas por día, así como el número de productos en promoción por familia
- ➔ Test.csv : Al igual que el archivo de "Train.csv" se muestran las tiendas divididas entre las familias de los productos y la cantidad de productos en promoción, omitiendo la información de

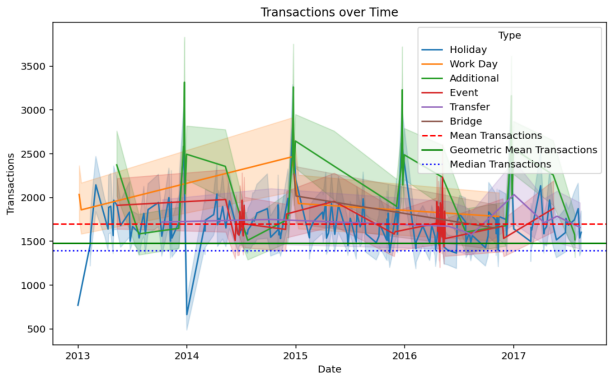
las ventas generados ya que es lo que se espera predecir

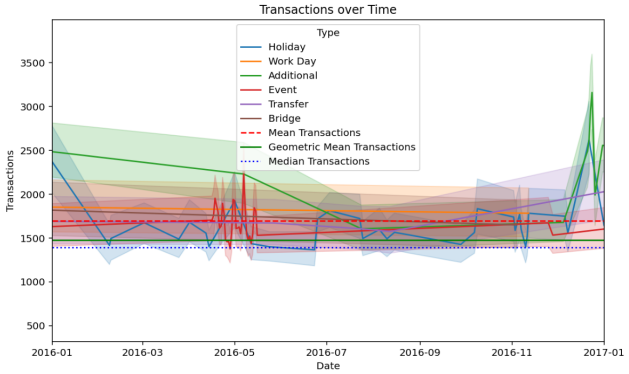
III. EXPLORACIÓN DE LOS DATOS

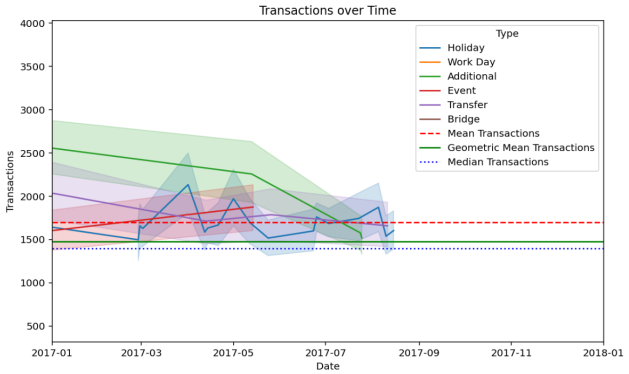
Al analizar los datos nos dimos cuenta que el tipo de fecha podía influir en las las ventas, por lo que decidimos que podría ser interesante saber si existía una región con mayor cantidad de festividades en el caso de que las mismas verdaderamente influyan.



En este caso podemos observar que hay una localidad que sobresale por mucho por encima de las demás, por lo que podría ser de interés nuestro ver cómo afectan los distintos "tipos" de días que existen en el dataset a las ventas.







Como podemos observar en las gráficas, hay dos indicadores estadísticos que nos dicen que las ventas por festividades tienden a estar por encima del promedio (geométrico) y la media, mientras que solo uno nos dice que las ventas están por debajo del promedio (aritmético). En conclusión, nos inclinamos a creer que las ventas se ven afectadas positivamente por las festividades.

IV. EXTRACT

Para la extracción de los datos consideraremos algunos criterios de exclusión e inclusión, para los cuales en caso de los datos incluidos decidimos tomar en cuenta los datos que tengan un impacto directo o indirecto sobre las ventas en la tienda, así como la información que pueda variar en el tiempo dado que las ventas de las tiendas tienden a variar en el tiempo.

Para la exclusión de los datos tomamos en cuenta la información que no pueda variar en el

tiempo y que claramente no tenga ninguna relación con las ventas de las tiendas, además de retirar las columnas que tengan grandes cantidades de información faltante.

