Sales Time Series Forecasting

Juan Felipe Vásquez Uribe
Jhon David Ballesteros Vargas
Natalia de Jesús Polo Peña

Introducción a la Inteligencia Artificial para ciencias e ingeniería.

Facultad de Ingeniería
Medellín
2023

RESUMEN:

En este informe se aborda las primeras impresiones del análisis de un modelo de series de tiempo , para esto se hace uso de un Dataset de la página de competencias www.kaggle.com , para llevar a cabo el desarrollo de esta competición se explora el dataset , visualizan los datos en búsqueda de tendencias y se limpia el conjunto de datos para evitar ruido en las predicciones , finalmente se lleva a cabo una estructuración de los datos de forma que sean más fácil para el modelo entenderlos y generar mejores predicciones.

Como parte importante del proyecto se indaga sobre diferentes modelos y se realiza un estudio para identificar si estos resuelven el problema de series de tiempo que se enfrenta.
Para llevar a cabo el análisis de los datos se hará uso del lenguaje de programación Python y sus librerías de gráficas y manejo de datos .

PALABRAS CLAVE: Modelo predictivo , Series de tiempo, random forest, algoritmos supervisados y no supervisados

1 INTRODUCCIÓN

Pronosticar la demanda de productos es una tarea común para científicos de datos actualmente. En efecto, un pronóstico más preciso por medio del aprendizaje automático permite garantizar a las empresas una mayor satisfacción a sus clientes, al realizar mejores predicciones de sus productos teniendo en cuenta las nuevas necesidades y gustos que surjan del mercado.

En este orden de ideas, de eso se trata este dataset, de encargarse de realizar predicciones en una tienda de Ecuador, esta tienda llamada Corporación Favorita es una gran minorista de dulces con varios puntos en el país.

El dataset seleccionado es de una competencia de www.kaggle.com que proporciona datos desde el 31 de diciembre de 2012 hasta el 15 de agosto de 2017.

Para este dataset se utiliza la métrica de desempeño Error logarítmico cuadrático medio.

2 METODOLOGÍA

2.1 *Métricas de evaluación*

La métrica de desempeño usada en este modelo será el error logarítmico cuadrático medio (RMSLE) el cual se calcula mediante la fórmula:

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

2.2 *Exploración de datos*

La competencia presenta 5 archivos que conforman la bases de datos para entrenar el modelo:

2.2.1 *train.csv*
se describe el total de ventas de cada familia de productos de cada tienda de cada día.

2.2.2 *transaction.csv*
especifica cuántas ventas se realizaron por día en cada tienda

2.2.3 *oil.csv*
contiene el precio del barril de petróleo en los días analizados

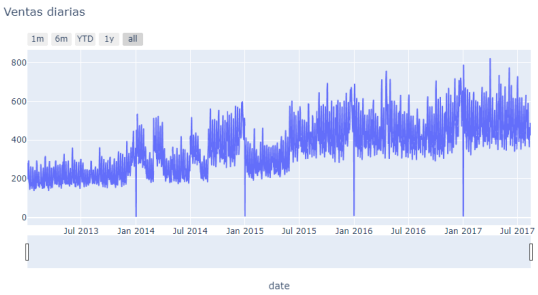
2.2.4 *test.csv*
pretende predecir el total de ventas en cada familia de productos en cada tienda en cada día

2.2.5 *holidays_events.csv*
describe los días de celebración o días especiales en Ecuador

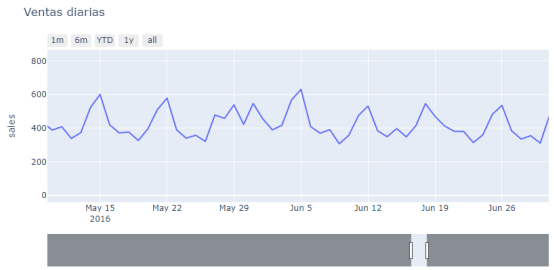
2.2.6 *stores.csv*
en este archivo a cada tienda se le asocia la ciudad, el tipo y el cluster (grupo de tiendas similares)

A continuación se hará una descripción breve de los hallazgos encontrados en cada archivo de datos.

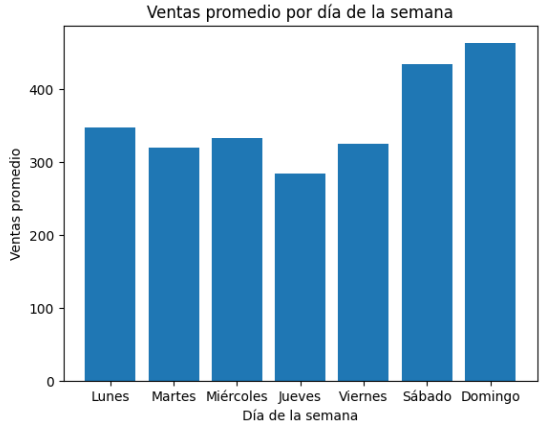
Train:



Cada registro de train data contiene las ventas de una familia de productos en una tienda específica para un dia.

