Métodos de Optimización en Regresión

Universidad Nacional del Altiplano Puno
Facultad de Ingeniería Estadística e Informática
Presentado por: MARIO WILFREDO RAMIREZ PUMA
Curso: Métodos de Optimización
Docente: Ing. TORRES CRUZ FRED

28 de mayo de 2025

Resumen

Este trabajo presenta un análisis comparativo de métodos de optimización aplicados a problemas de regresión, utilizando un dataset de empresas agroindustriales de la región Ica, Perú. Se evaluaron tres enfoques: Regresión Lineal Múltiple como baseline, Random Forest estándar, y Random Forest optimizado con Optuna tras aplicar transformación logarítmica. Los resultados demuestran que la transformación de datos y la optimización automática de hiperparámetros son factores críticos para el éxito en datasets con distribuciones extremas. El modelo final logró un coeficiente de determinación (R^2) de 0.9987, representando una mejora sustancial respecto a los métodos tradicionales que obtuvieron valores negativos de R^2.

Palabras clave: Métodos de optimización, Random Forest, Optuna, transformación logarítmica, regresión, machine learning.

Índice

1. Introducción

Los métodos de optimización constituyen una rama fundamental de las matemáticas aplicadas y la ciencia de datos, con aplicaciones críticas en la resolución de problemas complejos del mundo real. En el contexto de machine learning, la optimización se manifiesta en múltiples niveles: desde la minimización de funciones de pérdida durante el entrenamiento de modelos hasta la búsqueda de hiperparámetros óptimos que maximicen el rendimiento predictivo.

El presente trabajo aborda un problema específico de regresión utilizando datos reales de empresas agroindustriales de la región Ica, Perú. Este dataset presenta características desafiantes típicas de datos empresariales reales: distribuciones extremadamente asimétricas, presencia de outliers significativos, y alta variabilidad en la variable objetivo.

1.1. Problemática

Los datos empresariales frecuentemente exhiben distribuciones que violan los supuestos fundamentales de los métodos estadísticos tradicionales. En particular, cuando una pequeña proporción de observaciones domina la variabilidad total del dataset, los algoritmos de regresión convencionales pueden fallar completamente, produciendo modelos con capacidad predictiva nula o negativa.

1.2. Objetivos

Objetivo General: Evaluar y comparar la efectividad de diferentes métodos de optimización aplicados a un problema de regresión con datos empresariales reales.

Objetivos Específicos:

1. Analizar las características del dataset y identificar problemas inherentes en los datos.
2. Implementar y evaluar métodos de regresión tradicionales como baseline.
3. Aplicar técnicas de transformación de datos para abordar problemas de distribución.
4. Utilizar algoritmos de ensemble (Random Forest) con optimización automática de hiperparámetros.
5. Comparar el rendimiento de diferentes enfoques metodológicos.
6. Extraer insights empresariales relevantes del análisis de importancia de variables.

2. Marco Teórico

2.1. Regresión Lineal Múltiple

La regresión lineal múltiple constituye el método fundamental para modelar relaciones lineales entre una variable dependiente y y múltiples variables independientes $x_1, x_2, \ldots, x_p$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{1}$$

donde $\beta_0, \beta_1, \ldots, \beta_p$ son los coeficientes del modelo y ε representa el término de error. El método de Mínimos Cuadrados Ordinarios (MCO) estima los parámetros minimizando la suma de cuadrados de los residuos:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \tag{2}$$

2.2. Random Forest

Random Forest [1] es un algoritmo de ensemble que combina múltiples árboles de decisión mediante bootstrap aggregating (bagging). El algoritmo construye B árboles de decisión, cada uno entrenado en una muestra bootstrap del dataset original, y para cada división considera únicamente un subconjunto aleatorio de m variables.

La predicción final se obtiene promediando las predicciones individuales:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\boldsymbol{x}) \tag{3}$$

donde $T_b(\boldsymbol{x})$ representa la predicción del b-ésimo árbol.

