Hause Price Competition

Sergio Carrasco Márquez

Correo: sergiocmarq@gmail.com

Contenido

1 Introducción al problema

El problema consiste en aplicar técnicas de regresión para intentar aproximar el precio de venta final de una vivienda. El conjunto de datos sobre el que se aplicarán dichas técnicas consta de 77 variables y de 1460 entradas.

En la web de la asignatura se facilita un script que ejecuta dos algoritmos de regresión sobre el conjunto de datos y utiliza la media de los precios obtenidos por cada algoritmo para establecer el precio final de la vivienda.

2 Resolución del problema

El script mencionado con anterioridad se usará como punto de partida para crear los scripts usados para realizar los sucesivos experimentos. La parte sobre la que se realizarán las modificaciones será la parte de preprocesado de los datos y análisis de variables. La parte correspondiente a la ejecución de la regresión será la misma.

2.1 Preprocesado general

Antes de analizar que variables se descartan por la correlación que guardan con la variable que determina el precio se deben preprocesar los datos.

En primer lugar, se deben descartar las variables que tengan demasiados valores perdidos. Luego será necesario rellenar los valores perdidos de cada una de las instancias para cada variable no eliminada. En el caso de las variables numéricas se usa el valor medio, y para las variables categóricas se usa el valor más frecuente.

Previamente se ha modificado la forma de tratar con algunas variables. Existen variables que expresan calidad o algún tipo de puntuación, dichas variables se han extrapolado a puntuación numérica. De esta forma se puede entender que a mayor puntuación en alguno de los aspectos, el precio tendería a subir. Otras variables que se expresan en forma numérica, no deben ser tratadas de esta forma, sino como variables categóricas, como por ejemplo el mes en el que se vende una casa.

2.2 Estudio de la correlación de variables

Para comenzar el análisis de variables, se ha comenzado analizando la correlación de las variables con el precio final de la vivienda. La idea que subyace es que cuando una variable guarda poca correlación con el precio de la vivienda no interesa utilizarla para la regresión. La correlación de variables se muestra a continuación.

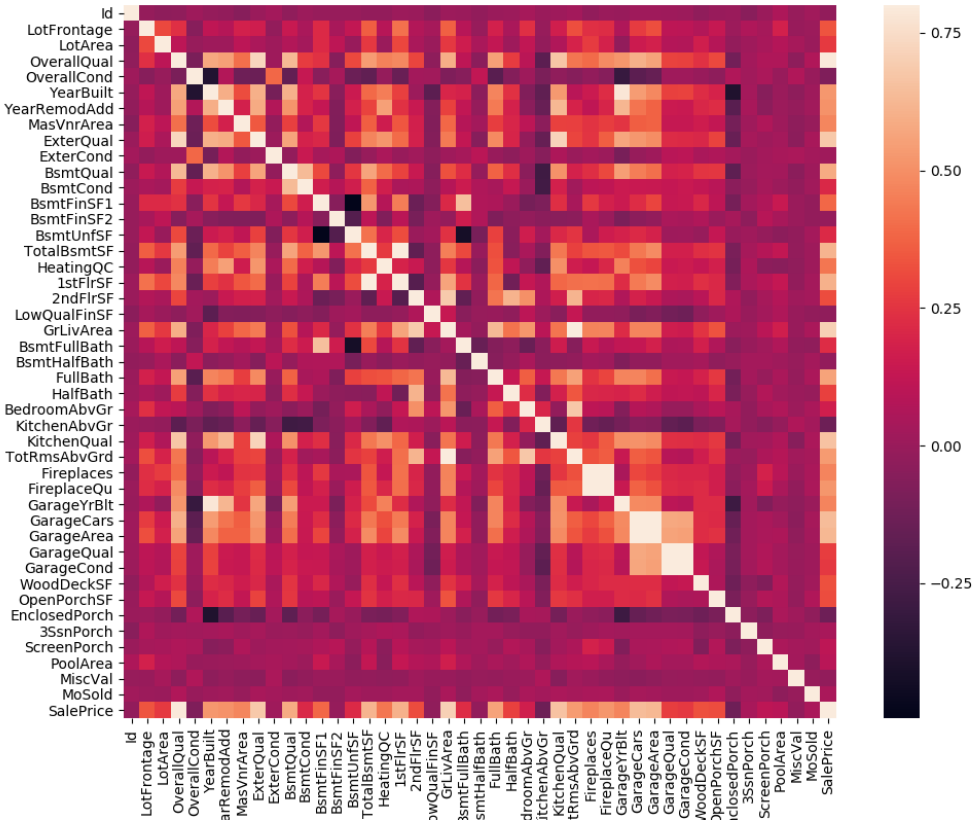

Ilustración 1: Matriz de correlación de variables

En la última fila de la matriz es dónde encontramos la parte que corresponde a la correlación de las variables con el precio final de la vivienda. Algunas variables como OverallCond o GrLivArea guardan mucha correlación con el precio final. Para los primeros experimentos se evalúan los algoritmos y la solución creada en el script sobre el que se trabaja podando algunas variables que guardan poca correlación con el precio final de la vivienda. Los resultados de dichos experimentos se muestran a continuación.

Descripción	Puntiación Kaggle
Variables con un valor distinto a 0 en correlación	0.11766
Variables con un valor mayor a 0.7 en correlación	0.13813
Variables con un valor mayor a 0.3 en correlación	0.12969

Tabla 1: Resultados en Kaggle de la poda de variables en función de la correlación de las variables con el precio

Como el objetivo es minimizar la puntuación, la siguiente gráfica muestra la calidad de cada resultado a partir de la inversa de dicha puntuación.

