Universidad de los Andes

VIVIAN CABANZO FERNÁNDEZ
LAURA DANIELA DIAZ TORRES
CRISTIAN FELIPE MUÑOZ GUERRERO
ZENNETH OLIVERO TAPIA

Problem Set 3 – Making Money with ML?

BIG DATA Y MACHINE LEARNING PARA ECONOMIA APLICADA

2025-02

best_equipo_08

MEJOR MODELO RANDOM FOREST CV ESPACIAL

Universidad de los Andes

Razonamiento detrás del ensamble Random Forest con CV espacial

- **Datos usados:**

$$P_i = f\left(X_i^{estructura}, X_i^{amenidades}, X_i^{distancia}, X_i^{locación}, X_i^{tiempo}, X_i^{tipo}\right) + u_i$$

La función f captura no linealidades e interacciones entre estructura, amenidades, distancias y ubicación.

- **Complementariedad estructural:**

i. Se quitaron columnas de **varianza cero** y se **alinearon niveles** entre train/test.

ii. $X_i^{amenidades}$ ➡ dummies indicando si cuenta con variables como garaje, etc.

iii. $X_i^{estructura}$ ➡ cantidad de habitaciones, baños, superficie, etc.

iv. $X_i^{distancia}$ ➡ distancia a colegios, restaurantes etc.

v. $X_i^{locación}$ ➡ localidad y barrio.

vi. X_i^{tiempo} ➡ mes y año.

vii. X_i^{tipo} ➡ casa o apartamento.

Universidad de los Andes

Random Forest CV espacial

Elemento	Descripción
Tipo de modelo	Random Forest con validación cruzada espacial (ensamble por folds)
Objetivo / Métrica	Minimizar MAE (validación 5-fold)
Hiperparámetros	num.trees = 1000; mtry ≈ √p; min.node.size = 10; sample.fraction = 0.80; splitrule = variance; seed = 2025
Esquema de validación	Regular 5-fold: MAE = 120,886,984; Espacial 5-fold: MAE = 131,308,700
Estrategia de predicción	Promedio de las K predicciones por fold espacial (ensamble por folds)
Top variables	surface_total, bedrooms, estrato, dist_centros_comerciales, has_parqueadero_garaje, localidad
Kaggle (public LB)	209,389,203

Usando el CV espacial se generalizaron mejor las zonas no vistas. El ensamble por folds reduce varianza y el bosque capta no linealidades e interacciones entre tamaño, ubicación, distancias y amenidades, entregando predicciones estables sin sobreajuste.

Universidad de los Andes

Validación por CV espacial

- Valida por $X_i^{distancia}$, creando pliegues que no comparten área entre entrenamiento y validación (5 folds).
- **Evidencia en nuestros resultados (MAE):**
  1. RF regular (CV aleatoria) → Kaggle 213,23 M.
  2. RF espacial (CV por zonas) → Kaggle 209,39 M.
- Estimación interna más realista debido a la importancia de la ubicación en el espacio para este contexto.
- Es major usar CV espacial cuando las features incluyen ubicación y se espera que el test provenga de zonas distintas al train.
- Contra: algo más costosa y con mayor varianza entre pliegues, pero resulta mejor el entrenamiento para generalizar fuera de muestra.

Universidad de los Andes

Por qué este modelo fue el mejor

- Generalizó mejor a áreas no vistas, reflejado en el menor MAE en Kaggle frente al RF regular.
- El ensamble por folds espaciales redujo la varianza de las predicciones.
- El bosque capturó no linealidades e interacciones entre tamaño, ubicación, distancias y amenidades sin sobreajuste.
- Pipeline simple y robusto (sin varianza cero y dummies consistentes) aplicado dentro de cada fold.