Universidad de Antioquia

Jhonatan Felipe Sossa Rojo – Fundamentos de Deep Learning

Maestría en ingeniería de telecomunicaciones

Contexto del problema

La predicción de la demanda de productos es una tarea habitual para las empresas y recientemente quienes están a su cargo son los profesionales del área de la ciencia de datos. Las previsiones son esencialmente importantes para tiendas de comestibles, pues pronosticar por encima implica un exceso de inventario de productos perecederos y pronosticar por debajo finalmente se traduce en pérdida de ventas, reputación y clientes. En general, las empresas quieren conocer cuál es la demanda de sus productos en cada uno de sus puntos de venta para hacer una distribución correcta de los mismos y además llegar a los puntos de venta en el tiempo correcto.

Descripción de la solución

Para pronosticar la demanda (ventas) de los productos en cada uno de los puntos de venta, se hizo uso de información histórica de ventas, información geoespacial de cada punto de venta, información histórica del precio del petróleo y finalmente información de los días festivos ocurridos durante este tiempo. Cruzando toda esta información, se esperaba obtener un modelo de aprendizaje profundo que fuera capaz de identificar cuántos productos de cada familia se van a vender en cada punto de venta.

La solución consistió en tres fases, entre las cuales se encuentran: ingeniería de datos, análisis de datos y ciencia de datos.

1. **Ingeniería de datos:**

   Durante esta fase se realizó todo el proceso de limpieza de los datos así:

   i) Eliminación de datos duplicados
   ii) Remover caracteres especiales
   iii) Eliminar espacios en blanco al inicio y al final de cada dato
   iv) Eliminar acentos
   v) Dar formato a las fechas
   vi) Convertir en minúsculas todas las cadenas de caracteres

   Una vez los datos estaban limpios, se consolidaron para crear un maestro transaccional que tuviera toda la información que venía de las fuentes de datos limpias.

   El pronóstico que se deseaba realizar era agregado por semana (demanda por familia por punto de venta por semana) y la información obtenida del origen era diaria. Por lo tanto, se debía agregar la información por semana. Esta información de ventas por semana contenía ya información del número de productos de cada familia que estaban en promoción cada semana en cada punto de venta. Se extrajo el rezago de cada semana con respecto a las

cuatro semanas anteriores, pues se consideró que conocer dichos estados anteriores podía otorgar información relevante con respecto al aumento o disminución de ventas en cada semana.

Luego se cruzó la información de ventas semanales con la información geográfica de cada punto de venta, pues sin duda la ciudad, el estado y el tipo de cada punto de venta influye en la demanda de cada familia de productos.

Se cruzó también con el precio promedio del petróleo por semana, ya que Ecuador (país donde están los puntos de venta) es un país dependiente del petróleo y los precios de los productos cambian frecuentemente con respecto al precio de este. Al precio del petróleo semanal se le extrajo el rezago de las cuatro semanas anteriores.

Desde el origen se contaba con la información de todos los festivos ocurridos en el lapso de tiempo de estudio. El tipo de festivo, la localización y su nombre son información importante que puede influir en el patrón de venta de los productos. Por semana se contaron la cantidad de festivos del mismo tipo y que tuvieran el mismo alcance. Esto porque no todos los festivos son comerciales, por el contrario, hay muchos de ellos en los que la demanda de los productos disminuye. Por esta razón es importante hacer una buena segmentación de lo festivos.

Con la información anterior, ya se tienen un total de 52 características con las que se puede hacer un pronóstico de ventas tal y como se desea.

