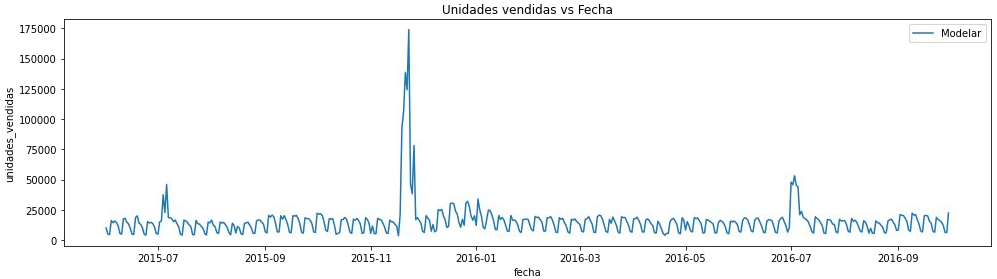
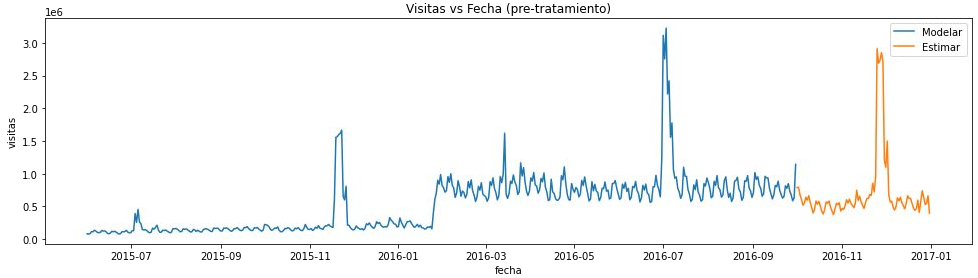
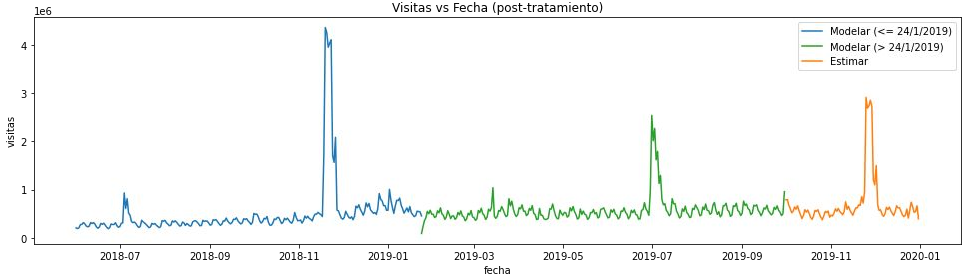
1. Tratamiento de datos y EDA

Universitat de Barcelona
Equip Autoestima







Tratamiento de datos

- Las fechas son erróneas. Los datos son del 2018-2019, no del 2015-2016.
- Esto es muy relevante, porque los días de la semana, que son determinantes, son incorrectos.
- Además, el 2016 fue bisiesto y el 2019 no, lo que genera un desfase con dos patrones diferentes.
- Las visitas, también determinantes, están multiplicadas por 5 en los datos de modelar, pero sólo en un rango. También en los de estimar.
- Las unidades vendidas están multiplicadas por 3 para todos los datos de modelar.
- La antigüedad del producto es estática y no cambia con el paso del tiempo, y además está desfasada entre los datos de modelar y los de estimar.

