Universidad del Valle de Guatemala
Minería de Datos - Sección 10
María Fernanda Estrada 14198
Christopher Kevin Sandoval 13660
20/03/2020

Hoja de Trabajo 4

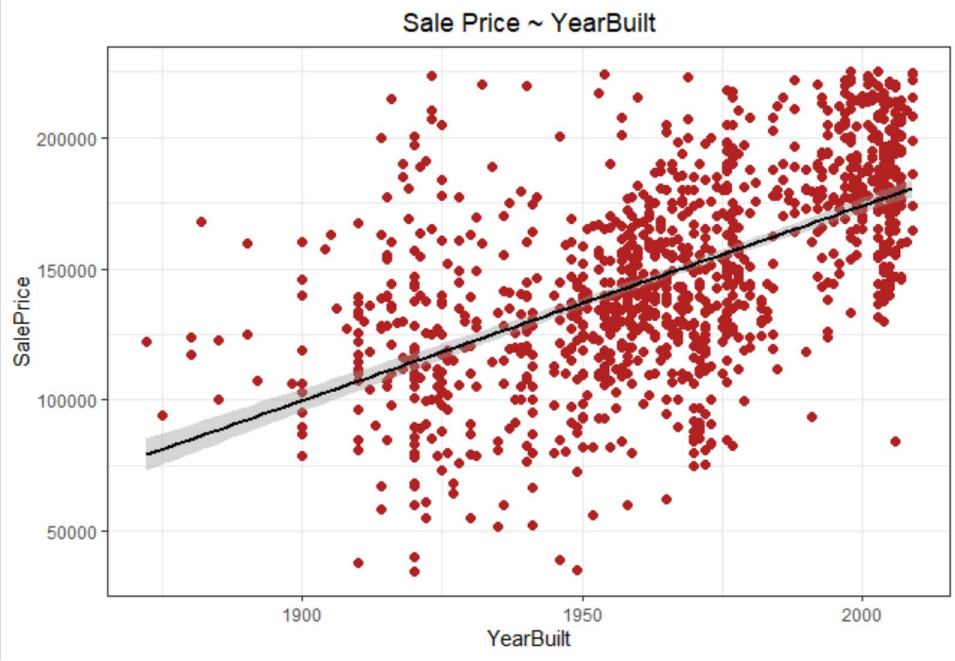
Análisis del modelo

Para determinar correctamente qué variables utilizar en el modelo, se procedió primero a ver la correlación entre variables. Se eliminaron las que tenían mucha correlación entre ellas y las que tenían poca significancia en el modelo. Las variables que quedaron son: MSSubClasss, OverallCond, YearBuilt, BsmtFinSF1, X2ndFlrSF, BsmtFullBath, BedroomAbvGr y SceenPorch.
Al aplicar el modelo de regresión lineal al conjunto de datos, se obtuvo la siguiente información:

```
> summary(fitLMPW)

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

Residuals:
    Min      1Q  Median      3Q     Max
-103768  -16882   -1826   15438   85795

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.553e+06  5.721e+04 -27.149  < 2e-16 ***
MSSubClass  -1.305e+02  1.887e+01  -6.916 7.75e-12 ***
OverallCond  5.995e+03  7.048e+02   8.507  < 2e-16 ***
YearBuilt    8.336e+02  2.835e+01  29.403  < 2e-16 ***
BsmtFinSF1   9.548e+00  2.448e+00   3.901 0.000102 ***
X2ndFlrSF    3.240e+01  2.564e+00  12.639  < 2e-16 ***
BsmtFullBath 7.264e+03  1.893e+03   3.838 0.000131 ***
BedroomAbvGr 7.234e+03  1.099e+03   6.584 7.02e-11 ***
ScreenPorch  8.229e+01  1.586e+01   5.188 2.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25830 on 1132 degrees of freedom
Multiple R-squared:  0.5353,    Adjusted R-squared:  0.532
F-statistic:   163 on 8 and 1132 DF,  p-value: < 2.2e-16
```

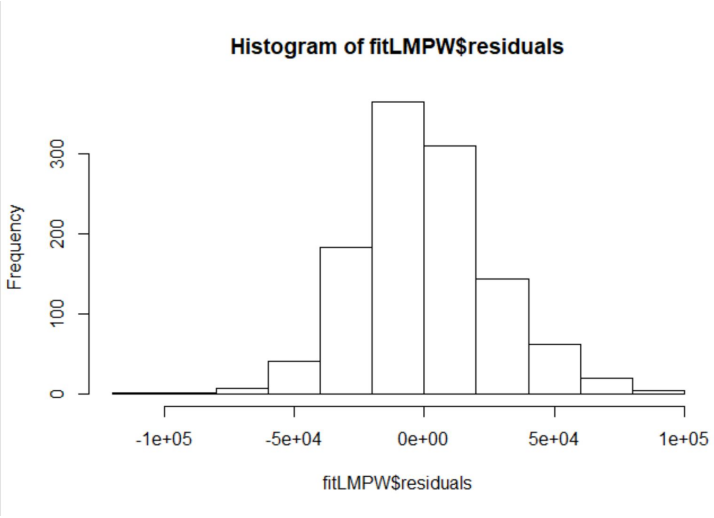
Según esta información del modelo, el nivel de significancia de todas las variables es 0 y se tiene un R^2 de 0.53, el cual es bastante bajo. Sin embargo, al tener un R^2 más elevado y con más variables, sucedió el caso de Overfitting.