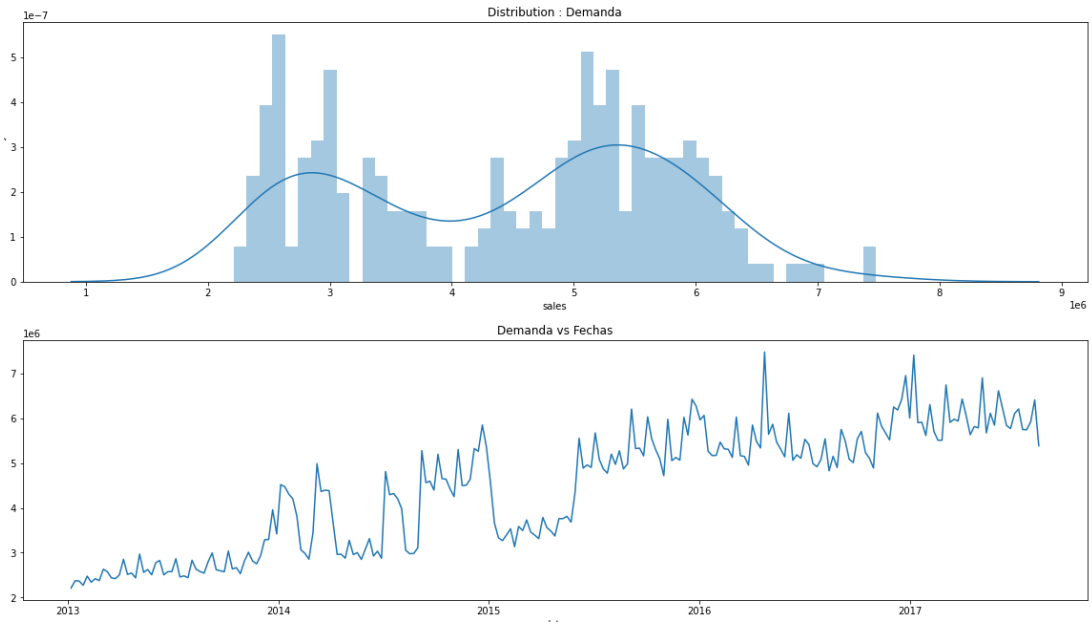
2. Análisis de datos

Durante la fase de análisis de los datos se realizó una revisión a los datos para confirmar datos tales como: El número de los puntos de venta en los que se venden los productos (54) y el número de familias de productos que son vendidos (33).

Los datos anteriores son de relevante importancia ya que dicen la cantidad de combinaciones puntos de venta – familias de productos para las cuales se debe realizar un pronóstico. En total, se tienen 1782 combinaciones de demandas las cuales se debían pronosticar y cada una de ellas tiene un patrón diferente.

Durante el proceso de análisis se descartó que el número de familias de productos o puntos de venta cambiaran con el tiempo, pues esto agregaría un grado más de dificultad al pronóstico. El cambio con el tiempo de productos o tiendas implica una canibalización de esa demanda por parte de otra combinación diferente.

La gráfica siguiente muestra la distribución de la demanda (agregada por semana, es decir, la demanda de TODOS los productos y puntos de venta sumados) y el patrón de ventas a lo largo del tiempo.

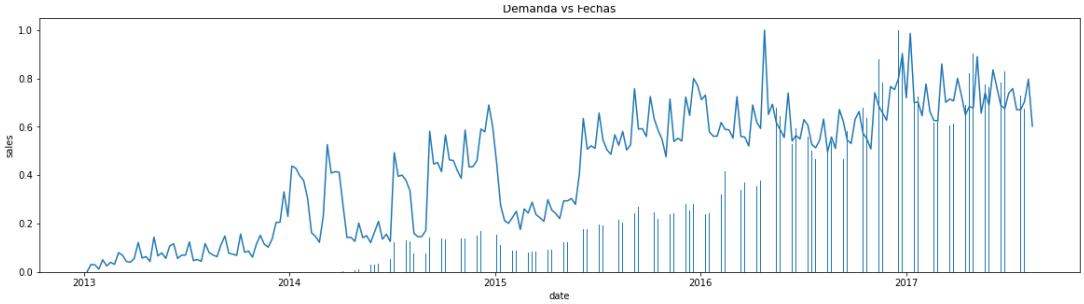


La suma de las ventas agregadas, si bien no es lo que finalmente se desea predecir (al fina l se desea obtener las ventas de cada familia por punto de venta), muestr- a la tendencia de las ventas en el tiempo. En el 2013 las ventas estaban por el orden de la s dos millones de venta por SEMANA y en el 2017 por los seis millones, lo que muestra un a tendencia positiva.

