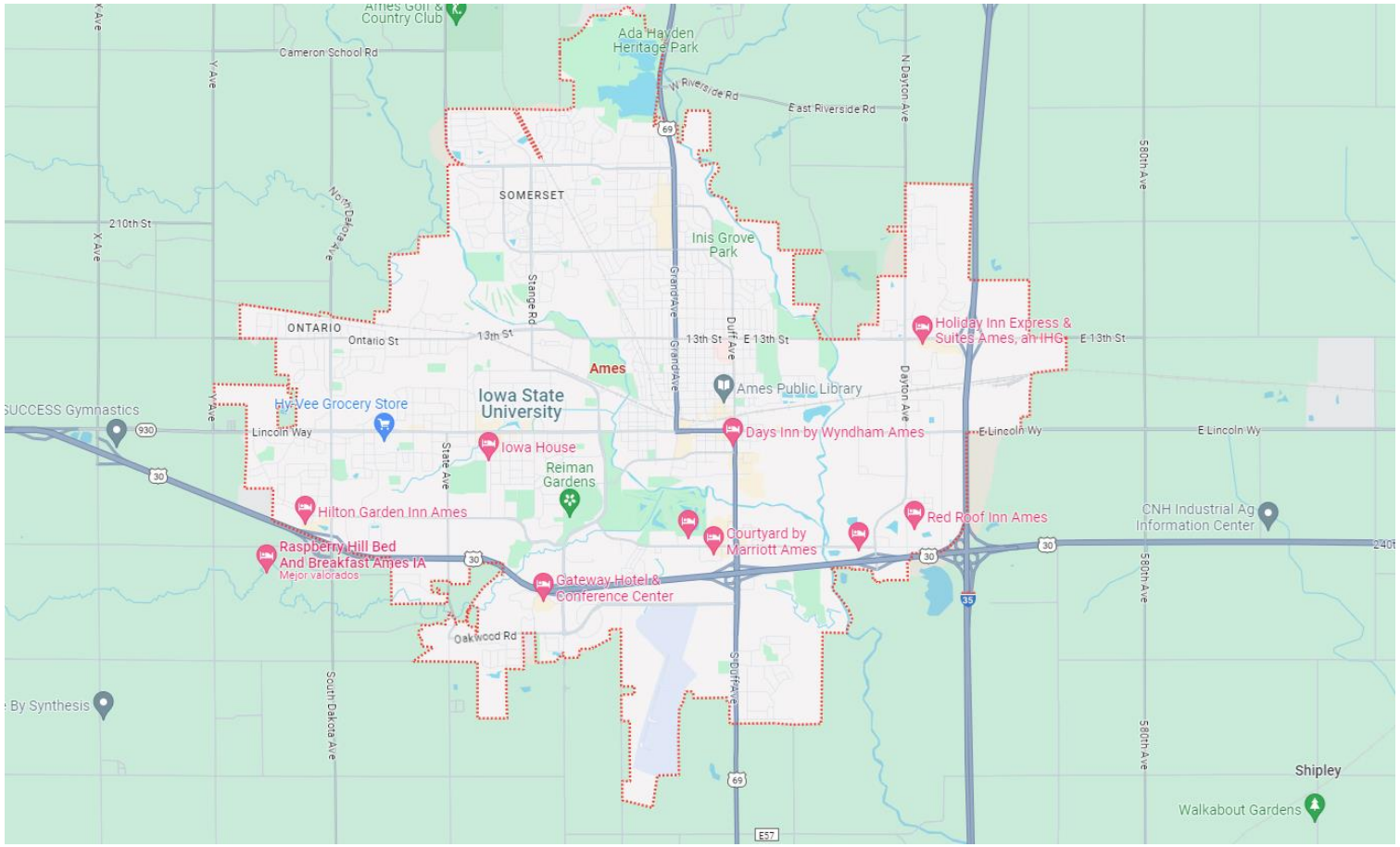
House Prices - Advanced Regression Techniques

POR JAVIER MARTÍNEZ Y JORGE LÓPEZ

Descripción del proyecto



Objetivo: Predecir el precio de venta para cada casa.

