

UNIVERSIDAD DEL VALLE DE GUATEMALA

Data Science

Lynette García



Laboratorio 2

Algoritmos de aprendizaje de máquinas

Jonathan Álvarez, 15842

GUATEMALA, 23 de agosto de 2018

1) El análisis inicio escogiendo el dataset de train que contiene la información de las casas en venta, conjuntamente con su precio. Debido a que se busca realizar una modelo de regresión lineal de la variable que describe el precio, es necesario incluirla en el análisis para su modelación y predicción. Posteriormente se dividió este dataset en 2 partes, una con el 60% de los datos para el entrenamiento y el 40% restante para el test.

2) Para obtener el modelo de regresión lineal, se realizó un análisis con todas las variables numéricas del conjunto de datos. Cotejando la significancia que tiene cada una de las variables en la variable *SalePrice*, que representa el precio de venta de la casa. En donde se obtuvo que:

Significancia	Variables
0.0001	MSSubClass OverallQual OverallCond YearBuilt MasVnrArea X1stFlrSF X2ndFlrSF BsmtFullBath BedroomAbvGr GarageCars
0.001	LotFrontage KitchenAbvGr
0.05	TotRmsAbvGrd WoodDeckSF

Por lo que en el modelo se incluyen las variables mencionadas en la tabla de significancia, debido a que son las variables que afectan directamente al precio de la casa.

3) Se realizó la selección de las variables significativas, eliminando las que no son significativas para el modelo del dataset. En el modelo generado se obtuvo un coeficiente *Multiple R-squared* de 0.81 y *Adjusted R-squared* de 0.8091, lo que nos indica que el modelo presenta ser representativo para la muestra. Debido a que no solo el *R-squared* es cercano a 1, si no que también el *Adjusted R-squared* presenta ser representativo.

```

Call:
lm(formula = SalePrice ~ ., data = trainHomeI)

Residuals:
    Min       1Q   Median       3Q      Max
-403671 -18833   -1435   14937  297280

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.377e+05  1.187e+05  -7.058 4.08e-12 ***
MSSubClass  -1.972e+02  3.835e+01  -5.143 3.52e-07 ***
OverallQual   1.913e+04  1.622e+03  11.792 < 2e-16 ***
OverallCond   4.965e+03  1.309e+03   3.793 0.000162 ***
YearBuilt     3.989e+02  6.060e+01   6.583 9.06e-11 ***
MasVnrArea    3.776e+01  8.101e+00   4.661 3.76e-06 ***
X1stFlrSF     6.089e+01  6.293e+00   9.676 < 2e-16 ***
X2ndFlrSF     5.677e+01  5.686e+00   9.984 < 2e-16 ***
BsmtFullBath   1.710e+04  2.674e+03   6.393 2.97e-10 ***
BedroomAbvGr  -9.531e+03  2.323e+03  -4.103 4.56e-05 ***
GarageCars     1.154e+04  2.263e+03   5.102 4.33e-07 ***
LotFrontage  -1.106e+02  6.562e+01  -1.685 0.092370 .
KitchenAbvGr  -1.065e+04  7.213e+03  -1.476 0.140327 .
TotRmsAbvGrd   3.081e+03  1.699e+03   1.814 0.070171 .
WoodDeckSF     2.783e+01  1.172e+01   2.373 0.017899 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34800 on 699 degrees of freedom
(162 observations deleted due to missingness)
Multiple R-squared:  0.8128,    Adjusted R-squared:  0.8091
F-statistic: 216.8 on 14 and 699 DF,  p-value: < 2.2e-16

```

Imagen No.1: Salida de consola con la información de la Regresión Lineal Múltiple.

4) Al obtener el precio estimado y guardarlo en la variable *PricePred*, se realizó el cálculo entre la diferencia del *SalePrice* y el *PricePred*. Para observar la efectividad del modelo obtenido con los datos de entrenamiento.