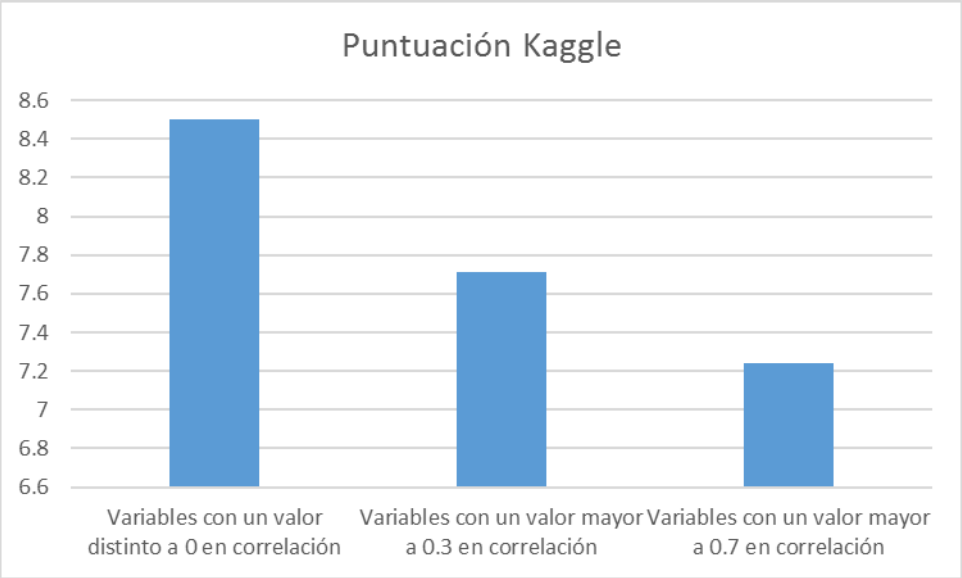

Ilustración 2: Comparativa de resultados del primer experimento

En la gráfica se observa de forma directa que minimizar el número de variables podadas en función de la correlación es una buena estrategia, por lo que todos los demás experimentos partirán de esta base. El script muestra el número de variables podadas, y en el caso de podar solo las que tengan 0 en el valor de correlación no se poda ninguna variable. En este caso en concreto parece que todas las variables aportan información valiosa a la hora de realizar la regresión.

2.3 Estudio de las variables categóricas

El siguiente aspecto importante para investigar es la relación de las variables categóricas con el precio final de la vivienda. Este análisis es más complejo que el anterior, pues requiere observar cada variable una a una. En el experimento anterior se ha logrado un buen resultado al hacer uso de todas las variables disponibles, por lo que es lógico pensar que dicho fenómeno se puede extrapolar también a las variables categóricas. El script facilitado en la web propone una serie de variables categóricas que deben ser eliminadas. Dichas variables han permanecido eliminadas en los experimentos anteriores. Eliminar más variables, aparte de las sugeridas también es otra opción a tener en cuenta.

Ahora es necesario distinguir qué variables deben ser eliminadas, y cuáles no. De forma visual se puede ver qué relación guarda cada variable con el precio final. En primer lugar se observa la relación entre el precio y las variables sugeridas por el script inicial, para confirmar si realmente estas variables merecen ser eliminadas del conjunto de variables o si por el contrario deben tenerse en cuenta a la hora de calcular el precio final de la vivienda.