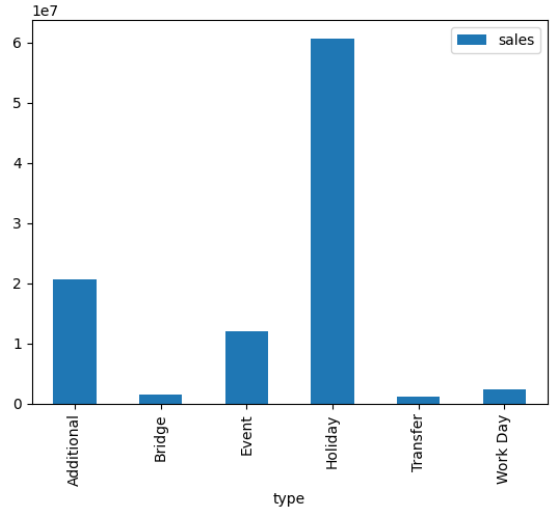


Mirando profundamente en los datos y sus comportamientos se tiene un indicio de que hay un patrón de ventas dependiendo del día .
Por lo general los fines de semana son los días de más ventas , lo que indica que una buena característica por añadir al dataset podría ser el dia de la semana

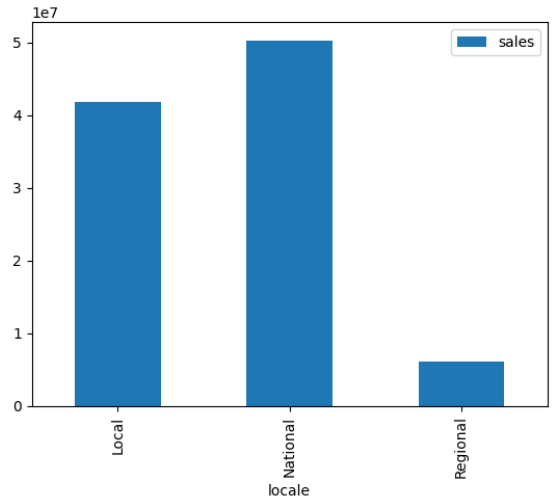


Holidays:

Esta base proporciona los días festivos en Ecuador desde el 2012 hasta el 2017, existe una columna “type” que los clasifica como: puente, evento, día de trabajo, transferido, día feriado o adicional. Otra columna que indica si se trata de un feriado local, regional o nacional, la descripción o nombre del evento y una última columna que indica si el día festivo fue transferido al siguiente lunes. A continuación se presentan las relaciones del tipo de evento con las ventas



La gráfica indica que la mayoría de las ventas se relacionan con los eventos feriados o Holidays. Los tipos de evento puente y transferido se les asocia muy pocas ventas, seguramente porque en proporción del total de festividades son los más bajos.



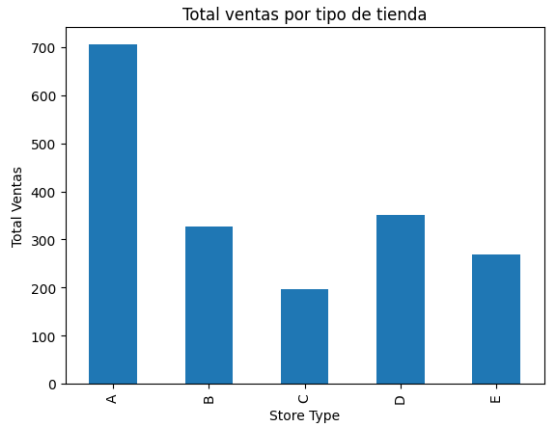
La gráfica indica que la mayoría de las ventas se realizan en los eventos nacionales que representan aproximadamente el 50% del total de eventos.

Ahora bien, analizando las fechas de la predicción: del 16 al 31 de agosto de 2017, solo existen dos días feriados: la fundación de Riobamba el 15 de agosto y la fundación de Ambato el 24 de agosto. Ambos eventos son de tipo feriado de clasificación regional. Por lo tanto, en primera instancia se cree que la influencia de los días feriados en las fechas a predecir no serán un factor determinante.