Hay semanas en las que se presentan picos grandes de ventas con respecto a sussemana s colindantes, así que es interesante saber qué pasó en esas semanas es- pecíficas. Pudo haber días festivos, descuentos o puede deberse al precio del petróleo.

A principios del 2015 hubo una reducción significativa de las ventas, es importante averigü ar a qué pudo deberse y saber si va a afectar el pronóstico futuro

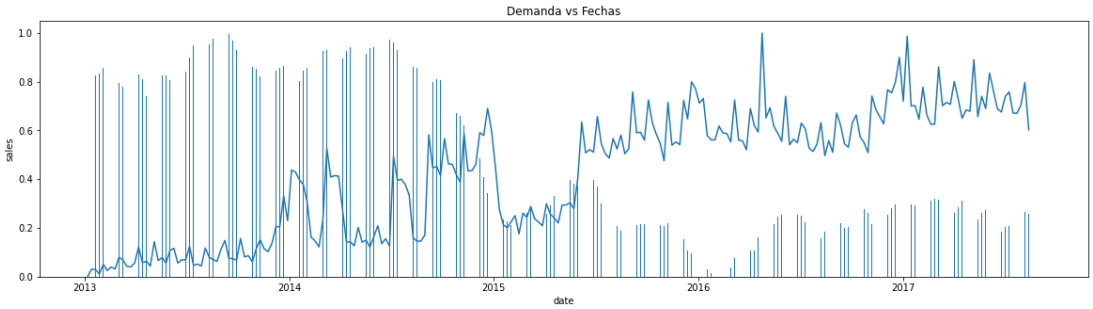
A continuación, se muestra la relación entre la demanda y el número de productos en descuento por semana



Con la gráfica anterior observamos que el número de productos en descuento en cada semana explican de forma coherente el patrón de ventas. Cuando hay una disminución en los productos promocionados entre una semana y otra, generalmente observa un valle en las ventas. Asímismo, cuando hay un aumento en los productos promocionados se observa un pico en las ventas.

A partir del primer cuarto del 2016 se observa que, aunque se aumente la cantidad de productos en promoción, ya la demanda no aumenta en la misma proporción con años anteriores. Esto puede significar que la demanda de los productos llegó a un límite o que hay otra variable importante afectando la demanda a partir de ese trimestre.

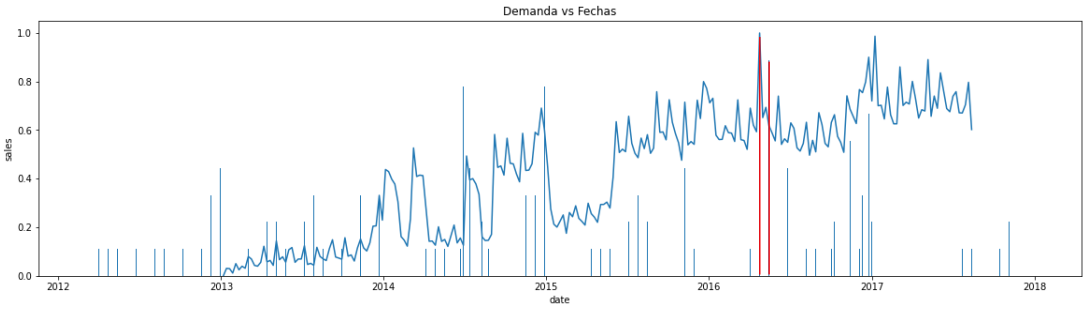
Se procede entonces a analizar la relación entre el precio del petróleo y el patrón de ventas agregado



En este caso, vemos una relación inversamente proporcional entre el precio del combustible y la demanda de los productos que ofrece la compañía.

