Comparación de modelos de regresión

Se evaluaron cuatro modelos de regresión para predecir los ingresos totales de clientes bancarios (Rev_Total): Elastic Net, Árbol de regresión, Random Forest y KNN regresión. La comparación se realizó sobre el conjunto de prueba, utilizando como métricas principales el RMSE, el MAE y el coeficiente de determinación R^2.

Los resultados se resumen en la siguiente tabla (conjunto de prueba):

Modelo	**RMSE**	**MAE**	**R^2**
Random Forest	2.7824	1.4533	0.23106
Árbol de regresión	2.8042	1.4903	0.21894
KNN	2.9178	1.5683	0.15441
Elastic Net	3.0056	1.6918	0.10275

Se observa que los modelos basados en árboles (Árbol de regresión y, especialmente, Random Forest) superan claramente al modelo lineal penalizado (Elastic Net) y al modelo KNN. El Random Forest presenta el mejor compromiso entre error de predicción y capacidad explicativa, con un R^2 cercano a 0.23.

Aunque el poder predictivo sigue siendo moderado (el mejor modelo explica alrededor de un 23% de la variabilidad de los ingresos), los resultados sugieren que existen relaciones no lineales y posibles interacciones entre variables que son mejor capturadas por modelos de tipo árbol. Esto también indica que podría ser necesario incorporar nuevas variables o realizar una mayor ingeniería de atributos para mejorar el desempeño global del modelo.

Flujo de Trabajo

Metodología aplicada

1. Exploración inicial del dataset: revisión de estructura, variables, datos faltantes, tipos.

2. Preprocesamiento: eliminación de ID, limpieza, factores, normalización donde correspondía.

3. Partición de datos: 80% entrenamiento / 20% prueba (estratificado por cuantiles de Rev_Total).

4. Entrenamiento de modelos:

- Elastic Net (glmnet)
- Árbol de regresión (rpart)
- Random Forest (randomForest / ranger)
- KNN regresión (caret)

5. Validación cruzada: CV 3-fold (optimización de hiperparámetros).

6. Evaluación

- RMSE, MAE, R^2
- Comparación de desempeño

7. Selección del modelo ganador: Random Forest

8. Resumen y conclusiones

Conclusión

El proyecto permitió construir y comparar distintos modelos de Machine Learning para la predicción de ingresos de clientes bancarios en R, siguiendo un flujo reproducible de trabajo: preprocesamiento, partición train/test, ajuste de hiperparámetros con validación cruzada y evaluación con métricas estándar (RMSE, MAE, R^2).

El modelo de referencia (Elastic Net) mostró un desempeño limitado, con un R^2 en torno a 0.10. Al incorporar modelos no lineales, en particular Random Forest, se consiguió una mejora significativa en el error de predicción y en la capacidad explicativa, alcanzando un R^2 cercano a 0.23 en el conjunto de prueba.

Si bien estos resultados son útiles para una primera aproximación a la estimación de ingresos y podrían apoyar tareas de segmentación y priorización comercial, el nivel de error obtenido sugiere que el problema es complejo y que sería recomendable profundizar en la selección e ingeniería de variables, así como explorar otros algoritmos (p. ej., Gradient Boosting, XGBoost, modelos aditivos, etc.) para seguir mejorando el desempeño.