Memoria. Ames Housing Dataset

The Ames Housing Dataset consta 2,919 observaciones de ventas de casas en Ames(Iowa) entre 2006 and 2010

Este dataset consta de tres ficheros:

- un fichero llamado train que contiene las features y la variable dependiente de los datos de entrenamiento
- un fichero llamado test que contiene las festures de los datos de test
- y un fichero llamado sample_submission que contiene los valores de la variable dependiente en test

Este dataset tiene los siguientes tipos de variables

* 20 variables continuas (numericas) relativas a varias elated to varios areas dimensions

* 14 variables discretas (numericas) quantify the number of items ocurring within the house

* 23 nominal variables nominales (categoricas) identican varios tipos de viviendas, garajes, maateriales y condiciones del entorno

* 23 ordinal variables (categorical) relativas a varios elementos dentro de la propiedad

Contiene información muy detallada de algunos elementos como el garaje, el sotano el Porch, las superficies, los baños, …..

Antes de empezar a analizar este dataset nos hemos preguntado que variables pueden ser determinantes para establecer el precio de una casa y hemos fijado tres variables como principales.

* Una el tamaño, las dimensiones de la casa. En este dataset

* Otra la localizacion, tenemos variables como la zona (agricola, industrial, comercial, residencial..), el barrio,…

* Por ultimo, el estado de la casa. Hay muchas variables en el dataset que nos informan de lavivienda calidad de elementos particulares y tambien existen otras variables generales como OverallQual y OverallCond indican el estado global de la casa.

Limpieza del dataset

Eliminamos la columna Id que es como un segundo indice, introducimos un columna SalePrice en el fichero de test con valor cero, para que ambos tengan la misma dimension y podamos juntarlos en un futuro, reemplazamos las variables que empiezan por un numero.

Data Exploration

Para analizar los datos nos creamos un dataframe donde podamos ver los tipos de las distintas variables, los valores nulos y su porcentaje respecto el total, la cantidad de valores distintos y unicos y la correlacion de las features con la variable SalePrice.

Vemos que las variables que aparecen en principio mas relacionadas con SalePrice son OverallQual

y GrLivArea. Tambien vemos que hay algunas variables que tienen muchos valores nulos como MiscFeature, PoolQC, Fence, Alley.. posteriormente procederemos a tratarlos.

Vemos en el notebook las variables continuas y visualizamos sus datos respecto a la variable SalePrice

Analizamos los outlier de SalePrice y para su tratamiento tendremos en cuenta las siguiente regla.

Consideraremos **outlier**

* aquellas observaciones cuyo precio sea muy superior a su valor en funcion de sus caracteristicas generales y

* aquellas otras que tienen caracteristicas muy superiores a las que les corresponderian según el precio que tienen.

En los dos outlier que hemos eliminado se cumplia esta regla. Eran dos valores que tenian unas caracteristicas general muy elevadas y el precio, sin embargo, era muy inferior a la media de las casas de las mismas cualidades. Existen otras observaciones que de acuerdo con alguna o algunas caracteristicas podrian ser outlier pero si nos fijamos detalladamente en la mayoria de sus cualidades no difieren mucho del precio de otras a las que son similares.

Hemos dedicado parte del notebook a analizar la variable **Neigborhood** que nos parecia importante para determinar la localizacion de la vivienda. Hemos extraido informacion por barrios de las casas vendidas, la superficie que tenian y el precio que se pago por ellas. Hay cinco barrios donde se vendieron muy pocas viviendas, menos de 25, y hay tres barrio donde se vendieron mas de 100. En un principio pensamos en balancear el dataser para que no hubiese tanta diferencia entre unos barrio y otros, pero como veremos al final estas variables según los modelos de prediccion tampoco influyen mucho a la hora de predecir el precio. Respecto al precio por area la mayoria estan entre diez y veinte mil. Destacar que en el barrio mas barato ClearCr el precio es de 6.114 y en el mas caro Blueste de 32.213

Las **variables ordinales** tendremos que mapearlas de menor a mayor cualidad de la variable.

