Maestría en Ciencia de Datos

Laboratorio de Implementación III

Trabajo Final

Docente

Gustavo Denicolay

Grupo:

Federico Grijalba

Florencia Santiago

Paola Szekieta

10-08-2025

Informe Trabajo Práctico Final

1. Contexto y Desafío

La empresa en cuestión es una multinacional líder en el sector de consumo masivo, con presencia en 170 países y una participación de mercado en Argentina que oscila entre el 50% y el 80%, según la categoría de producto. Actualmente, la compañía opera con una cartera de 780 productos activos y atiende a 550 clientes, entre los que se encuentran mayoristas, cadenas de supermercados y distribuidoras. Cabe destacar que solo 13 clientes concentran el 51% de las ventas totales.

El objetivo de este laboratorio es desarrollar un modelo de predicción de ventas para cada producto, estimando el volumen que se venderá dentro de dos meses a partir de la fecha de corte (12/2019). Es importante considerar que no se deben predecir productos en etapa de lanzamiento (menos de tres meses de ventas) ni aquellos en proceso de discontinuación (últimos tres meses antes de salir del mercado).

El contexto económico también juega un papel relevante: en 2018, Argentina atravesó una crisis que impactó negativamente en las ventas, lo que debe ser tenido en cuenta en el análisis.

La métrica de evaluación será el error absoluto medio, calculado como el cociente entre la suma de los errores absolutos y la suma de las ventas reales en el mes objetivo (mes + 2).

En resumen, el desafío consiste en estimar con precisión las ventas futuras de cada producto, considerando la dinámica del mercado, la concentración de clientes y las particularidades del portafolio, para apoyar la toma de decisiones comerciales y operativas de la empresa.

2. Hipótesis Experimental

Hipótesis: Se partió de la idea de evaluar si, mediante modelos de gradient boosting con ingeniería de variables y ajuste de hiperparámetros, era posible encontrar un enfoque que mejorara la precisión del método de referencia basado en el promedio, incrementando así la capacidad de predicción en un escenario con múltiples series temporales.

3. Diseño Experimental y Materiales

3.1 Granularidades Evaluadas

El diseño experimental se estructura en tres niveles de granularidad:

- **Granularidad por producto:** Predicción individual para cada uno de los 780 productos activos
- **Granularidad temporal:** Predicción a 2 meses vista desde diciembre 2019

3.2 Arquitectura del Experimento

Dataset Experimental:

- Registros con target válido: 20.789
- Número de features finales: 62
- Productos incluidos: 780
- Períodos de datos: 36 meses (2017-01 a 2019-12)

División Temporal:

- **Entrenamiento:** hasta agosto 2019 (18.458 registros, 756 productos únicos)
- **Validación:** desde agosto 2019 (2.331 registros, 780 productos únicos)

3.3 Materiales y Herramientas

Algoritmos Implementados:

1. **LightGBM con linear_tree=True:** Optimizado con 700 trials usando Optuna
2. **Regresión Lineal:** Implementación clásica con 11 lags mensuales
3. **Media Móvil Ajustada:** Promedio de 12 meses con factor de corrección 0.93

Feature Engineering Sistemático:

- 36 lags mensuales de ventas (tn_lag_1 a tn_lag_36)
- Lags de clientes únicos (num_customers_lag_1 a num_customers_lag_3)
- Lags de stock (stock_tn_mean_lag_1 a stock_tn_mean_lag_3)
- Features temporales (mes, trimestre, año)
- Features rolling (ventanas de 3 y 6 meses)
- Variables de contexto del producto (cant_clientes_unicos, meses_vida)

Estrategia de Robustez:

- Entrenamiento con 20 semillas diferentes para LightGBM
- Validación cruzada temporal estricta
- Optimización de hiperparámetros con búsqueda exhaustiva

3.4 Scripts y Reproducibilidad

Repositorio: https://github.com/paolasz/LaboIII-TPFinal.git

Los scripts implementados:

- Reproducibilidad completa mediante semillas fijas
- Procesamiento automatizado de los 780 productos
- Validación de consistencia entre modelos
- Exportación estandarizada de resultados

4. Procedimientos y Recopilación de Resultados

4.1 Ejecución del Experimento

Parámetros Iniciales LightGBM:

- lambda_l1: 0.8638900372842315
- lambda_l2: 0.5824582803960838
- num_leaves: 81
- feature_fraction: 0.6557975442608167
- learning_rate: 0.022059149678071027
- bagging_fraction: 0.7465447373174767
- bagging_freq: 5
- min_child_samples: 33
- max_bin: 160

Procedimiento de Ensemble:

1. Carga de predicciones individuales de los tres modelos
2. Verificación de consistencia en product_id entre modelos
3. Merge de salidas generando columnas: tn_lgbm, tn_regresion, tn_media
4. Cálculo de predicción final mediante promedio simple: tn = (tn_lgbm + tn_regresion + tn_media) / 3

4.2 Resultados Capturados

Public Score del Modelo Ensemble: 0.249

Métricas de Entrenamiento:

- Métrica de optimización: MAE (Mean Absolute Error)
- Estrategia de múltiples semillas para reducir varianza
- 20 entrenamientos independientes con promediado final

5. Análisis de los Resultados

5.1 Interpretación del Rendimiento

El score público de 0.249 representa una mejora significativa respecto a métodos baseline tradicionales (en el orden de los 0.27-0.34). Este resultado confirma la efectividad del enfoque ensemble propuesto.

