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

Problem Set 3 – Making Money with ML?

BIG DATA Y MACHINE LEARNING PARA ECONOMIA APLICADA

2025-02

othermodels_equipo_08

OTROS MODELOS

Universidad de los Andes

Best model overview – 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

Comparativa de modelos

	Características del modelo			Desempeño	
Modelo	**Parametros**	**Preprocesamiento**	**Validacion**	**MAE_Kaggle**	**MAE_interno**
RF espacial (ganador)	1000 árboles; mtry≈√p; frac.=0.8; min.node=10; ensamble espacial	Eliminación variables varianza 0; imputación NA (Si hay); eliminación ZV; distancias geográficas	CV espacial por zonas	209,389,203	131,308,700
RF var. 1	1000 árboles; mismos hiperparámetros base	Mismo pipeline; sin ensamble espacial	CV aleatoria	213,229,154	120,886,984
RF var. 2	1000 árboles; cambio leve min.node	Pipeline estándar	CV aleatoria	213,381,370	120,973,000
SL (lm + RF)	Regresión + RF; mtry bajo	Solo apartamentos; dummies; interacción año–parques	CV aleatoria	218,792,827	46,237,960
SL (glmnet–RF–XGB)	GLMnet + RF reg. + XGB reg.	Categóricas a dummies; año x distancia y tipo_propiedad x distancia ; filtrado colinealidad	CV aleatoria (10 folds)	219,447,036	93,851,747
SL espacial (lm + RF)	Regresión + RF	Categóricas a dummies; interacción año–parques	CV por localidad	222,174,503	46,288,722
SL (RF–glmnet–XGB)	GLMnet + RF + XGB	Transformaciones; pm2; densidad habitacional, pm2 localidad; filtrado de outliers; dummies model.matrix	CV aleatoria (5 folds)	225,716,496	65,845,521
Boosting MAE	eta=0.05; max_depth=5; subsample=0.7; MAE explícita	model.matrix; interacción año–dist.	CV aleatoria (5 folds)	225,957,622	94,090,177
Caret (RF–GBM–XGB)	Grillas amplias; selección por RMSE log-price	Eliminación outliers 1%; log1p(price)	CV aleatoria (5 folds)	226,879,601	91,367,282

Universidad de los Andes

Resumen de los 9 modelos

Los 9 modelos representan tres familias:

- **Bosques aleatorios (RF)**
- **SuperLearner (ensambles híbridos)**
- **Boosting y modelos entrenados en Caret**

Hallazgo general:

- La diferencia entre *validación aleatoria* y *validación espacial* explica gran parte de las caídas en Kaggle.
- Los modelos entrenados con validación aleatoria tuvieron un MAE interno muy bajo, pero fallaron al generalizar.

¿Por qué los Random Forest alternativos no ganaron?

Modelos: RF variante 1 y variante 2

- Mismo algoritmo y casi los mismos hiperparámetros que el RF ganador:

Problema clave:

Validación aleatoria 5-fold

- mezcla viviendas cercanas.
- Fuga espacial → MAE interno artificialmente optimista (≈ 120M).

Conclusión:

- Con la misma arquitectura, solo cambiar *CV aleatoria → CV espacial* define la diferencia entre “bueno” y “mejor del equipo”.

¿Por qué los Superlearners no ganaron?

Problema estructural:

- SuperLearner no puede optimizar MAE internamente. El paquete Superlearner de R no lo permite directamente y se entrena bajo la familia gaussiana.
- Trabajó bajo familia gaussiana → pérdida cuadrática (MSE/RMSE) = ajusta errores cuadrados

Implicación directa:

- Penaliza más unos pocos errores grandes en vez de reducir el MAE.

Otros problemas según versión:

- **SL básico:**
  - filtra apartamentos, genera pocas interacciones
- **SL regularizado (GLMnet + RF + XGB):**
  - Interacciones automáticas, eliminación de colinealidad
  - CV aleatoria
- **SL con CV espacial por localidad:**
  - validación espacial muy gruesa (localidad = zona enorme)
- **SL avanzado:**
  - Mejor ingeniería de variables de toda la clase
  - Pero CV aleatoria + RMSE interno → **sobreajuste micro-local**

Conclusión:

- Métrica equivocada, espacialdad por localidad muy amplia.

¿ Por qué el boosting y caret no ganaron?

XGBoost con métrica MAE

Problemas:

- validación aleatoria 5-fold
- aprende patrones micro-locales
- MAE interno muy bajo → sobreoptimismo

Caret (RF–GBM–XGB)

- Se evaluaron 3 modelos con el mismo tratamiento de datos pero se seleccionó el que tuviera menor RMSE → CARET

Problemas:

Pipelines limpios y grillas amplias.

Pero:

- optimiza RMSE del log-precio
- transformación logarítmica desalineada con MAE real
- CV aleatoria