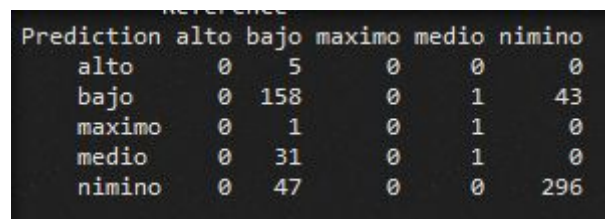
El resultado obtenido, no es precisamente el deseado. Debido a que el modelo presenta variedad en el cálculo de correcto del precio de las casas. En pocos casos, el precio estimado es bastante acertado al precio de venta de la casa, pero en la mayoría de los casos este precio tiene una diferencia alta. Por lo que se debe realizar un análisis sobre las diferencias entre el estimado y el real.

Es importante resaltar, que en este caso, una diferencia de \$10,000 puede no ser significativa, por lo que establecer una diferencia mínima de \$15,000 puede ser acertado. Debido a las cantidades que se manejan en las ventas de las casas. Por lo que puede que el modelo encontrado pueda ser representativo. Sin embargo, no cabe duda que las variables cualitativas, que se omitieron en el análisis, tienen algún peso en el cálculo del precio y deben ser tomadas para la realización de un mejor modelo.

5) Para la elección del parámetro k , se consideró utilizar un $k=23$, el cual se obtuvo de la raíz cuadrada del total de registros en el dataset de entrenamiento. Debido a que son muchos datos, comparar cada elemento con los 23 elementos más cercanos podría ser una buena opción para realizar la clasificación. Otra opción para seleccionar el valor de k es el de elegir un valor aleatorio que pueda ser representativo de la cantidad de categorías que se espera obtener.

6) Al realizar la comparación con los valores del precio que ya conocemos, debemos tomar en cuenta que ahora estamos comparando categorías y no valores numéricos. Por lo que una diferencia mínima no hace sentido.

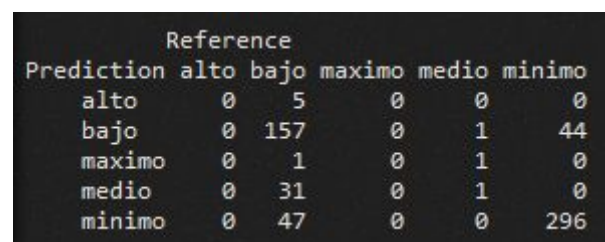
En los datos, se encontró que el algoritmo del KNN clasificó correctamente el 77% de los datos. Lo que significa que colocó correctamente el 77% de las casas en las categorías que se esperaban. El valor del *Accuracy* del modelo es de 0.78 y la matriz de confusión es la siguiente:



Prediction	alto	bajo	maximo	medio	minimo
alto	0	5	0	0	0
bajo	0	158	0	1	43
maximo	0	1	0	1	0
medio	0	31	0	1	0
minimo	0	47	0	0	296

Imagen No.2: Salida de consola con matriz de confusión del KNN.

7) Al realizar los modelos utilizando la validación cruzada, se encontró con un modelo bastante parecido al realizado sin la validación cruzada. La diferencia entre la clasificación de los precios realmente no es importante, debido a que en los métodos se obtuvo un *Accuracy* del 77% y la matriz de confusión es básicamente la misma:



Prediction	alto	bajo	maximo	medio	minimo
alto	0	5	0	0	0
bajo	0	157	0	1	44
maximo	0	1	0	1	0
medio	0	31	0	1	0
minimo	0	47	0	0	296

Imagen No.3: Salida de consola con matriz de confusión del KNN utilizando validación cruzada.

8) Al comparar los resultados de los métodos utilizados para estimar el precio de las casas en ventas, podemos decir que la Regresión Lineal Múltiple es la mejor solución para este problema. Debido a que nos proporciona un mejor acercamiento a la solución, otorgándonos un valor numérico de referencia para el precio estimado. Mientras que el método de KNN solo nos brinda una clasificación de la casa.