Fortalezas del Modelo:

- **Robustez:** El uso de 20 semillas diferentes en LightGBM reduce la varianza y mejora la estabilidad
- **Complementariedad:** La combinación de gradient boosting, regresión lineal y media móvil captura diferentes aspectos de los patrones temporales
- **Feature Engineering Exhaustivo:** Los 62 features derivados capturan patrones estacionales, tendencias y contexto del producto

Análisis por Componente:

- **LightGBM:** Captura patrones no lineales complejos y interacciones entre variables
- **Regresión Lineal:** Proporciona estabilidad y interpretabilidad, especialmente efectiva en productos con patrones lineales
- **Media Móvil Ajustada:** Aporta robustez como baseline y suaviza predicciones extremas

5.2 Validación de la Hipótesis

Los resultados experimentales validan parcialmente la hipótesis original. El ensemble efectivamente supera métodos baseline simples, demostrando que la combinación de gradient boosting optimizado con otros predictores mejora la precisión predictiva.

6. Resultados, Conclusiones y Discusión

6.1 Conclusión Principal

La hipótesis experimental se cumple: El enfoque ensemble basado en gradient boosting con ingeniería de variables optimizada logra mejorar significativamente la precisión predictiva comparado con métodos de referencia basados únicamente en promedios históricos.

6.2 Limitaciones y Consideraciones

Limitaciones Identificadas:

- Dependencia de datos históricos de 36 meses que pueden no capturar cambios estructurales futuros
- Sensibilidad a crisis económicas (como la de 2018) que pueden requerir ajustes adicionales
- Concentración en top 13 clientes que podría requerir modelos específicos por segmento

Robustez del Diseño:

- La estrategia de múltiples semillas mitiga la variabilidad del modelo
- El ensemble reduce el riesgo de sobreajuste de modelos individuales

6.3 Propuestas de Experimentos Futuros

Extensiones Metodológicas:

1. **Ensemble Dinámico:** Implementar pesos adaptativos por producto según rendimiento histórico de cada componente
2. **Segmentación Avanzada:** Desarrollar modelos específicos para los 13 clientes principales vs. clientes menores
3. **Features Macroeconómicas:** Incorporar indicadores externos (inflación, tipo de cambio) para mejorar robustez ante crisis

Validación Adicional:

1. **Backtesting Extendido:** Evaluar performance en múltiples períodos de validación
2. **Análisis de Estacionalidad:** Estudiar rendimiento diferencial por temporadas y productos

3. **Análisis de Sensibilidad:** Evaluar impacto de cada componente del ensemble

6.4 Respuestas a Cuestionamientos Potenciales

"¿Por qué usar ensemble en lugar del mejor modelo individual?" Nuestros resultados confirman que los ensemble reducen varianza y mejoran robustez. Aunque LightGBM podría ser el mejor modelo individual, el ensemble proporciona mayor estabilidad predictiva, especialmente importante en contextos comerciales donde la consistencia es crucial.

"¿Es suficiente el feature engineering implementado?" Los 62 features derivados capturan múltiples dimensiones temporales (1-36 meses), contextuales y de stock. Sin embargo, futuras extensiones podrían incluir features macroeconómicas y de competencia para mejorar la robustez ante shocks externos.

"¿Cómo se garantiza la reproducibilidad?" El diseño experimental incluye semillas fijas, scripts automatizados

6.5 GitHub

Repositorio: https://github.com/paolasz/LaboIII-TPFinal.git

Anexo Técnico: Especificaciones del Modelo

Modelo 1: LightGBM Optimizado

Configuración Final:

Se desarrolló un modelo de predicción de ventas (en toneladas) utilizando LightGBM con linear_tree=True.
El enfoque estuvo orientado a granularidad por producto y a predecir el valor de ventas dos meses a futuro (target).

Parámetros optimizados con 700 trials Optuna
Entrenamiento con 20 semillas independientes
Target: tn (ventas en toneladas) a +2 meses
Features: 62 variables derivadas

Validación temporal estricta

Modelo 2: Regresión Lineal

Configuración:

Granularidad: por product_id
Features: 11 lags mensuales de tn
Dataset reducido: 33 productos con historial completo

```
magicos = [20002, 20003, 20006, 20010, 20011, 20018, 20019, 20021,
20026, 20028, 20035, 20039, 20042, 20044, 20045, 20046, 20049, 20051,
20052, 20053, 20055, 20008, 20001, 20017, 20086, 20180, 20193, 20320,
20532, 20612, 20637, 20807, 20838]
```

Matriz: 33 filas × 13 columnas
Target: tn en periodo +2

Modelo 3: Media 12 Meses Ajustada

Configuración:

Base: promedio simple últimos 12 meses (enero-diciembre 2019)
Factor de ajuste: 0.93
Tratamiento: valor cero para productos sin historial
Output: predicción conservadora y estable

Ensemble Final

Metodología:

Predicción final = (tn_lgbm + tn_regresion + tn_media) / 3
Verificación de consistencia en product_id
Score público resultante: 0.249