El dataset tiene **missing value**, por lo que hay que proceder a su analisis. El tratamiento seguido el siguient

* con estos valores ha sido con las variables categóricas aplicar la moda y

* en el caso de variables continuas sustituirlas por la mediana.

El garaje y el sotano son las partes de la casa que mas variables tienen para describirlar y tambien son las que generaban mas casuisticas al llevar a cabo el proceso de limpieza.

Creamos variable sinteticas como combinacion de otras variables del dataset original. El principal objetivo es crear variables que agreguen otras variables que detallan una misma cualidad. Agrupamos los baños, las superficies, las calidades de los distintos elementos, ...y vemos la correlacion de estas nuevas variables con nuestra variable SalePrice. Tambien obtenemos la edad a finales del 2010 de las variables que tienen fecha.

Las **variables nominales** las tendremos que hacer un one hot encoding para crear por cada variable tantas columnas como elementos distintos tenga la variable. Quedando marcada con un uno la cualidad que disponga cada observacion y con un 0 las restantes.

Hasta aquí nuestro dataset esta compuesto de 2917 obsservacion todas ellas numericas, despues de quitar los outliers y 252 variables despues de incluir las nuevas columnas con el one hot encoding

Luego comprobamos las **correlaciones entre las distintas variables** del dataset y eliminamos de aquellas que tienen una correlacion superior a 0,8 la que esta menos correlacionada con SalePrice. Alguna de las variables que en un principio podrian parecer que podrian determinar mas el valor de las casas son eliminadas como OverallQual, GrLivArea, LotArea, TotalArea pero veremos que como parte de variables sinteticas van a aparecer como variables destacadas. En total eliminamos 26 variables

Analizaremos la **variable dependiente** y vemos que tiene una positive-skew distribution. Tiene una larga cola a medida que aumenta el SalePrice. La media de los precios de las viviendas es 180.932 esta localizada a la derecha de la mediana 163.000 de los precios. Habría que haberla tratado la skewess con una Box-Cox transformation.

Al ver los graficos de las variable numericas vimos que el precio aumentaba mas que proprocionalmente con la mayoria de las features. Por lo tanto, hemos decidio aplicar **Polynomial Features** para ver si este tipo de variables predicen mejor el precio que las variables originales del dataset. Hemos utilizado las variables que han salido en el heatmap como las que tenian mas correlacion con la variable dependiente. Como resultado de este proceso hemos incorporado 30 variables mas, pasando de 226 a 256.

Para reducir el tamaño de nuestro dataset vamos a realizar s**elección de variables**. Hemos utilizado para esta tarea Recursive Feature Elimination despues de escalar todos los datos con MinMaxScaler Como resultado de esta operación pasamos de 256 a 37 variables

Para comprimir un poco mas los datos vamos a aplicar **Principal Componente Analysis**. Como resultado vemos que con 30 variables explicamos el 99,59% del ratio de varianza. Por lo tanto, son las variables que utilizaremos en la fase de modelado

Después de aplicar los modelos a los datos tratados, los r**esultados obtenidos** nos indican que los elementos que mas influyen en el precio de las vivienda son:

-la ContructArea que se refiere a las superficies de los distintos elementos de la casa

-los TotalPoints que recogen los elementos del dataset que detallan la calidad de la vivienda.

Las variables mas destacadas en todos los modelos son combinaciones de estos valores que hicimos con Polynominals Features, aplicando un degree de 3 que nos ha generado variables cubicas, cuadráticas y combinaciones con el resto de variables utilizadas.

Los modelos que hemos utilizado para la prediccion son la media, una regresion linear, Ridge Regularization, Elastic Net, Gradient Boost, Random Forest. Los que mejor funcionan son Ridge y Elastic Net, pero seguramente que entrenando con otros hiperparametros estos dos ultimos los resultados tendran que mejorar las cifras obtenidas por los anteriores