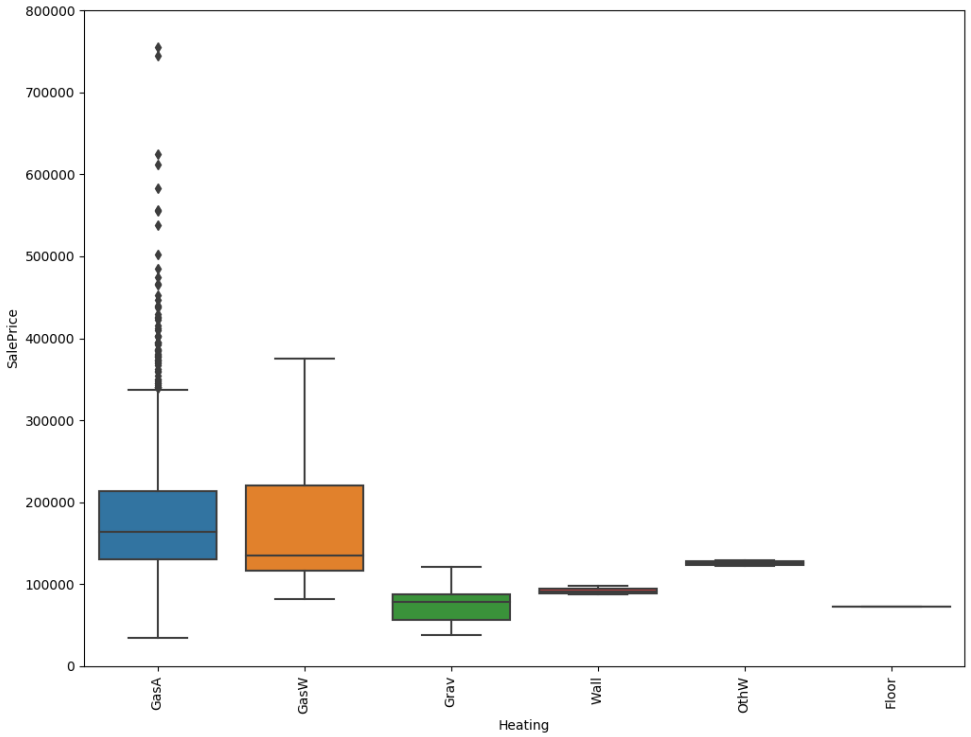

Ilustración 3: Relación de Heating con el precio final

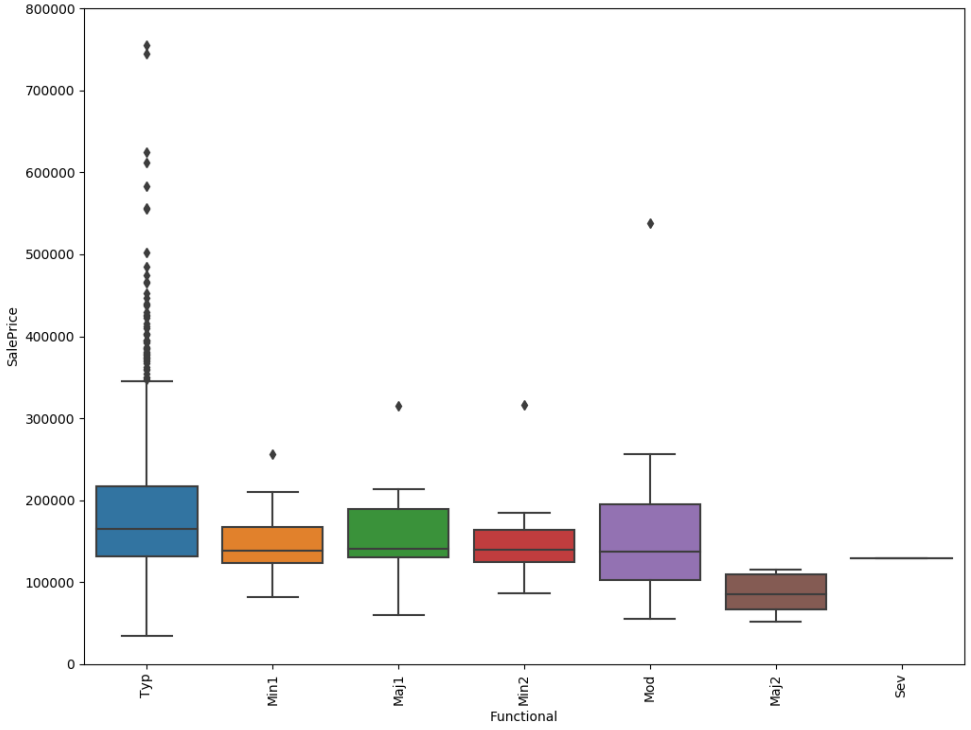

Ilustración 4: Relación de Functional con el precio final

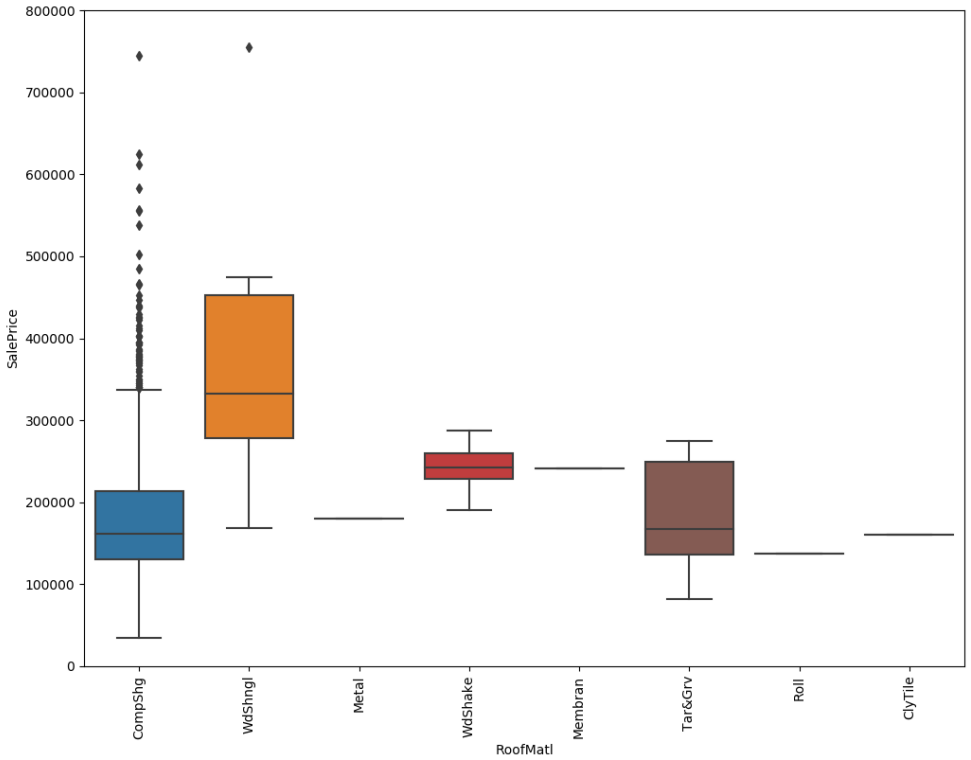

Ilustración 5: Relación de RoofMalt con el precio final

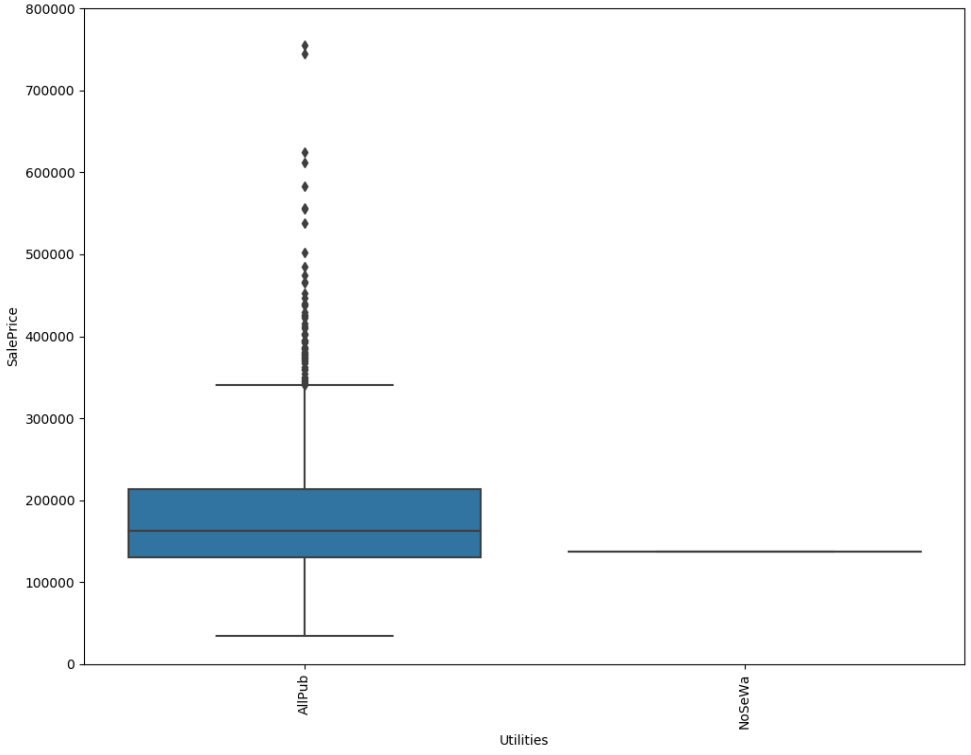

Ilustración 6: Relación de Utilities con el precio final

En cuanto a las figuras anteriores no se aprecia de forma clara una distinción de precios en función de los valores de las variables. Algunas de las variables tienen un rango muy alto de precios para un mismo valor, como Utilities. Otras ofrecen rangos de precios muy similares para distintos valores de la variable, como Functional. Tras analizar el resto de variables se han escogido algunas más.