2.3. Optimización Bayesiana con Optuna

Optuna [2] implementa optimización bayesiana utilizando Tree-structured Parzen Estimator (TPE) para la búsqueda eficiente de hiperparámetros. El algoritmo construye modelos probabilísticos de la función objetivo y utiliza estos modelos para sugerir configuraciones prometedoras de hiperparámetros.

2.4. Transformación Logarítmica

Para variables con distribuciones altamente asimétricas o con outliers extremos, la transformación logarítmica puede estabilizar la varianza y aproximar la distribución a la normalidad:

$$y^* = \log(y) \tag{4}$$

Esta transformación es particularmente útil cuando los datos exhiben crecimiento exponencial o cuando la variabilidad es proporcional al nivel de la variable.

3. Metodología

3.1. Descripción del Dataset

El dataset utilizado contiene información de 494 empresas agroindustriales de la región Ica, Perú, correspondiente al año 2023. Los datos fueron obtenidos del portal de Datos Abiertos del Gobierno Peruano y provienen del Ministerio de la Producción (PRODUCE) en colaboración con SUNAT.

Variables del dataset:

- **Variable objetivo:** `valor_estimado_maximo_venta` (ventas máximas estimadas en soles)
- **Variables predictoras:**
  - `ciiu`: Código de Clasificación Industrial Internacional Uniforme
  - `provincia`: Provincia donde se ubica la empresa
  - `distrito`: Distrito específico de ubicación
  - `descciiu`: Descripción detallada de la actividad económica
  - `tamanio_emp`: Categoría empresarial (micro, pequeña, mediana, gran empresa)
  - `exporta`: Indicador binario de actividad exportadora
  - `valor_estimado_minimo_venta`: Ventas mínimas estimadas

3.2. Análisis Exploratorio

El análisis exploratorio inicial reveló características problemáticas en la distribución de la variable objetivo:

Cuadro 1: Estadísticas descriptivas de la variable objetivo

Estadística	**Valor (S/)**
Media	29,340,774
Mediana	742,500
Mínimo	742,500
Máximo	13,365,000,000
Desviación Estándar	601,274,319
Coeficiente de Variación	2,049.3 %

La distribución mostró una asimetría extrema, con 457 empresas (92.5 %) presentando el mismo valor de ventas máximas (S/ 742,500), mientras que una sola empresa registró ventas de S/ 13,365,000,000, representando un outlier de magnitud 18,000 veces superior al valor modal.

3.3. Preprocesamiento de Datos

Codificación de Variables Categóricas: Se aplicó Label Encoding a las variables categóricas:

- `provincia_encoded`: 5 categorías únicas
- `distrito_encoded`: 34 categorías únicas
- `descciiu_encoded`: 14 categorías únicas
- `tamanio_emp_encoded`: 4 categorías únicas

Escalado de Variables (solo Regresión Lineal): Para el modelo de regresión lineal se aplicó StandardScaler para normalizar las variables numéricas, ya que este algoritmo es sensible a la escala de las variables. Random Forest no requiere escalado debido a su naturaleza basada en árboles de decisión.

Variable Binaria: La variable `exporta` se convirtió a formato binario (0/1), donde 22 empresas (4.5 %) reportaron actividad exportadora.

División del Dataset: Se aplicó una división aleatoria estratificada 80/20 para entrenamiento y prueba, resultando en 395 muestras para entrenamiento y 99 para evaluación.

3.4. Modelos Implementados

3.4.1. Modelo 1: Regresión Lineal Múltiple (Baseline)

Se implementó regresión lineal múltiple con standardización de variables mediante StandardScaler. Este modelo sirve como baseline para comparación.

3.4.2. Modelo 2: Random Forest Estándar

Se aplicó Random Forest con parámetros por defecto para evaluar el rendimiento sin optimización de hiperparámetros.