Además de la cantidad de productos en promoción, el precio del petróleo en cada semana logra explicar bastante bien el patrón de demanda. Definitivamente, el precio del petróleo es algo que afecta las ventas de la empresa. En los últimos años el bajo precio del petróleo ha significado un aumento en las ventas de los productos ofrecidos.

Por último, se analiza la relación entre las ventas y los días festivos ocurridos durante el tiempo de interés.



La cantidad de días festivos que ocurren en la semana no explican tan bien como las dos anteriores todo el patrón de demanda, pero sí logran explicar algunos de los picos más importantes de ventas.

El día 16 de abril del 2016 ocurrió un terremoto en Manabí, lo que ocasionó que muchas personas donaran bienes de primera necesidad a los afectados. Claramente esto aumentó la cantidad de venta durante los días posteriores al suceso. Las autoridades de panamá declararon festivos los siguientes 30 días hábiles después del terremoto.

Es normal que no todos los días festivos afecten de igual forma el patrón de demanda, pues no todos ellos son días comerciales.

Por lo anterior, esta es una variable que debe tratarse con cuidado, pues tratar todos los días festivos por igual podría confundir el modelo cuando solo una parte de ellos explican ventas.

3. **Ciencia de datos**

   Una vez realizados el análisis y la ingeniería de los datos, la última fase es preparar los datos para ser ingresados a un modelo de aprendizaje profundo y realizar el pronóstico de las ventas.

   El primer paso fue extraer 'dummies' de las variables categóricas para completar el conjunto de características. Durante la extracción de las dummies se descarta una de ellas para evitar la correlación que se genera del proceso de extracción.

   Para el pronóstico de la demanda fueron utilizados dos modelos: Uno basado en aprendizaje profundo con una LSTM y un XGBoost con el propósito de comparar el rendimiento de ambos. En la sección descripción de los resultados se presentan y comparan los resultados.

   El conjunto de datos histórico comenzaba en el año 2013 hasta el año 2017. El año 2017 fue utilizado para testear la capacidad de los modelos mientras que hasta el año 2016 se utilizó para el entrenamiento de estos.

   Los modelos resultantes se encuentran en la carpeta 05_model y las predicciones en 06_model_output.

Descripción de la estructura de los notebooks

Para la implementación se utilizaron 5 notebooks nombrados de la siguiente manera:

+ 01 - data_cleansing.ipynb
+ 02 - exploracion de datos
+ 03 - Agregación de datos y análisis de serie temporal
+ 04 - Obtener dummies
+ 05 - Modeling

Se diseñó también un lago de datos que contiene los datos con todas sus transformaciones. Los datos en raw son los datos tal cual como se obtuvieron del origen y suelen llamarse datos crudos, están en formato csv. A partir de las siguientes zonas los datos son almacenados con formato parquet y comprimidos con snappy. Lo anterior se decidió para conservar la estructura de los datos y ocupar menos espacio de almacenamiento. Para accederlos puede utilizarse la librería pandas para cargarlos.

