UNIVERSIDAD AUSTRAL

TRABAJO FINAL

LABORATORIO DE IMPLEMENTACIÓN III

INTEGRANTES

FIERRO ABEL – SALDANIA GUSTAVO - SANTESTEBAN NATALIA

CASO DE NEGOCIO.

La empresa objeto de este análisis es una multinacional Argentina que busca predecir las toneladas (tn) a vender de 780 productos para el mes de febrero 2020.

Para lograr la misma se cuenta con la información de ventas históricas desde enero 2017 hasta diciembre 2019.

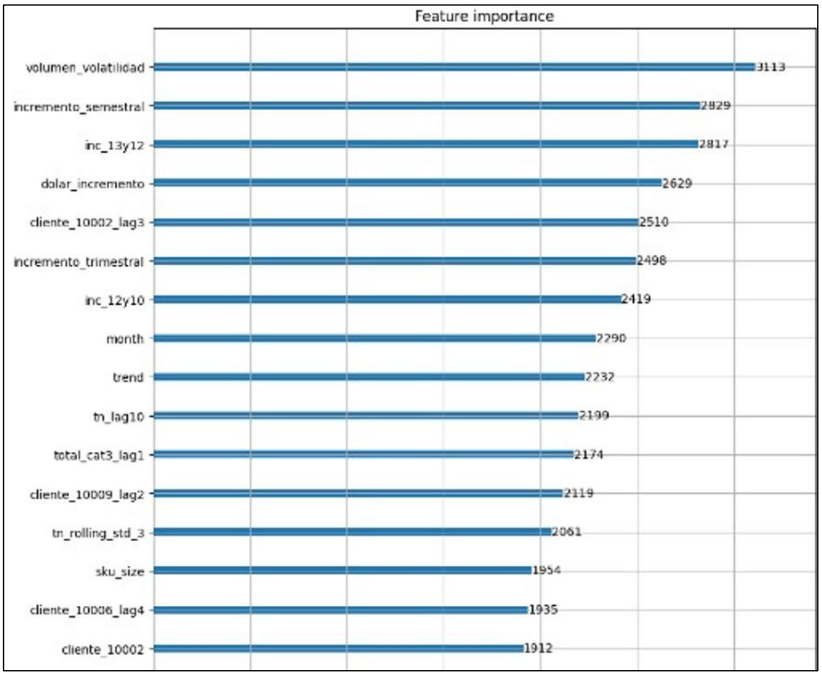
Diseño experimental.

Llegar al modelo final fue el resultado de aplicar los siguientes pasos y estrategias.

1. Procesamiento y análisis exploratorio (EDA).

2. Construcción de dataset base.
   Para la construcción de este se integró información de las ventas por periodo, producto y cliente. Se agregaron los volúmenes totales y incorporaron columnas para para indicar el consumo en toneladas [tn] por cada cliente del producto y en el periodo correspondiente. Además, se completaron con ceros a partir de la primera compra de cada producto los meses para los cuales no había datos de estos, tanto en relación a las toneladas totales como las respectivas a cada cliente.

3. Feature Engineering.
   Se incorporaron nuevas variables al dataset principal buscando garantizar continuidad temporal por producto y completando periodos faltantes. Se incluyó información de clientes, datos macroeconómicos, variables políticas y cluster generados a partir del comportamiento de la demanda por producto calculados a partir de la matriz de distancias de dtw (Dynamic Time Warping).
   También se sumaron la tendencia y la estacionalidad de los consumos obtenidos a través de Prophet y se incluyeron, además la elasticidad entre los productos, capturando relaciones de sustitución o complementariedad.
   Se analizó el ciclo de vida del producto generando variables de madurez, tendencia, estacionalidad y volatilidad.
   Se sumaron variables aleatorias “canaritos” para controlar la importancia de las features.



4. Particionado de datos.
   Se trabajó con un conjunto de entrenamiento hasta agosto 2019, otro conjunto de validación hasta octubre 2019, obteniendo las siguientes métricas:
   R2: 0.9537
   RMSE: 17.1635
   MAE: 6.2801
   Total Forecast Error: 28.53%

   El entrenamiento final se realizó con dataset que incluye datos de entrenamiento hasta octubre 2019 y se usó para la predicción los datos hasta diciembre 2019.
   Total Forecast Error (lista 780): 24.19%
   Total Forecast Error (general): 28.30%

5. Estrategias.

   AutoGluon.

   LightGBM con Features Engineering.

   Optimización de parámetros con Optuna.

   Se utilizó un LightGBM base con el objeto de establecer una línea de referencia para la performance de manera de poder compara contra modelos ajustados.

   En una segunda instancia se incorporó un LightGBM con poda de variables para poder reducir la complejidad y el sobreajuste manteniendo mejora de la performance, con el objeto de poder eliminar variables con baja importancia y todos las “canaritos”, optimizando interpretabilidad y velocidad.

   Al final se utilizaron modelos con features específicas entre las cuales se destacan aquellas vinculadas a las de elasticidad y el ciclo de vida de los productos.

6. Ensamble y selección del modelo final.

   Se hizo uso de un semillero para promediar diferentes ligthGBM, tratando de encontrar una mejora de los valores obtenidos y en última instancia se aplicó el promedio entre el semillero de LightGBM y AutoGluon.

   Clase. Toneladas + 2

   Mejores parámetros:

   num_leaves= 94,
   learning_rate= 0.015293495120602733,
   n_estimators= 967,
   min_data_in_leaf= 8,
   reg_alpha= 12.320557276828628,
   reg_lambda= 12.390693664291984,
   feature_fraction= 0.730906609846672,
   bagging_fraction= 0.6955968409148202,
   bagging_freq= 2,

min_split_gain= 0.5826709316464246,
random_state=42

Conclusión.

Las fortalezas del modelo aplicado están asociadas al uso de variables vinculadas al ciclo de vida del producto y market share que mejoraron notablemente la precisión.

Las elasticidades captaron relaciones competitivas que no eran evidentes en los datos originales.

Aplicar el método de poda elimina ruido y redujo la complejidad sin pérdida de rendimiento.

Respecto a las limitaciones o puntos que no favorecieron al modelo se destacan la presencia de productos con bajo niveles de historia en ventas que presentan predicciones pocos confiables.

La predicción no parece ser muy buena para periodos de alta volatilidad o ante la presencia de eventos exógenos no contemplados.

Tanto las variables de clusterización, de tendencia y estacionalidad obtenidas con Prophet se ubicaron dentro de las features más importantes.

Las pruebas realizadas con Sample_weight y linear_tree = True en nuestro caso, empeoraban las métricas del modelo.