Métrica: La Raíz del Error Cuadrático Medio (RMSE)

Análisis de variables

Dataframes de train y test

79 variables explicativas, 37 numéricas y 43 categóricas.

Nuestra variable objetivo es el precio de venta

Usaremos las librerias **ydata_profiling** para agilizar el análisis.

https://github.com/ydataai/ydata-profiling

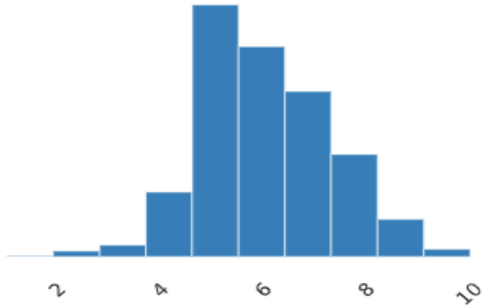
Análisis de variables

OverallQual

Real number (ℝ)

HIGH CORRELATION

Distinct	10	**Minimum**	1
Distinct (%)	0.7%	**Maximum**	10
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	6.0788211	**Memory size**	11.5 KiB



More details

Statistics | Histogram | Common values | Extreme values

Quantile statistics

Minimum	1
5-th percentile	4
Q1	5
median	6
Q3	7
95-th percentile	9
Maximum	10
Range	9
Interquartile range (IQR)	2

Descriptive statistics

Standard deviation	1.4368116
Coefficient of variation (CV)	0.23636353
Kurtosis	0.037640747
Mean	6.0788211
Median Absolute Deviation (MAD)	1
Skewness	0.18119602
Sum	8869
Variance	2.0644277
Monotonicity	Not monotonic

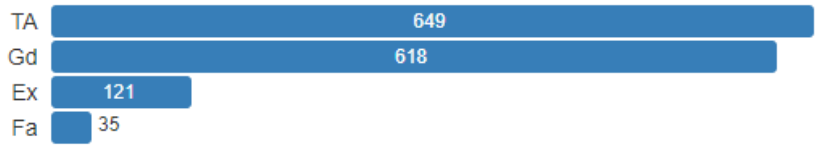
Análisis de variables

BsmtQual

Categorical

HIGH_CORRELATION MISSING

Distinct	4
Distinct (%)	0.3%
Missing	37
Missing (%)	2.5%
Memory size	11.5 KiB



More details

Overview | Categories | Words | Characters

Length

Max length	2
Median length	2
Mean length	2
Min length	2

Characters and Unicode

Total characters	2846
Distinct characters	8
Distinct categories	2
Distinct scripts	1
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	0
Unique (%)	0.0%

