

RESUMEN EJECUTIVO: PRUEBA TÉCNICA

INFORMACIÓN DEL PROYECTO

- **Fecha:** Agosto 2025
- **Autor:** David Rodríguez Arrauth
- **Tipo de Análisis:** Modelado predictivo de demanda de productos
- **Metodología:** Machine Learning y Series Temporales
- **Alcance:** 135 productos, 3 años de datos históricos (2020-2023)

RESUMEN EJECUTIVO

El proyecto desarrolló un sistema integral de pronóstico de demanda para optimizar la gestión de inventarios y planificación comercial. Se implementaron múltiples metodologías avanzadas de machine learning y análisis de series temporales para predecir la demanda de 135 productos, considerando el impacto competitivo y patrones estacionales.

RESULTADOS CLAVE























- **Precisión del Modelo:** Se logró una precisión promedio del 92% en las predicciones con los mejores modelos seleccionados por producto.
- **Impacto del Competidor:** La entrada del competidor (julio 2021) redujo las ventas en un 20.03% promedio, con variaciones significativas por categoría.
- **Tendencia de Largo Plazo:** Sin considerar el impacto competitivo, la demanda se mantiene estable (crecimiento no significativo estadísticamente).

Modelos Óptimos: Se seleccionó automáticamente el mejor modelo para cada producto:

- ARIMA/SARIMA: 45% de productos (mejores para series con patrones estacionales)
- Ridge Regression: 30% de productos (efectiva para productos con variables exógenas)
- LSTM: 15% de productos (superior en series complejas no lineales)
- Lasso Regression: 10% de productos (óptima para productos con pocas variables relevantes)

ESTRUCTURA DEL PROYECTO

Prueba/

- | —  Data/ # Datos del proyecto
 - | | — demanda.csv #  Dataset principal de demanda
 - | | — catalogo_productos.csv #  Catálogo de productos
 - | | — demanda_test.csv #  Dataset de prueba (out-of-sample)
 - | | — data_clean.csv #  Datos limpios procesados
 - | | — df_global.csv #  Dataset global unificado
 - | | — exog_variables_complete.csv #  Variables exógenas completas
 - | | — outlier_dummies.csv #  Variables dummy para outliers
 - | | — ts_log.csv #  Series de tiempo transformadas
- | —  src/ # Código fuente modularizado
 - | | — EDA_tools.py #  Herramientas de análisis exploratorio
 - | | — econometric_tools.py #  Herramientas econométricas
 - | | — models_data_preprocessing.py #  Preprocesamiento de datos
 - | | — arima_model_tools.py #  Modelos ARIMA
 - | | — linear_model_tools.py #  Modelos lineales (Ridge/Lasso)
 - | | — lstm_model_tools.py #  Modelos LSTM (Deep Learning)
 - | | — prophet_model_tools.py #  Modelos Prophet (Facebook)
 - | | — model_comparison_tools.py #  Comparación de modelos
 - | | — predict_tools.py #  Herramientas de predicción
- | —  Modelos registrados/ # Modelos entrenados y guardados
 - | | — best_model_producto_1_arima.pkl #  Mejor modelo ARIMA producto 1
 - | | — best_model_producto_1_lasso.pkl #  Mejor modelo Lasso producto 1

```
| |— best_model_producto_1_lstm.pkl    # 🏆 Mejor modelo LSTM producto 1
| |— best_model_producto_1_ridge.pkl  # 🏆 Mejor modelo Ridge producto 1
| |— ... (para todos los productos)
|
|— 📁 data_by_product/                # Datos organizados por producto
| |— train/                          # 💠 Datos de entrenamiento por producto
| | |— producto_1_train.csv
| | |— producto_10_train.csv
| | |— ...
| |— test/                           # 💠 Datos de prueba por producto
| | |— producto_1_test.csv
| | |— producto_10_test.csv
| | |— ...
|
|— 📊 output/                        # Resultados y reportes
| |— model_comparison_results.csv    # 📋 Comparación de modelos
| |— best_models_summary.csv        # 🏆 Resumen de mejores modelos
| |— predictions_report.csv         # 🎯 Reporte de predicciones
| |— visualizations/                # 📈 Gráficos y visualizaciones
|
|— 🛠️ env/                           # Ambiente virtual de Python
|— 📓 Desarrollo prueba.ipynb       # 🚀 NOTEBOOK PRINCIPAL
|— 🏆 Best_model_selection.py        # Script de selección del mejor modelo
|— 📋 requirements.txt               # Lista de dependencias
```

└─ 📖 Instrucciones de instalación.txt

...

Nota: Las instrucciones para instalar el ambiente están en "Instrucciones de instalación.txt".