Por lo tanto la información que será incluida es la siguiente:

- ➔ Holidays_events.csv : Se mantendrán las columna "type" ya que dependiendo de la festividad celebrada aumentan o no las ventas; y la columna "date" para poder integrar más fácilmente la información dentro del dataframe ya que este es el indicador de las distintas celebraciones celebradas en sus respectivos días.
- ➔ Stores.csv : Las columnas "state" y "city", ya que nos ayudan a relacionar la locación geográfica de las tiendas con sus ventas ya que las tiendas ubicadas en lugares más concurridos tienen más posibilidades de ventas; y la columna "store nbr" ya que facilita el agregar las dos columnas anteriores al dataframe que integrara toda la información.
- ➔ Transactions.csv : Todas las columnas dentro de este archivo son de utilidad ya que nos indica el número de ventas que tuvieron las tiendas a través del tiempo.
- ➔ Oil.csv : En este archivo solo hay dos columnas, las cuales indican el precio del barril del crudo durante el día, y como Ecuador es un país que depende de esto significa que el precio de los barriles de petróleo crudo afectan las ventas de las tiendas indirectamente.
- ➔ Train.csv y Test.csv : Para el caso de estos archivos ambos contienen prácticamente la misma información con la excepción de la columna "sales" que se encuentra únicamente en el archivo "Train.csv"; por lo tanto la información que mantendrá en ambos archivos será virtualmente la misma. Obviando la columna "sales" del archivo "Train.csv", las columnas "store_nbr", "onpromotion" y "date" son las seleccionadas para ser mantenidas dado que dependiendo de la fecha,la tienda y los artículos en promoción serán las ventas que se realicen.

Para el caso de la información obviada esta simplemente será excluida dado que cumplió con los criterios de exclusión definidos anteriormente.

V. TRANSFORM

Para esta parte de la limpieza de los datos se realizará con ayuda de la librería de pandas one-hot encoding para convertir los datos categóricos en datos numéricos, e imputación de datos para los archivos que tengan información faltante, cuyo único caso después de haber filtrado las columnas de los archivos es "oil.csv", para el cual se usará el promedio para rellenar esa información faltante, por último después de haber realizado los cambios pertinentes a las columnas se alineará la información para que toda concuerde adecuadamente en relación a la fecha y la tienda para poder construir el dataframe final que contendrá toda la información para realizar las predicciones.

VI. LOAD

En esta última parte tenemos ya un dataframe limpio y listo para ser utilizado, por lo que por seguridad este será guardado como un archivo ".csv" a manera de respaldo, para así tener también una sencilla manera de acceder y revisar la información cada vez que sea necesario. Al momento no hay los conocimientos necesarios para llevar la información a una base de datos en

la nube o algún servidor, por lo que la principal fuente de consulta será el git sobre el que se ha trabajado el proyecto donde toda la información que se ha utilizado se encuentra disponible, tanto los archivos originales como el nuevo archivo con el dataframe limpio y la notebook sobre la que se programó toda esta parte de ETL's.

VII. Conclusión

En conclusión el proceso realizado para la limpieza de los datos fue laborioso, pero con la exploración adecuada de los datos así como su correcta limpieza hará que el modelo de predicción que se diseñe para resolver el reto funcione lo mejor posible, de manera que podamos predecir satisfactoriamente las ventas de las tiendas.

VIII. Anexos

Github:
https://github.com/3mi1ian0/EquipoJADE_Reto1_IA_ETL
Cocalc:
https://cocalc.com/projects/1bac3df3-ca14-45c5-885c-db71d3b236fb/files/#id=1fd97d

IX. Referencias

[1] 🐼 Simple ETL w/ Pandas. (2022, 18 agosto). Deepnote. https://deepnote.com/@rickyharyanto14-3390/Simple-ETL-w-Pandas-7c198322-fad0-4af7-a6f8-a7ac4c020b0d

[2] Informatica. (s. f.). What is ETL? https://www.informatica.com/se/resources/articles/what-is-etl.html

[3] Learn data cleaning tutorials. (s. f.). https://www.kaggle.com/learn/data-cleaning

[4] Learn pandas tutorials. (s. f.). https://www.kaggle.com/learn/pandas

[5] MatplotLib — visualization with Python. (s. f.). https://matplotlib.org/

[6] NumPy. (s. f.). https://numpy.org/

[7] Numpy and scipy documentation — Numpy and scipy documentation. (s. f.). https://docs.scipy.org/doc/

[8] Pandas.DataFrame — Pandas 2.0.3 documentation. (s. f.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[9] Pandas.DataFrame.to_sql — Pandas 2.0.3 documentation. (s. f.). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_sql.html

[10] SciKit-Learn: Machine Learning in Python — SciKit-Learn 1.3.0 documentation. (s. f.). https://scikit-learn.org/stable/

[11] Seaborn: Statistical Data Visualization — Seaborn 0.12.2 documentation. (s. f.). https://seaborn.pydata.org/

[12] StatQuest with Josh Starmer. (2018, 2 abril). StatQuest: Principal Component Analysis (PCA), Step-by-Step [Vídeo]. YouTube. https://www.youtube.com/watch?v=FgakZw6K1QQ

[13] Working with missing data — Pandas 2.0.3 documentation. (s. f.-a). https://pandas.pydata.org/docs/user_guide/missing_data.html#filling-with-a-pandasobject

[14] Working with missing data — Pandas 2.0.3 documentation. (s. f.-b). https://pandas.pydata.org/docs/user_guide/missing_data.html#cleaning-filling-missing-data