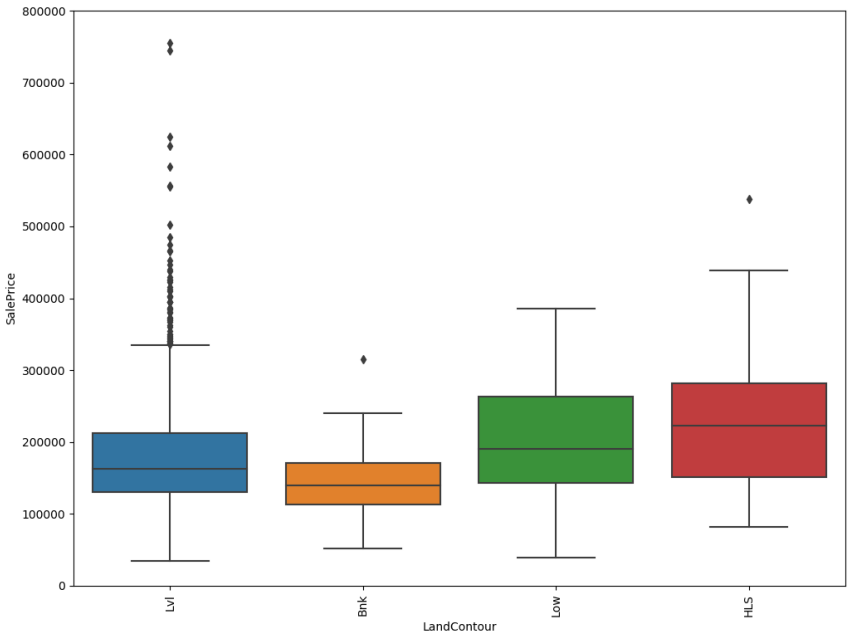

Ilustración 7: Relación de LandContour con el precio final

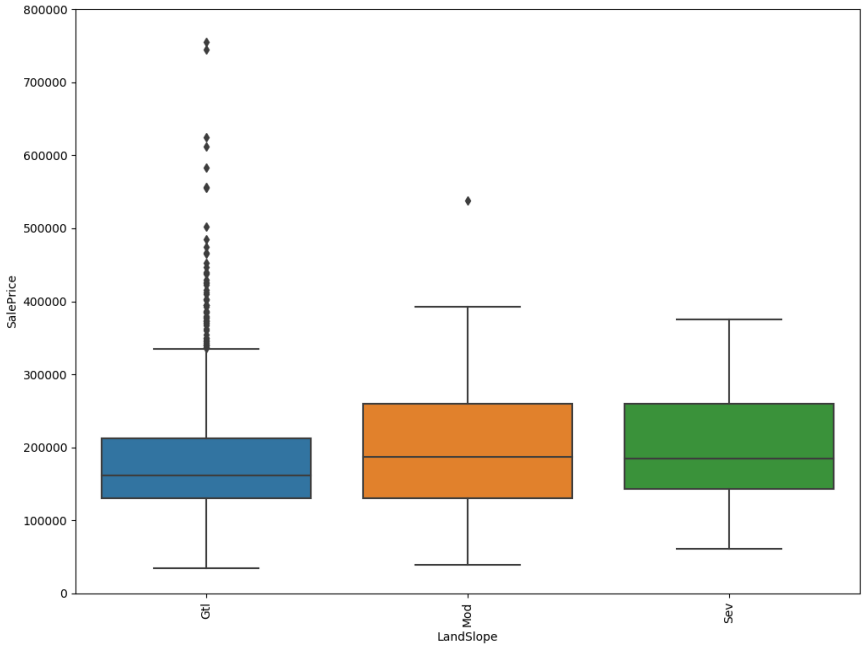

Ilustración 8: Relación de LandSlope con el precio final

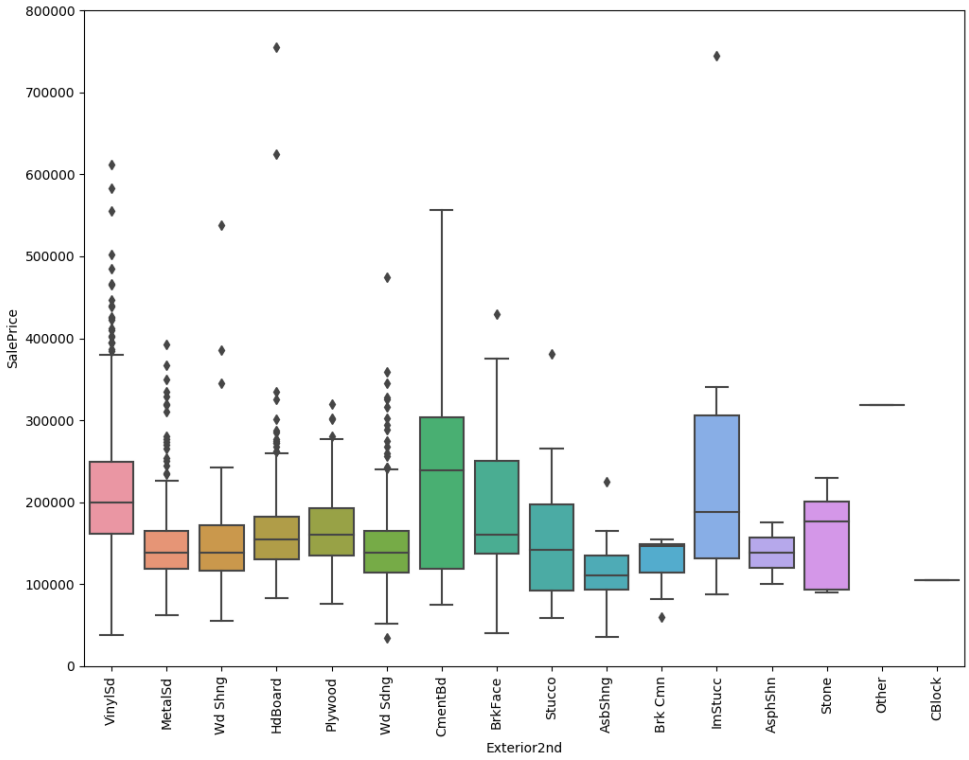

Ilustración 9: Relación de Exteriornd con el precio final

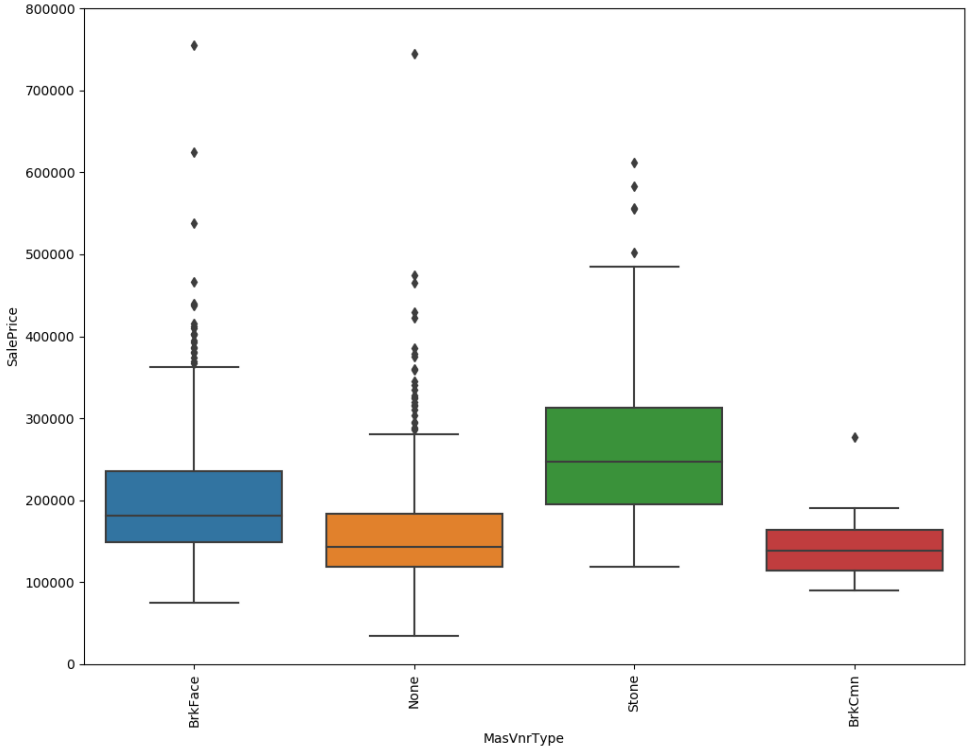

Ilustración 10: Relación de MasVnrType con el precio final

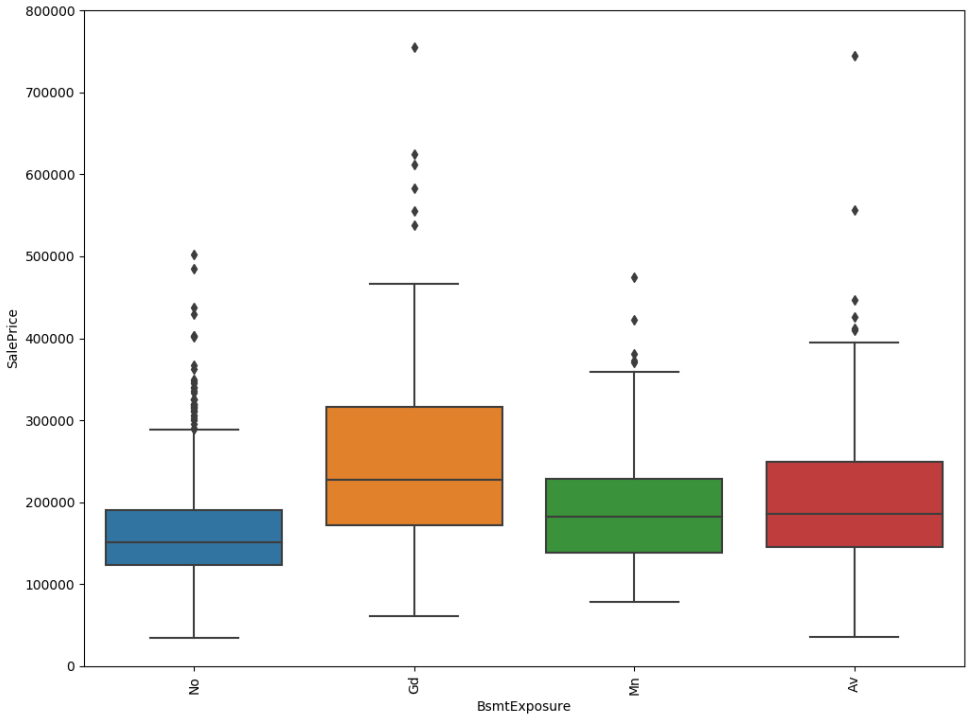

Ilustración 11: Relación de BsmtExposure con el precio final

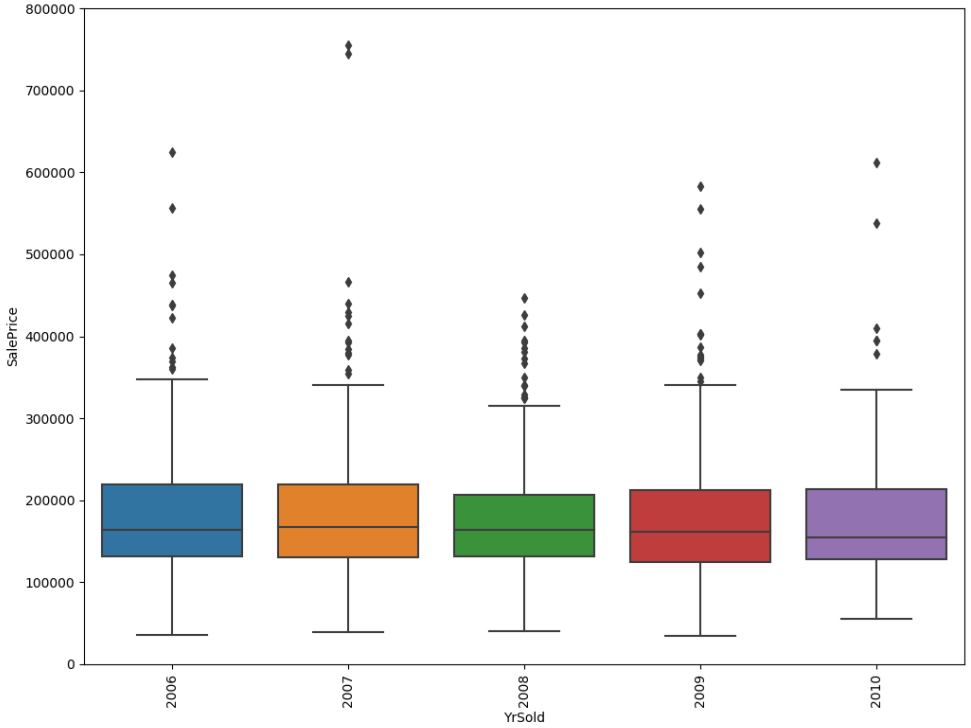

Ilustración 12: Relación de YrSold con el precio final

Una vez que se tiene una idea aproximada de que variables pueden ser eliminadas, es necesario comprobar de forma empírica que se pueden eliminar. Se han realizado 2 experimentos más partiendo de la mejor solución del experimento anterior. En uno de los experimentos se eliminan todas las variables mencionadas anteriormente, en el otro ninguna. Una vez realizados dichos experimentos se comparan los resultados con la mejor solución hasta el momento que solo eliminaba las 4 primeras variables mencionadas con anterioridad.

Descripción	Puntuación Kaggle
Todas las categóricas	0.11775
Conjunto reducido de categóricas	0.12048
Mejor resultado hasta el momento	0.11766

Tabla 2: Resultados en Kaggle de la experimentación con variables categóricas

Al comparar la inversa de las puntuaciones se obtiene la siguiente gráfica.