3.4.3. Modelo 3: Random Forest con Transformación Logarítmica y Optuna

Dado el fracaso de los modelos anteriores, se implementó:

1. **Transformación logarítmica** de la variable objetivo
2. **Random Forest** como algoritmo base
3. **Optuna** para optimización automática de hiperparámetros

Hiperparámetros optimizados:

- `n_estimators`: [50, 300]

- `max_depth`: [5, 20]
- `min_samples_split`: [2, 15]
- `min_samples_leaf`: [1, 8]
- `max_features`: ['sqrt', 'log2']
- `bootstrap`: [True, False]

Función objetivo: Se minimizó el RMSE mediante validación cruzada 3-fold:

$$\text{Objetivo} = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{1}{n_k}\sum_{i\in S_k}(y_i - \hat{y}_i)^2} \tag{5}$$

donde $K = 3$ es el número de folds y S_k representa el conjunto de prueba en el fold k.

4. Resultados

4.1. Comparación de Modelos

Cuadro 2: Comparación de rendimiento de modelos

Modelo	R^2 **Train**	R^2 **Test**	**RMSE Test (S/)**	**Viable**
Regresión Lineal	0.8980	-16,737.21	195,455,534	No
Random Forest Estándar	0.0632	-1,605.52	60,553,142	No
RF + Log + Optuna	0.9763	0.9987	54,013	Sí

4.2. Impacto de la Transformación Logarítmica

La transformación logarítmica produjo una mejora dramática en la tratabilidad de los datos:

Cuadro 3: Efecto de la transformación logarítmica

Métrica	**Datos Originales**	**Datos Log-Transformados**
Coeficiente de Variación	2,049.3 %	6.6 %
Rango	S/ 13,364,257,500	9.80
Desviación Estándar	S/ 601,274,319	0.91
Mejora en CV	**-**	**310.1x**

4.3. Optimización con Optuna

El proceso de optimización exploró 50 configuraciones diferentes de hiperparámetros en aproximadamente 3 minutos. Los hiperparámetros óptimos encontrados fueron:

Cuadro 4: Hiperparámetros óptimos encontrados por Optuna

Hiperparámetro	**Valor Óptimo**
`n_estimators`	150
`max_depth`	16
`min_samples_split`	16
`min_samples_leaf`	10
`max_features`	sqrt
`bootstrap`	True

4.4. Análisis de Importancia de Variables

El modelo final reveló la importancia relativa de cada variable predictora:

Cuadro 5: Importancia de variables en el modelo final

Variable	**Importancia**	**Porcentaje**
`valor_estimado_minimo_venta`	0.537	53.7 %
`tamanio_emp_encoded`	0.288	28.8 %
`exporta_encoded`	0.113	11.3 %
`ciiu`	0.035	3.5 %
`descciiu_encoded`	0.018	1.8 %
`distrito_encoded`	0.007	0.7 %
`provincia_encoded`	0.002	0.2 %

4.5. Métricas de Evaluación Final

El modelo optimizado final alcanzó las siguientes métricas de rendimiento:

- R^2 **en conjunto de prueba:** 0.9987 (99.87 %)
- **RMSE en escala original:** S/ 54,013
- **MAE en escala original:** S/ 20,579
- **Tiempo de entrenamiento:** 3 minutos

5. Discusión

5.1. Análisis de Resultados

Los resultados obtenidos demuestran de manera contundente la importancia crítica del preprocesamiento de datos en problemas de optimización aplicados. La mejora en el coeficiente de determinación desde valores negativos extremos (-16,737) hasta 0.9987 representa un cambio cualitativo fundamental en la viabilidad del modelo.

Fracaso de Métodos Tradicionales: La regresión lineal múltiple falló completamente debido a la violación severa de sus supuestos fundamentales. La presencia de un outlier 18,000 veces mayor que el valor modal generó una distribución que no puede ser modelada efectivamente mediante relaciones lineales simples.