INSIGHTS ESTRATÉGICOS PRINCIPALES

1. ANÁLISIS DEL IMPACTO COMPETITIVO

Hallazgos Críticos:

- Impacto Heterogéneo: No todos los productos fueron afectados igual
- Productos Resilientes: 23 productos (17%) incrementaron ventas post-competidor
- Categorías Más Afectadas: Productos lácteos (-52%), carnes (-35%), panaderías (-30%)
- Categorías Resilientes: Productos enlatados (+0.6%), alimentos envasados (+0.6%)

Implicaciones de Negocio:

- Necesidad de estrategias diferenciadas por categoría
- Oportunidad de reforzar categorías resilientes
- Priorizar innovación en categorías más vulnerables

2. PATRONES ESTACIONALES IDENTIFICADOS

Descubrimientos Clave:

- Estacionalidad Semanal: Patrones claros cada 7 días
- Ciclos Anuales: Picos en junio-julio, valle en febrero-marzo
- Volatilidad Reducida: Post-competidor, menor variabilidad pero niveles más bajos
- Productos Discontinuos: 13 productos con ventas esporádicas

3. SEGMENTACIÓN ESTRATÉGICA DE PRODUCTOS

Clasificación por Comportamiento:

- Productos Premium: Menor volumen pero mayor resistencia competitiva
- Productos Estacionales: Requieren gestión de inventario específica
- Marcas Exclusivas: Mantuvieron mejor desempeño post-competidor
- Productos por Tamaño: Diferencias marginales, no factor determinante

METODOLOGÍA Y DECISIONES TÉCNICAS

1. PROCESO DE LIMPIEZA Y PREPARACIÓN

Decisiones Tomadas:

- Datos Faltantes: Imputación estratégica basada en análisis de patrones
- Variables Eliminadas: Subcategoría (23% faltantes) y NIT proveedor (irrelevante)
- Encoding: Aplicado a variables categóricas para modelado
- Transformaciones: Logaritmos para reducir volatilidad

Justificación: Maximizar calidad de datos manteniendo representatividad

2. ANÁLISIS EXPLORATORIO INTEGRAL

Enfoques Implementados:

- Análisis Univariado: Distribuciones y frecuencias de cada variable
- Análisis Bivariado: Relaciones entre variables predictoras y demanda
- Análisis Multivariado: Interacciones complejas y correlaciones
- Series Temporales: Tendencias, estacionalidad y autocorrelación

Valor Agregado: Comprensión profunda del comportamiento de datos para mejor modelado

3. MODELADO ECONOMETRICO PARA MEDIR IMPACTO DE COMPETIDOR Y ANÁLISIS DE TENDENCIA

ANÁLISIS DEL IMPACTO DEL COMPETIDOR

El análisis del impacto del competidor se realizó mediante modelos SARIMA con variables exógenas, siguiendo una metodología rigurosa de econometría de series temporales.

3.1 IDENTIFICACIÓN DEL EVENTO EXÓGENO

- **Fecha del Shock:** 2 de julio de 2021 (entrada del competidor)
- **Variable Dummy:** Entrada_competidor (0: pre-competidor, 1: post-competidor)
- **Justificación:** Evento claramente identificable que permite separar los efectos

3.2 ESPECIFICACIÓN DEL MODELO BASE

Modelo SARIMA(1,1,1)(1,0,1)[7] con variable exógena + errores estándares robustos:

$$\log(\text{demanda}_t) = \phi_1 \log(\text{demanda}_{t-1}) + \theta_1 \varepsilon_{t-1} + \phi_7 \log(\text{demanda}_{t-7}) + \theta_7 \varepsilon_{t-7} + \beta * \text{Entrada_competidor}_t + \varepsilon_t$$

3.3 DECISIONES TÉCNICAS CLAVE

A) Transformación Logarítmica

- **Decisión:** Aplicar $\log(1+x)$ a la serie de demanda
- **Justificación:** Reducir heterocedasticidad y permitir interpretación de coeficientes como elasticidades
- **Resultado:** Mejora significativa en la distribución de residuales

B) Especificación SARIMA(1,1,1)(1,0,1)[7]

- AR(1): Captura la persistencia temporal (coeficiente = 0.88)
- I(1): Una diferenciación para lograr estacionariedad
- MA(1): Corrige autocorrelación residual
- Componente Estacional [7]: Refleja patrones semanales identificados en ACF/PACF
- SAR(1): Autorregresivo estacional para patrones recurrentes
- SMA(1): Media móvil estacional para suavizar fluctuaciones

C) Errores Estándares Robustos

- **Problema:** Residuales no pasaron pruebas de normalidad y homoscedasticidad
- **Solución:** HAC (Heteroscedasticity and Autocorrelation Consistent) standard errors
- **Implicación:** Inferencia estadística válida pese a violaciones de supuestos clásicos

3.4 RESULTADOS DEL ANÁLISIS DE IMPACTO

- **Coefficiente del Competidor:** $\beta = -0.1603$ (no significativo con errores robustos)
- **Interpretación:** Reducción estimada del 16% en la demanda promedio post-competidor
- **Intervalo de Confianza:** [-0.35, 0.03] al 95%
- **Implicación:** Existe incertidumbre estadística sobre la magnitud exacta del efecto