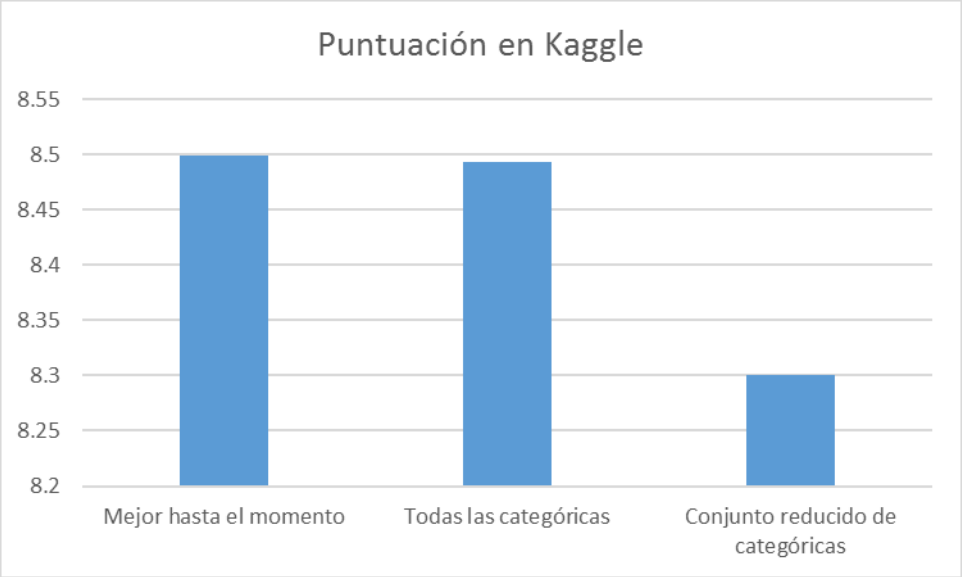

Ilustración 13: Comparativa de resultados del segundo experimento

Aparentemente dejar o quitar las variables que el script original sugiere eliminar es indiferente a grandes rasgos. Aunque las puntuaciones de la competición son bastante similares, por lo que cualquier mejora, por pequeña que sea puede avanzar varios puestos. Lo que sí está claro es que reducir el conjunto de variables en gran medida parece perjudicar bastante la puntuación.

2.4 Estudio de la poda de valores perdidos

Hasta ahora no se ha revisado si la imputación de variables por tener demasiados valores perdidos es la correcta. Algunos de los tutoriales disponibles en Kaggle sugieren podar aquellas variables que tengan más de la mitad de valores perdidos. Aparentemente realizar una poda con un umbral del 0.5 en valores perdidos puede dejar demasiadas variables que no puedan reponer los valores perdidos manteniendo la calidad de los datos. En este caso los valores perdidos se han rellenado usando la media para las variables numéricas y la moda para las categóricas. Existen muchas otras técnicas para reparar los valores perdidos que no han llegado a probarse, pero que podrían haber dejado mejores resultados. Los tutoriales de Kaggle sobre esta competición usan la media y la moda para rellenar valores perdidos, por lo que no se ha hecho hincapié en la cuestión de cómo rellenar los datos, sino que se ha experimentado con la cantidad de datos que se pueden reponer sin perder calidad usando solamente la media y la moda.

Descripción	Puntuación Kaggle
0.5 de valores perdidos	0.11892
0.8 de valores perdidos	0.11842
Mejor resultado hasta el momento	0.11766

Tabla 3: Comparación entre los distintos umbrales de valores perdidos

La inversa de la puntuación ofrece el siguiente resultado.

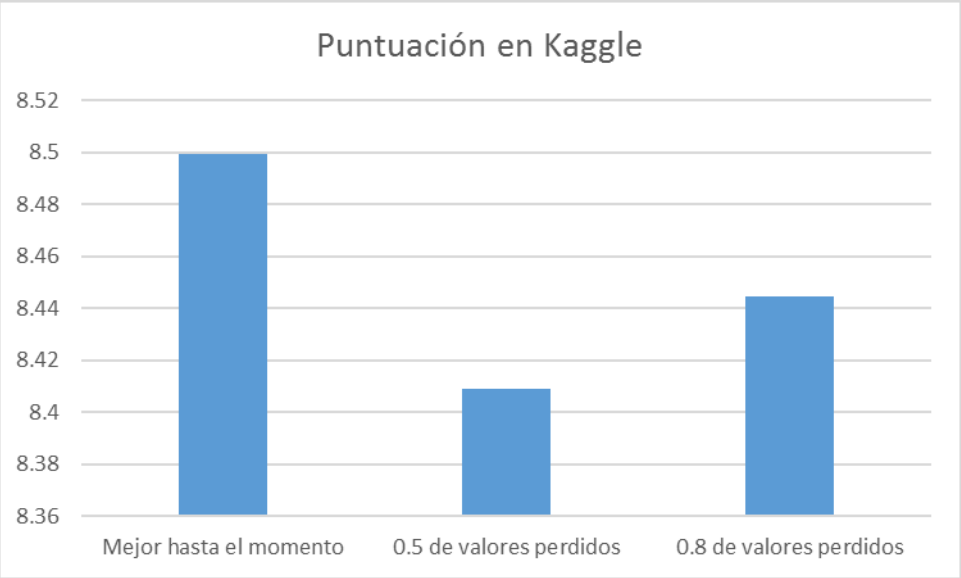

Ilustración 14: Resultados experimento 3

Variando el umbral de valores perdidos para eliminar variables no se obtiene ninguna mejora sobre el primer resultado obtenido.

2.5 Otras consideraciones

Aparentemente los algoritmos convergen en el mejor resultado del primer experimento. Pero aún puede hacerse algo más sobre las variables. En Kaggle hay un tutorial (Gawlik, s.f.) que sugiere aplicar transformaciones sobre los datos para ajustarlos a una distribución normal. Existen varios métodos y dependiendo de la distribución de la variable algunos se ajustan más que otros. En este caso se ha aplicado el logaritmo a cada una de las variables numéricas para aproximar la distribución a una normal. En la siguiente tabla se muestran los resultados obtenidos.

Descripción	Puntuación en Kaggle
Mejor hasta el momento	0.11766
Transformación logarítmica	0.11524

Tabla 4: Resultado de la transformación logarítmica

La mejora es notable y los siguientes experimentos partirán de este último. Con el objetivo de aumentar la puntuación obtenida en Kaggle se aumentan las iteraciones del algoritmo de ElasticNetCV. Así mismo, se prueba a realizar otro experimento para saber si la poda de variables en función de la correlación tiene efecto ahora. Los resultados a los 3 experimentos anteriores se muestran en la siguiente tabla.

