**Conclusión de la práctica de Betlem Ferrer**

Para realizar esta práctica he hecho 4 hipótesis diferentes en cuanto a selección y tratamiento de características. El motivo es que una de las características más importantes del modelo (Square Feet) sólo tiene unas 314 filas con datos, de las 12893 filas que tiene el modelo.

Los casos considerados son:

1. No tener en cuenta el que haya tan pocos valores con Square Feet y rellenándolos con la media de los Square Feet de los que sí que tienen datos.
2. He modelado el modelo con sólo los datos que tienen Square Feet 314 filas que se dividen en train y test.
3. División train: las 314 filas que contienen Square Feet y test el resto de filas (haciendo una división manual).
4. Para el cuarto modelo, he eliminado la característica Square Feet, porque tiene muy pocos datos y he querido ver qué salía.

Estos son los resultados:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** |
| Random Forest | 0,592 | 0,281 | 0,408 | 0,588 |
| Lasso | MSE = 1437,54 | MSE = 1064,8 | MSE = 1888,66 | MSE = 1439 |
| Regresión Lineal | - | MSE = 1105,71 | MSE = 1798,15 | MSE = 1458 |
| Cross Validation (con Random Forest) | Me daba error | Me daba error | Me daba error | Me daba error |

Viendo estos resultados, he sacado la conclusión de que no vale mucho la pena tratar el Square Feet, por lo tanto, he intentado mejorar el modelo 1, quitando Outliers y añadiendo características.

También puede ser que el modelo 2 presente underfitting, por el hecho de que se han tomado muy pocas muestras.

Los resultados se pueden ver en **“Practica Betlem Ferrer1 ok.ipynb”**

Y el resto de notebooks son las otras pruebas, no hace falta que las corrijas.

Después de aplicar las mejoras, el modelo ha mejorado un poco con un resultado de: **Random Forest: 0,621 y un MSE con Lasso de: 1279.**