Executive Summary

Emiliano Islas y Raúl Mora

Dentro de 1C company existe una problemática grave con relación al inventario de nuestras tiendas. Esto porque alrededor del 23% de nuestro inventario está en sobrestock, obligándonos a liquidar con descuentos, y también tenemos quiebres de stock en productos clave el 18% del tiempo, generando pérdidas millonarias. Cabe aclarar que contamos con 60 tiendas y 22,170 distintos productos, y nuestra meta es predecir las ventas por producto en cada tienda durante cada mes. Actualmente ajustamos inventarios cada 14 días, mientras que nuestra competencia directa lo hace en 48 horas. Esto puede cambiar con la implementación de un modelo de aprendizaje de maquina para forecasting (predicción) a nivel producto-tienda-mes. Para la correcta implementación de este modelo se cuenta con una base de datos lo suficientemente significativa, pues hay alrededor de 2.9 millones de registros a lo largo de 3 años. A grandes rasgos, el objetivo es desarrollar un modelo de aprendizaje de maquina ent to end que prediga ventas fututas en retail para poder maximizar nuestras ganancias.

Nuestra solución como científicos de datos es usar un modelo conocido como “Gradient Boosting”. De esta manera lograremos reducir una métrica muy importante conocida como RMSE, la cual actualmente se tiene con un valor de 11 unidades, lo que representa un turnover de 6.2x, y con la ayuda de este modelo se llega a bajar considerablemente hasta 1 unidad y con un turnover mucho mayor. De manera resumida, estos modelos nos ayudan a comprender comportamientos pasados, para poder predecir el futuro, esto a partir de las bases de datos ya existentes, aunque con algunas modificaciones realizadas a nuestra conveniencia, claro.

Evaluaciones

Se evaluaron una regresión lineal (Ridge) como baseline y un modelo no lineal (cómo la naturaleza de estos datos) de Gradient Boosting, el cual fue elegido modelo principal. Con este modelo de obtuvieron:

- RMSE = 0.96, lo que indica un error promedio cercano a una unidad
- MAE = 0.33, por lo que el error típico es significativamente menor a una unidad
- R2 = 0.29, por lo que el modelo explica el 29% de la variabilidad de ventas mensuales
- MAPE = 71%, que es alta volatilidad relativa en escenarios de ventas bajas

El modelo que usamos captura de manera efectiva la persistencia temporal mediante variables rezagadas (lags) , aprovechando el historial reciente como principal fuente

de señal predictiva. La baja MAE indica que, en la mayoría de los casos, el error absoluto es menor a una unidad, lo cual es adecuado dado el bajo volumen típico de ventas , ya que alrededor del 75% de las veces se compran solo un producto por tienda al mes. De igual manera El clipping del target estabiliza el entrenamiento y reduce el impacto de valores extremos que pudieron haber influenciado de más al modelo. No obstante, el R2 tan bajo es importante pues muestra que la variabilidad no puede ser explicada únicamente con datos históricos de ventas.

Escenarios con ventajas y desventajas del modelo

Este modelo funciona mejor cuando los productos presentan patrones de ventas recurrentes cada mes. También si existen ventas recientes, es decir que no haya cero ventas prolongadamente, lo que le permite a los lags aportar señales útiles. No obstante, no es bueno para predecir picos de demanda excepcionales, o productos con ventas esporádicas, o combinaciones de producto y tienda con poco historial, ya que este es un modelo que se basa fuertemente en la experiencia.

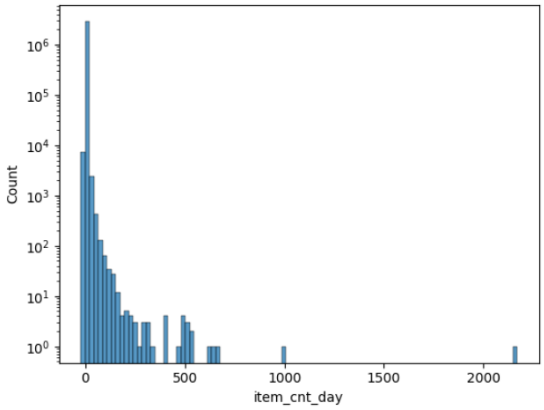


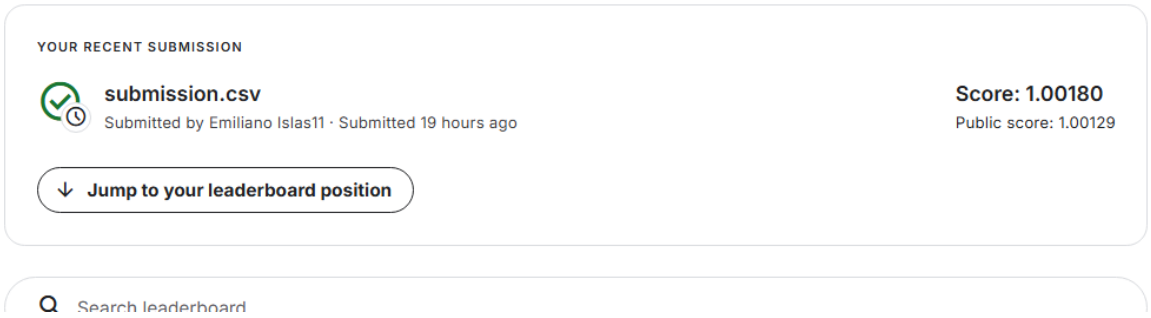
Figura 1. Distribución de ventas diarias por producto–tienda

Figura 2. Ventas reales vs ventas predichas

En la Figura 1 se puede ver que hay outliers en valores altos por lo que la gráfica está sesgada a la derecha. La mayoría de los valores se encuentran en valores pequeños. Para un producto específico, en una tienda específica, en un día específico, lo más común es vender 0 o 1 unidades. De igual forma, en la Figura 2, con las predicciones realizadas para el mes 34, se muestra una alta concentración de observaciones en el rango de 0 a 5 unidades, consistente con la distribución del dataset. El modelo presenta un buen alineamiento en este rango, que representa la mayoría de los casos. Para valores más altos, se observa mayor dispersión, lo que indica menos precisión, lógico con valores atípicos.

En conclusión, algunas recomendaciones adicionales en torno a los próximos pasos sería incorporar nuevas variables que nos fueran de ayuda como precios, promociones y campañas para mejorar la explicación de picos de demanda. Otra meta a futuro para la recepción de los datos sería segmentar productos por nivel de rotación (alta, media, baja) y entrenar modelos diferenciados para cada nivel.

Anexo 1



Resultado de RMSE en competencia Kaggle