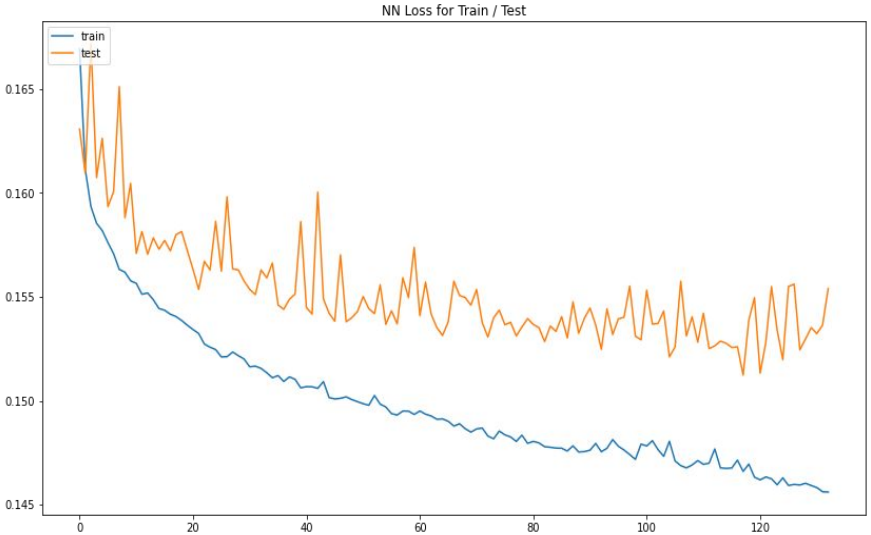
Análisis Exploratoria de Datos

- Fines de semana y festivos no associados a compra de regalos muestran disminución de visitas y ventas.
- Hay una correlación significativa entre el día de la semana y las ventas.
- Los datos tienen una fuerte periodicidad semanal a la que se le suma picos i mínimos ocasionales.
- Los clientes tienen más tendencia a visitar y no comprar los fines de semana.

2. Modelo y Entrenamiento

Para entrenar el modelo hemos alterado en gran medida los datos. Hemos realizado todas las correcciones de fechas y visitas que el análisis de datos han descubierto.

Todas estas transformaciones han resultado en una mejora de la métrica obtenida inicialmente.



El tipo de modelo que hemos escogido para computar las estimaciones es el de redes neuronales, apoyándonos en la librería *Keras* de python. Para entrenar hemos utilizado aceleración de hardware por GPU.

Hemos probado otros modelos, incluyendo regresiones lineales, *Random Forest Regressors*, *Autogluon*, *LightGBM* y similares, pero las mejores métricas han sido con redes neuronales.

Como se puede apreciar en la gráfica, se trata de unos datos difícilmente generalizables debido al desviamiento del error en entrenamiento vs. validación, pero en general podemos aprender levemente de los datos. Para las predicciones finales utilizamos el modelo que mejor ha trabajado en los datos de validación.

En un futuro vemos oportunidades de mejora utilizando librerías específicas para Time-Series como por ejemplo Dart.

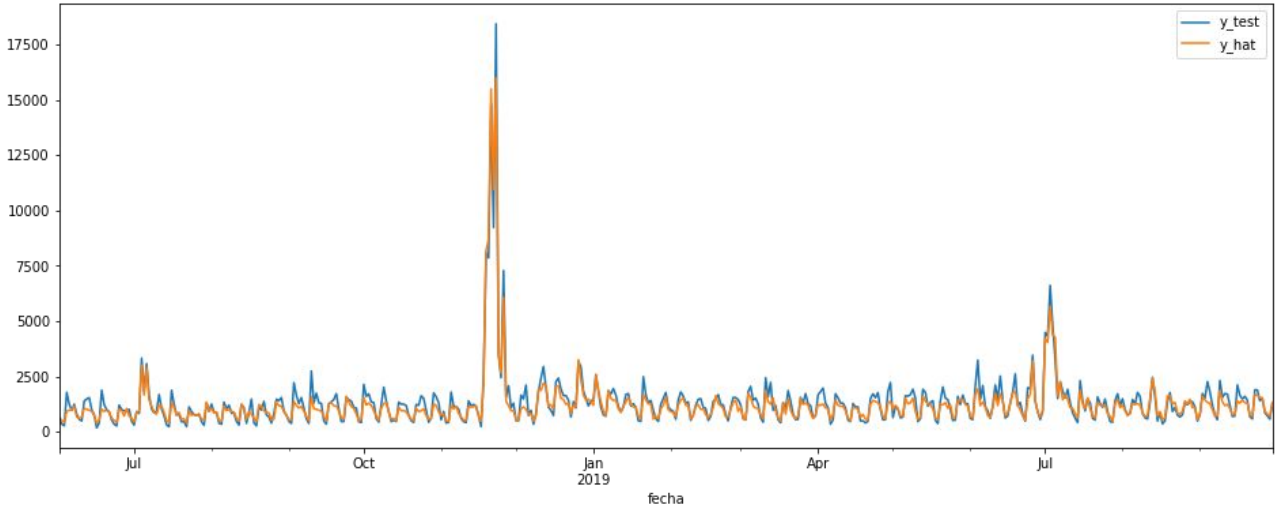
3. Resultados Obtenidos

Para evaluar el modelo obtenido hemos evaluado sobre un gràfico cómo de cercana es la suma total diaria de ventas real (azul) y estimada (naranja).

Debido al factor de casos favorables, encontramos una mejora en la métrica del score si modificamos la estimación para reportar siempre una unidad vendida de más

En el gráfico podemos observar que, al menos diariamente, las predicciones tienen sentido y se adhieren a los datos reales.

Ponemos especial atención al periodo de más ventas entre noviembre y final de año, donde la métrica se mantiene estable pese al pico de ventas.



Metrica de Score estimada:

Caso General:

2.208

Periodo Black Friday:

2.247