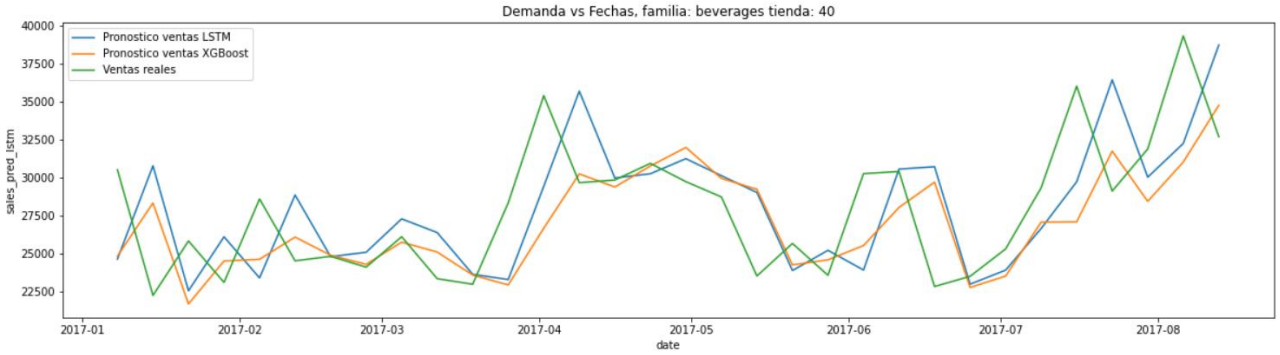
+ 01_raw
+ 02_intermediate
+ 03_primary
+ 04_model_input
+ 05_model
+ 06_model_output

Descripción de los resultados

El resultado del pronóstico realizado con LSTM arrojó un RMSE de 1695.785 mientras que el pronóstico de XGBoost arrojo un RMSE de 1335.651. Estos resultados sugieren que los modelos aún no cuentan con una alta capacidad para adaptarse a los datos pero aún así, pueden logran una predicción decente.

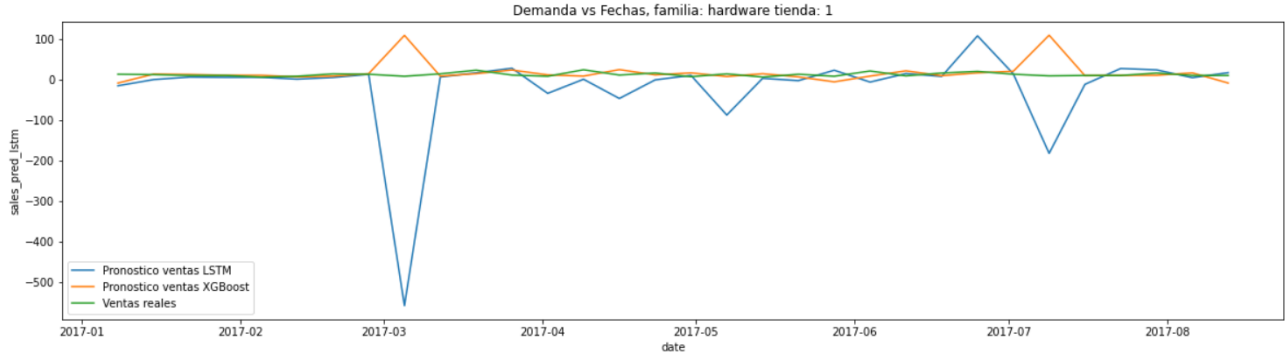
Era de esperarse que los modelos no pudieran adaptarse a todos los patrones de ventas de las familias y las tiendas combinadas, pero, como lo veremos a continuación, pueden seguir muy bien la forma de muchas de las combinaciones y del patrón agregado

Patrón de ventas de la familia ‘beverages’ en la tienda número 40:



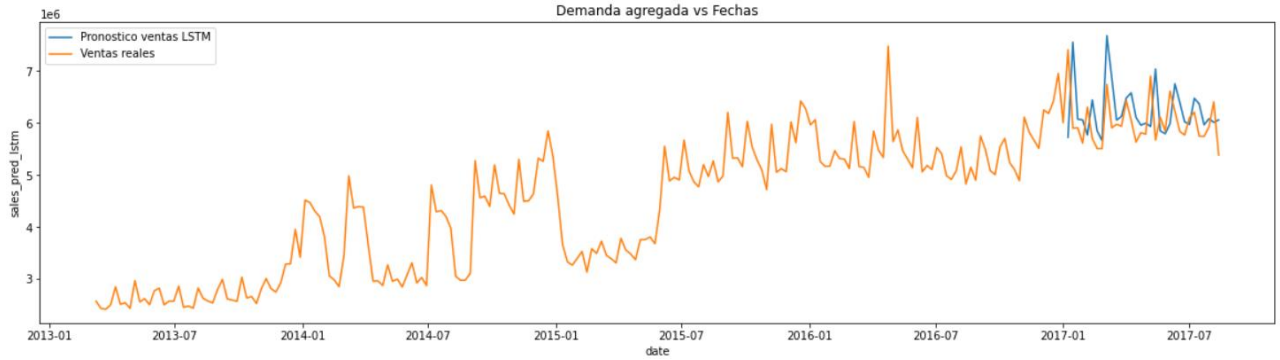
Ambos modelos tienden a seguir de forma adecuada la forma del patrón de ventas y mantienen una magnitud aproximada. El LSTM sigue casi a la perfección el patrón, pero está un poco desfasada con respecto a las ventas original. El XGBoost, aunque lo intenta, no puede seguir todos los cambios que presenta el patrón.

Patrón de ventas de la familia ‘hardware’ en la tienda número 1:

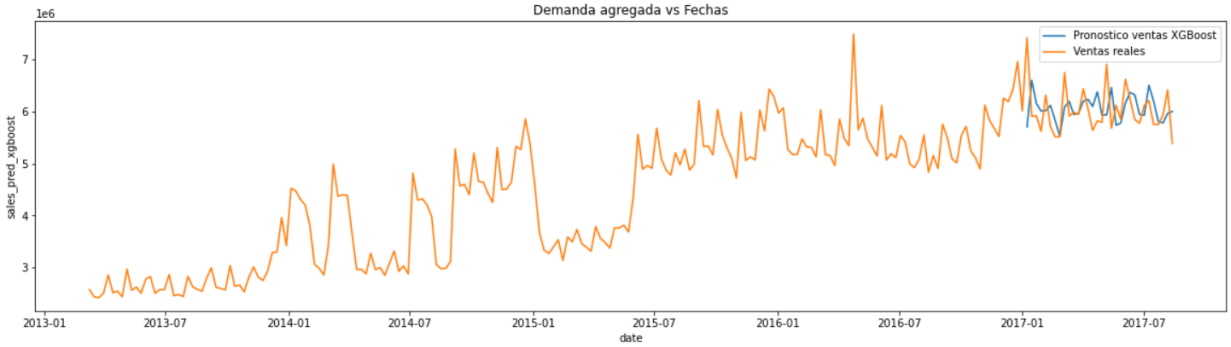


En este caso, por ejemplo, ambos modelos tienen patrones que no acompañan fielmente el patrón de demanda real. Como este, hay muchos casos en los que LSTM, principalmente, pierde mucho la magnitud. XGBoost no sigue los cambios de forma acertada pero no se aleja tanto de la magnitud total.

Finalmente, el patrón de demanda sumado para LSTM



Patrón de demanda sumado para XGBoost



Como se percibía con los patrones individuales, el patrón de demanda sumado muestra como el modelo con XGBoost mantiene más la magnitud, pero no está en la capacidad de seguir todos los cambios presentados por el patrón. El modelo con LSTM, aunque tiene un retraso y un offset positivo con respecto al patrón de ventas original, podemos decir que sigue bastante bien el patrón de demanda y si se corrigen de forma adecuada, presenta unos resultados coherentes y beneficiosos para el pronóstico de las ventas.

Oportunidades de mejora

- Durante la etapa de entrenamiento de los modelos, no se implementó optimización de parámetros para el XGBoost y se utilizaron pocas épocas para el entrenamiento de la LSTM debido a la alta volumetría de los datos, hacerlo podría aumentar significativamente el rendimiento de ambos modelos y la corrección de los problemas actuales.

- Adicionar información demográfica al modelo podría mejorar la descripción del patrón de ventas en cada punto de venta pues el estrato, la cantidad de población, sus costumbres, etc, influyen directamente en sus patrones de consumo.