Sample

1st row	Gd
2nd row	Gd
3rd row	Gd
4th row	TA
5th row	Gd

Transformación de variables

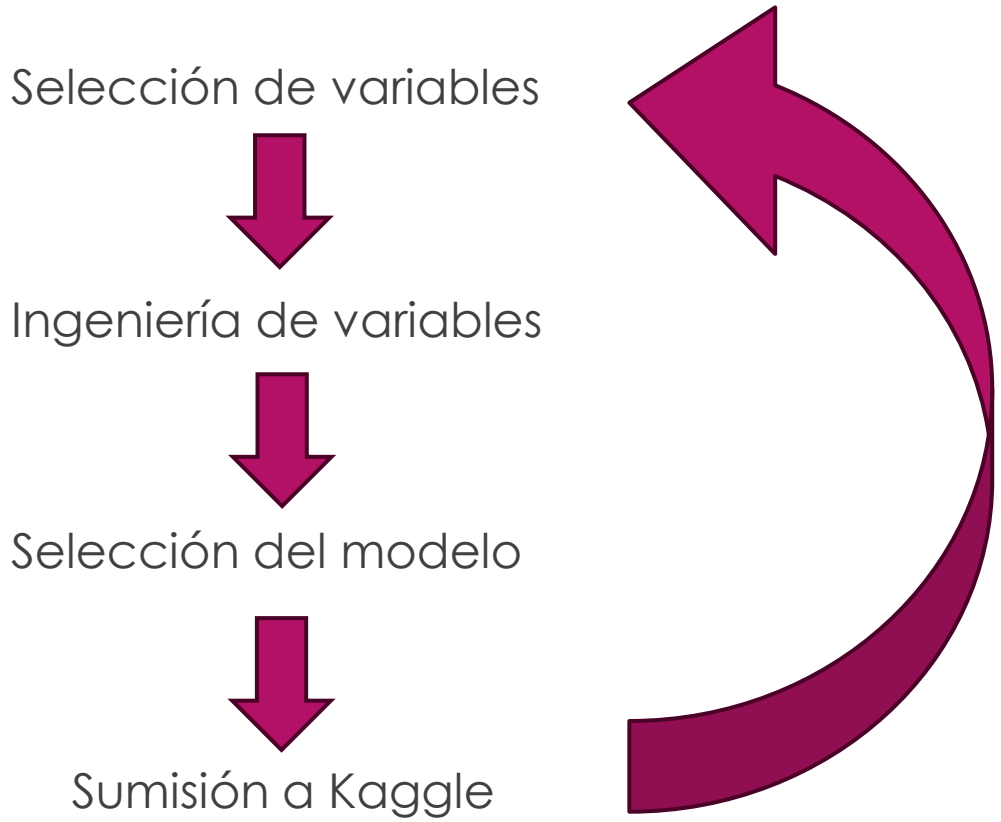
Hacemos una primera transformación de variables que mantendremos durante todo el proyecto.

Rellenamos nulos para variables numéricas, por ejemplo, para aquellas casas con area de garage nula les asignamos el valor 0, puesto que esto significa que no tienen garage. Pero para otras como la distancia de la propiedad a la calle asignamos la media del dataset a los nulos.

Para las variables categóricas rellenamos los nulos con el valor respectivo que para cada campo hace referencia a la ausencia de valor para esa variable (NA, No_Garage, None…)

También eliminamos aquí los outliers.

Loop de trabajo



Selección de modelos

Usaremos **Lazy Predict** para obtener el major modelo para las variables seleccionadas

1. Selección automática de modelos de entre 30-40 modelos
2. Mínima configuración
3. Versatilidad
4. Informe completo de resultados

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
GradientBoostingRegressor	0.88	0.89	28263.66	0.39
ExtraTreesRegressor	0.88	0.89	28266.95	0.47
HistGradientBoostingRegressor	0.88	0.89	28304.96	0.31
LGBMRegressor	0.88	0.89	28393.12	0.10
PoissonRegressor	0.87	0.88	29090.16	0.54
RandomForestRegressor	0.86	0.87	30564.48	0.76
BaggingRegressor	0.85	0.87	31073.26	0.09
XGBRegressor	0.85	0.86	31442.68	0.25

https://lazypredict.readthedocs.io/en/latest/

Primera sumisión

Escogemos las 10 variables numéricas más correlacionadas.

1. Calidad general de la casa
2. Superficie de la propiedad en pies cuadrados
3. Número de plazas de garage
4. Área de garage
5. Área del sotano
6. Área de la primera planta
7. Número de baños
8. Número de habitaciones
9. Año de construcción
10. Año de remodelación

Primera sumisión

Lazy predict → XGBRegressor → Fine tunning de hiperparametros

```
param_grid={'n_estimators':[50,100,150,200],
            'learning_rate': [0.1,0.075,0.05],
            'alpha': [0.5,0.9],
            'max_depth':[2,3,4],
            'min_samples_leaf':[1,3,5],
            'max_features':[1.0,2.0,3.0] }
```