Descripción	Puntuación en Kaggle
Escala Log	0.11524
Aumento de iteraciones	0.11435
Variables(log) con un valor mayor a 0.3 en correlación	0.11551

Tabla 5: Resultados 3 últimos experimentos

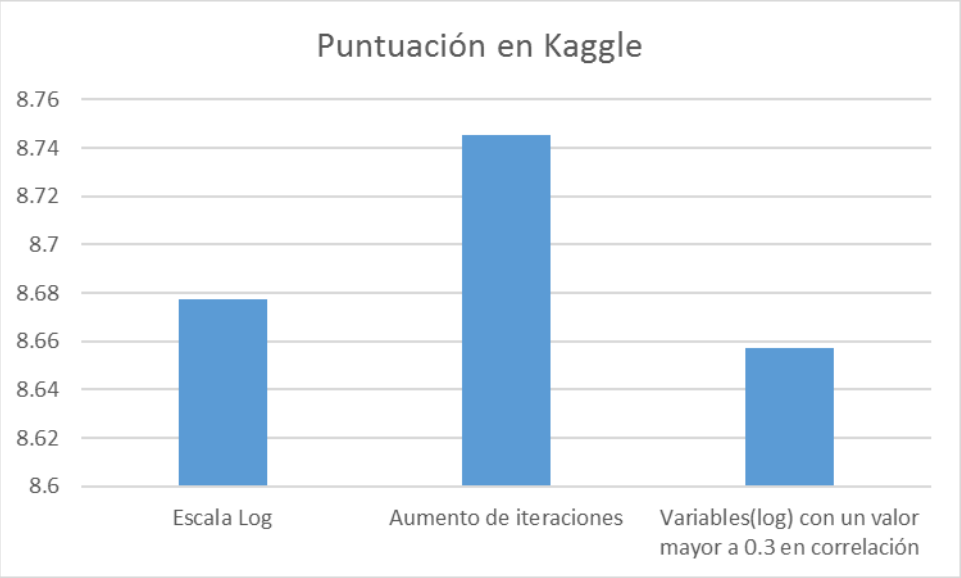

Ilustración 15: Últimos 3 experimentos en Kaggle

Como en los gráficos anteriores, se usa la inversa da la puntuación para la visualización de los resultados de una forma más intuitiva. Esta tabla refleja que el aumento de iteraciones es bastante efectivo, mientras que eliminar variables numéricas sigue siendo contraproducente.

3 Tabla resumen

A continuación, se muestra una tabla que resume todos los experimentos realizados hasta el momento. Salvo que se indique en la tabla, los parámetros de los algoritmos no han sido modificados del script original. Las métricas de train y test de R2 y RMSE se calculan de la misma forma que en script original y no cambia su significado. En cuanto a la métrica de RMSE, nótese que se calcula sobre el logaritmo de los precios, por lo que equivale a la RMSLE.

	ElasticNetCV				
Descripción	max_iter	R2-Score	RMSE	R2-Score_test	RMSE_test
Variables con un valor distinto a 0 en correlación	5000	0.903685899	0.118146559	0.903780031	0.108051592
Variables con un valor mayor a 0.3 en correlación	5000	0.894358484	0.123138724	0.880293797	0.118381569
Variables con un valor mayor a 0.7 en correlación	5000	0.862007847	0.137865846	0.838650465	0.131630579
Todas las categóricas	5000	0.916573854	0.110230165	0.920312541	0.099682247
Conjunto reducido de categóricas	5000	0.900291742	0.12005141	0.905770288	0.108129169
0.5 de valores perdidos	5000	0.903685924	0.11814656	0.903780101	0.10805157
0.8 de valores perdidos	5000	0.903685924	0.11814656	0.903780101	0.10805157
Escala Log	5000	0.910769	0.114170571	0.907863909	0.109369212
Aumento de iteraciones	500000	0.910769	0.114170571	0.907863907	0.109369213
Variables(log) con un valor mayor a 0.3 en correlación	500000	0.910863496	0.114117835	0.911011461	0.105975073

	Gradient Boosting			
Descripción	R2-Score	RMSE	R2-Score_test	RMSE_test
Variables con un valor distinto a 0 en correlación	0.968237655	0.069553717	0.915606186	0.102055917
Variables con un valor mayor a 0.3 en correlación	0.966168738	0.071421185	0.88989843	0.113715849
Variables con un valor mayor a 0.7 en correlación	0.940840529	0.093312278	0.865739531	0.123395199
Todas las categóricas	0.969293563	0.068446239	0.925274495	0.096206455
Conjunto reducido de categóricas	0.968776495	0.069018375	0.926443817	0.095923822
0.5 de valores perdidos	0.969325022	0.06835494	0.916993947	0.100639734
0.8 de valores perdidos	0.968813697	0.068962536	0.921386714	0.098981936
Escala Log	0.968427769	0.069372148	0.922358514	0.097664517
Aumento de iteraciones	0.968086732	0.069655505	0.927324396	0.094315309
Variables(log) con un valor mayor a 0.3 en correlación	0.970385821	0.06726876	0.921478925	0.098409836

Descripción	Puntuación Kaggle	Dia de subida	Hora de subida
Variables con un valor distinto a 0 en correlación	0.11766	01/01/2018	19:10
Variables con un valor mayor a 0.3 en correlación	0.12969	01/01/2018	20:52
Variables con un valor mayor a 0.7 en correlación	0.13813	01/01/2018	21:19
Todas las categóricas	0.11775	02/01/2018	19:00
Conjunto reducido de categóricas	0.12048	04/01/2018	16:22
0.5 de valores perdidos	0.11892	04/01/2018	16:33
0.8 de valores perdidos	0.11842	04/01/2018	16:39
Escala Log	0.11524	04/01/2018	19:23
Aumento de iteraciones	0.11435	04/01/2018	19:56
Variables(log) con un valor mayor a 0.3 en correlación	0.11551	04/01/2018	20:47

Bibliografía

Gawlik, D. (s.f.). *Kaggle*. Obtenido de https://www.kaggle.com/dgawlik/house-prices-eda