Limitaciones de Random Forest sin Transformación: Aunque Random Forest es conocido por su robustez ante outliers, incluso este algoritmo falló cuando se enfrentó a la distribución extrema sin preprocesamiento. Esto subraya que ningún algoritmo es completamente inmune a problemas de calidad de datos.

Efectividad de la Transformación Logarítmica: La transformación logarítmica redujo el coeficiente de variación de 2,049 % a 6.6 %, una mejora de 310 veces. Esta transformación permitió que los algoritmos de machine learning identificaran patrones genuinos en los datos en lugar de ser dominados por el outlier extremo.

5.2. Importancia de Variables

El análisis de importancia de variables revela patrones empresariales coherentes:

1. **Ventas mínimas como predictor principal (53.7 %):** Este resultado es empresarialmente lógico, ya que existe una correlación natural entre los rangos mínimo y máximo de ventas de una empresa.
2. **Tamaño empresarial (28.8 %):** La categorización oficial del tamaño empresarial captura efectivamente la capacidad operativa y el potencial de ventas.
3. **Actividad exportadora (11.3 %):** Las empresas que exportan tienden a ser más sofisticadas y tener mayores volúmenes de ventas.
4. **Ubicación geográfica (impacto mínimo):** La provincia y distrito tienen impacto negligible, sugiriendo que factores intrínsecos de la empresa son más determinantes que la ubicación.

6. Conclusiones

6.1. Conclusiones Principales

1. **La transformación de datos es crítica:** La transformación logarítmica fue el factor determinante para el éxito del proyecto, demostrando que el preprocesamiento adecuado puede resolver problemas aparentemente intratables.

2. **La optimización automática aporta valor significativo:** Optuna identificó una configuración de hiperparámetros que resultó en un modelo prácticamente perfecto ($R^2 = 99.87\,\%$).

3. **Random Forest supera métodos tradicionales en datos complejos:** Cuando se combina con preprocesamiento adecuado, Random Forest demostró capacidades superiores para manejar relaciones no lineales complejas.

4. **Los insights empresariales son consistentes:** Las variables más importantes identificadas por el modelo tienen interpretaciones empresariales claras y lógicas.

7. Referencias

Referencias

[1] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[2] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623-2631.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12, 2825-2830.

A. Código Principal

A.1. Implementación de Random Forest con Optuna

Listing 1: Función objetivo para optimización con Optuna

```
def objective_log(trial, X_train, y_train):
    """Funci n objetivo optimizada para datos log-transformados"""

    params = {
        'n_estimators': trial.suggest_int('n_estimators', 50, 300, step
            =25),
        'max_depth': trial.suggest_int('max_depth', 5, 20),
        'min_samples_split': trial.suggest_int('min_samples_split', 2,
            15),
        'min_samples_leaf': trial.suggest_int('min_samples_leaf', 1, 8),
        'max_features': trial.suggest_categorical('max_features',
                                                  ['sqrt', 'log2']),
```

```
        'bootstrap': trial.suggest_categorical('bootstrap', [True, False
            ]),
        'random_state': 42
    }

    model = RandomForestRegressor(**params)

    # Cross-validation 3-fold
    cv_scores = cross_val_score(
        model, X_train, y_train,
        cv=3,
        scoring='neg_root_mean_squared_error'
    )

    return -cv_scores.mean()
```

A.2. Transformación Logarítmica

Listing 2: Aplicación de transformación logarítmica

```
# Aplicar transformaci n logar tmica
df['log_ventas_maximas'] = np.log(df['valor_estimado_maximo_venta'])

# Verificar mejora en variabilidad
cv_original = (original_target.std()/original_target.mean())*100
cv_log = (log_target.std()/log_target.mean())*100
mejora = cv_original / cv_log

print(f"CV␣Original:␣{cv_original:.1f}%")
print(f"CV␣Log:␣{cv_log:.1f}%")
print(f"Mejora:␣{mejora:.1f}x␣mejor")
```