RESULTADO

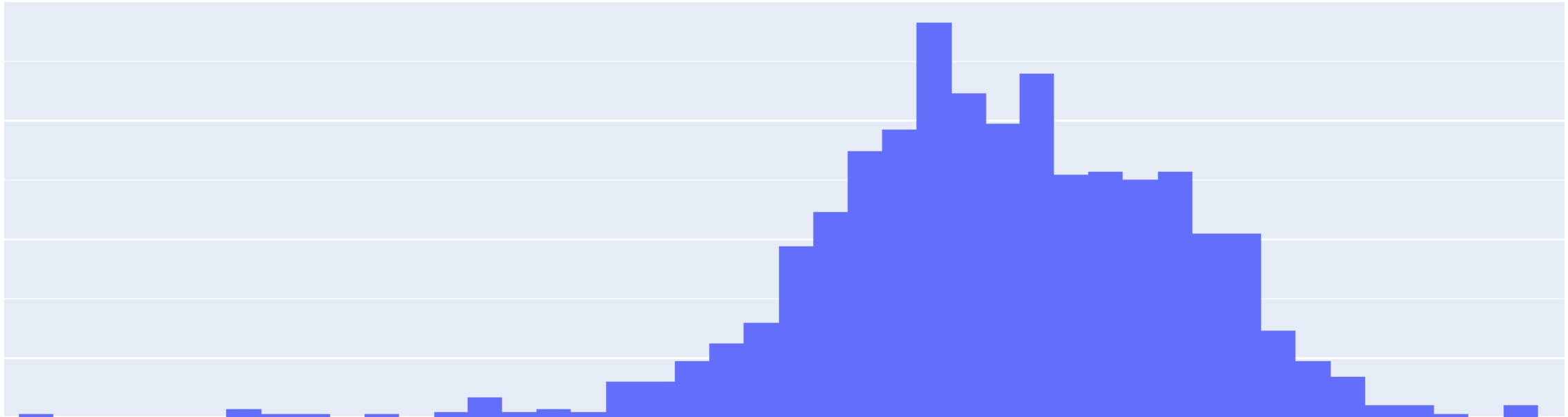
kaggle_submission.csv **0.15938**

Complete · 2mo ago · submission-01

Segunda sumisión

Para la segunda sumisión cogemos las 30 variables numéricas más correlacionadas y las 10 categóricas más correlacionadas.

Normalizamos las variables numéricas usando la transformación logarítmica y hacemos un One-Hot-Encoding para las variables categóricas.



Segunda sumisión

Con lazy predict obtenemos que el major modelo es GradientBoostingRegressor.

Lazy predict → GradientBoostingRegressor → Fine tunning de hiperparametros

RESULTADO

kaggle_submission.csv **0.14379**

Complete · 2mo ago · submission-01

Tercera sumisión

Usamos las mismas variables y transformaciones que en la segunda sumisión.

Como modelo haremos un model blending de los 5 mejores.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
GradientBoostingRegressor	0.88	0.89	28263.66	2.41
ExtraTreesRegressor	0.88	0.89	28266.95	1.35
HistGradientBoostingRegressor	0.88	0.89	28304.96	8.23
LGBMRegressor	0.88	0.89	28393.12	0.21
PoissonRegressor	0.87	0.88	29090.16	1.25

Tercera sumisión

El model blending con sus pesos.

```
modelos_name=['GradientBoostingRegressor','ExtraTreesRegressor', 'RandomForestRegressor', 'XGBRegressor','PoissonRegressor']
test['SalePrice']=weights[0]*y_models_list[0]+weights[1]*y_models_list[1]+weights[2]*y_models_list[2]+weights[3]*y_models_list[3]+weights[4]*y_models_list[4]
```

```
weights=[0.40,0.20,0.2,0.1,0.1]
```

RESULTADO

kaggle_submission.csv **0.14305**

Complete · 2mo ago · submission-01

MUCHAS GRACIAS