3.5 LIMITACIONES RECONOCIDAS

- Residuales no cumplen completamente supuestos de ruido blanco
- Posible confusión con otras variables no observadas (estacionalidad, tendencias macroeconómicas)
- Necesidad de modelado más sofisticado de la estructura de varianza

ANÁLISIS DE TENDENCIA DE LARGO PLAZO

3.6 MODELO CON TENDENCIA DETERMINÍSTICA CONSTANTE

Especificación: SARIMA(1,0,1)(1,0,1)[7] + intercepto

$$\log(\text{demanda}_t) = \mu + \varphi_1 \log(\text{demanda}_{t-1}) + \dots + \varepsilon_t$$

Resultados:

- Coeficiente de tendencia (μ): 0.0003 (0.03% crecimiento diario)
- Significancia estadística: No significativo ($p > 0.10$)
- Interpretación: No evidencia de crecimiento constante en el tiempo

3.7 MODELO CON TENDENCIA DETERMINÍSTICA LINEAL

Especificación: SARIMA(1,0,1)(1,0,1)[7] + tendencia lineal

$$\log(\text{demanda}_t) = \alpha + \beta \text{tiempo}_t + \varphi_1 \log(\text{demanda}_{t-1}) + \dots + \varepsilon_t$$

Resultados:

- Coeficiente de tendencia lineal (β): -0.0003 (-0.03% por período)
- Significancia estadística: No significativo ($p > 0.13$)
- Interpretación: Ausencia de tendencia lineal determinística

3.8 DECISIONES TÉCNICAS EN EL ANÁLISIS DE TENDENCIA

A) Especificación de la Tendencia

- **¿Por qué No Usar Diferenciación para Tendencia?**
 - Modelo Alternativo: SARIMA(1,1,1) implicaría tendencia estocástica
 - Decisión: Usar modelos en niveles con tendencia determinística
 - Justificación: Mayor control sobre la especificación de la tendencia y mejor interpretabilidad económica

B) Tratamiento de la Heterocedasticidad

- **¿Por qué Errores Robustos?**
 - Problema: Persistencia de heterocedasticidad en residuales
 - Solución: Mantener errores HAC para inferencia válida
 - Trade-off: Menor potencia estadística pero mayor confiabilidad en conclusiones

C) Validación de Estacionariedad

- Test ADF aplicado: Serie original ya es estacionaria ($p < 0.01$)
- Implicación: No requiere diferenciación obligatoria
- Ventaja: Permite modelar directamente los niveles de demanda

INTERPRETACIÓN ECONÓMICA FINAL

SOBRE EL IMPACTO DEL COMPETIDOR

- **Magnitud:** Reducción estimada del 16%, pero con alta incertidumbre estadística
- **Heterogeneidad:** Impacto variable por categoría (desde +0.6% hasta -52%)
- **Persistencia:** Efecto permanente en el nivel, no en la tasa de crecimiento
- **Recomendación:** Monitoreo continuo para detectar efectos dinámicos

SOBRE LA TENDENCIA INTRÍNSECA

- **Conclusión Principal:** Demanda se mantiene estable en ausencia de shocks externos
- **Implicación Comercial:** Crecimiento futuro dependerá de factores exógenos (nuevos productos, expansión geográfica, cambios demográficos)

ROBUSTEZ DE LOS HALLAZGOS

- Consistencia entre especificaciones alternativas (constante vs. lineal)
- Validación mediante múltiples pruebas de raíces unitarias
- Coherencia con análisis exploratorio descriptivo previo

4. MODELADO PREDICTIVO AVANZADO

Modelos Implementados:

4.1 Modelos de Regresión (Ridge/Lasso):

- Optimización bayesiana de hiperparámetros
- Regularización para prevenir overfitting
- Incorporación de variables exógenas

4.2 Modelos ARIMA/SARIMA:

- Modelado de patrones temporales complejos
- Componentes estacionales semanales
- Pruebas de estacionariedad rigurosas

4.3 Redes Neuronales LSTM:

- Captura de patrones no lineales
- Memoria de largo plazo para series complejas
- Secuencias de 15 períodos

4.4 Prophet (Facebook):

- Manejo automático de estacionalidad
- Robustez ante datos faltantes
- Interpretabilidad de componentes

Criterio de Selección: MAPE (Mean Absolute Percentage Error) por producto

5. VALIDACIÓN Y SELECCIÓN AUTOMATIZADA

Proceso Implementado:

- **Validación Cruzada:** Separación temporal train/test 80/20
- **Métricas Múltiples:** MAPE, MAE, RMSE para evaluación integral
- **Selección Automática:** Mejor modelo por producto basado en performance
- **Validación Final:** Predicción en datos completamente fuera de muestra