Stores:

La base de datos Stores clasifica las 54 tiendas participantes en los reportes de ventas, cada tienda se identifica por un número del 1 al 54, se le especifica su ciudad, su estado, el tipo de tienda

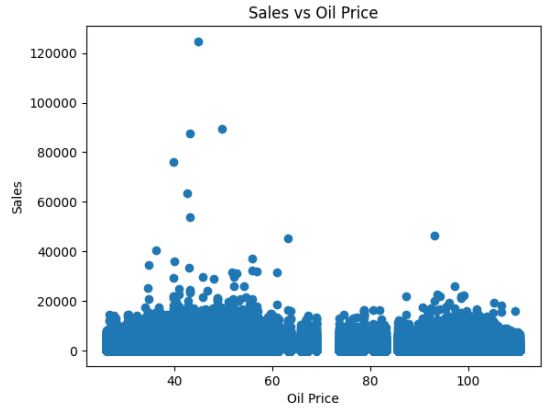
como A,B, C o D, y el cluster al que pertenece, habiendo en total 17 de estas categorías.



Oil:

Se obtuvo la correlación directa del precio del petróleo con las ventas en (train) , la correlación se detalla en la siguiente gráfica:

```
Correlation with Daily Oil Prices
-0.0790501602199275 4
```



Se tiene una correlación entre las ventas diarias y el precio del petróleo de (-0.079), al ser negativo indica que las ventas diarias disminuyen cuando el precio del petróleo aumenta pero no es una relación fuerte , inicialmente no se tendrá en cuenta este archivo , pero se tendrá en cuenta para futuros cambios al modelo

2.3 *Primeros algoritmos a considerar*

En machine learning existen varios grupos de algoritmos , este proyecto se centrará en la implementación de algoritmos supervisados y no supervisados .

Para modelos de pronóstico de ventas con series temporales se recomienda el uso de algoritmos supervisados tales como ARIMA o algoritmos de redes neuronales [2][3].

En este primer acercamiento de entrenamiento y predicción se hizo uso de un algoritmo supervisado llamado Random Forest Regressor con este se busca identificar patrones importantes en la serie temporal entrenando el modelo con datos históricos y buscando obtener predicciones sobre las tendencias en ventas futuras.

En avances futuros del proyecto se tendrán en cuenta otros algoritmos supervisados como SVM , RNN y ARIMA y algoritmos no supervisados como Clustering.

3 ALCANCES Y LIMITACIONES

Dentro de los primeros alcances de este proyecto:

- Se encontró poder predecir las ventas de los siguientes 15 días.
- Crear un entorno de desarrollo que permita comparar la predicción y precisión de cada algoritmo.
- Comparar desempeño de algoritmos supervisados y no supervisados

Las limitaciones que se han encontrado hasta la fecha en el desarrollo de este proyecto son :

- Dataset con muchos datos , esto en ocasiones puede detener la ejecución del notebook por consumo total de recursos.
- Tiempo de entrenamiento del modelo es demasiado largo , en promedio 25 minutos para Random Forest Regressor
- Al tener muchos datos y variables categóricas que serán convertidas a variables binarias el se debe estudiar detalladamente cada archivo de datos para depurar la información más importante y disminuir el set de entrenamiento sin perder precisión.

4 SIGUIENTES ETAPAS

4.1 *Algoritmos supervisados*

Estudiar y probar diferentes algoritmos supervisados en especial algoritmos estadísticos como ARIMA y otros algoritmos que se están comenzando a utilizar últimamente para pronóstico de series temporales como RNN (redes recurrentes) o LSTM.

4.2 *Algoritmos no supervisados*

Investigar qué algoritmos no supervisados pueden tener un buen desempeño con predicciones de series de tiempo, estudiar cómo organizar y limpiar los datos para que este tipo de algoritmos sean capaces de comprenderlos .

4.3 *Preparación de datos*

Investigar más a fondo cada archivo que compone el dataset para identificar aquellas variables que ayudan al modelo a ser más preciso y que hasta el momento no se han tenido en cuenta .

4.4 *Despliegue en producción*

Investigar detalles y procesos para el despliegue en producción de este tipo de modelos , se busca resolver preguntas como :
¿ Cuál es el proceso de despliegue ?
¿ Validación y puesta a punto en el entorno de producción ?

5 RESULTADOS

El primer avance del proyecto arroja predicciones con RMLSE = 3.07 unidades, al momento de subirlo a la competición de Kaggle se obtiene un score de 2.4062.Se busca mejorar este score analizando nuevos algoritmos e investigando con más detalle la correlación de las variables de los distintos archivos del dataset con las ventas.

6 REFERENCIAS

[1] Store sales time series forecasting - https://www.kaggle.com/competitions/store-sales-time-series-forecasting/data?select=train.csv
[2] Algoritmos ML o estadisticos https://sitiobigdata.com/2019/05/01/two-thoughts-on-the-question-are-times-series-models-considered-part-of-machine-learning-or/#
[3]Time series Forecasting with Supervised Machine Learning https://towardsdatascience.com/time-series-forecasting-with-machine-learning-b3072a5b44ba