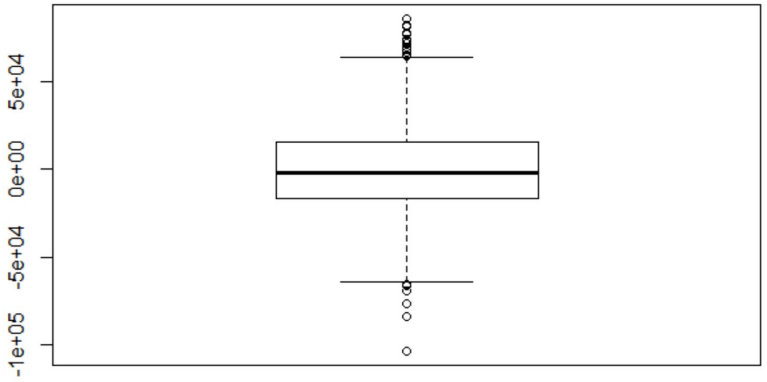
Como la gráfica del modelo solo soporta dos variables, se fue probando una por una y observando que tuvieran una relación lineal con el precio de la casa. Algunos ejemplos son:



El análisis de residuos nos indica qué tan bueno será el modelo para predecir futuros datos. Se obtuvieron los siguientes resultados de residuos, indicando que el modelo predecirá bastante bien porque los datos son aleatorios:

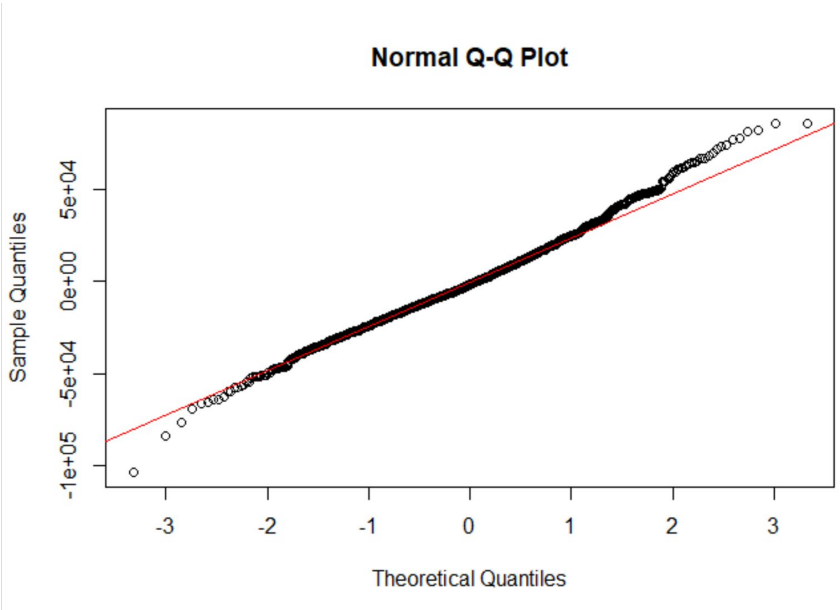
```
> head(fitLMPW$residuals)
        1         2         3         4         6         8
6422.573  11063.484  24801.536  20440.953 -28495.049  11294.902
```

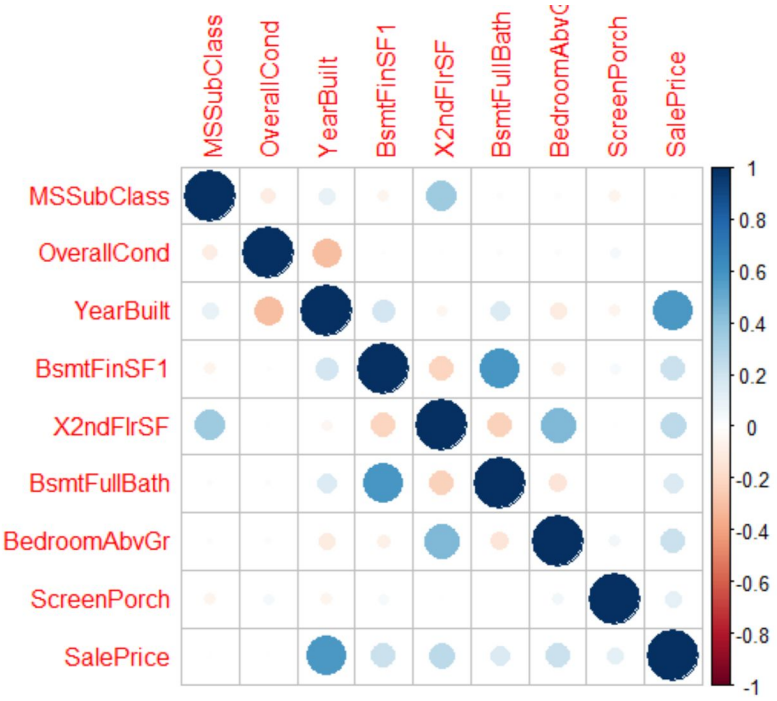




Análisis de las variables

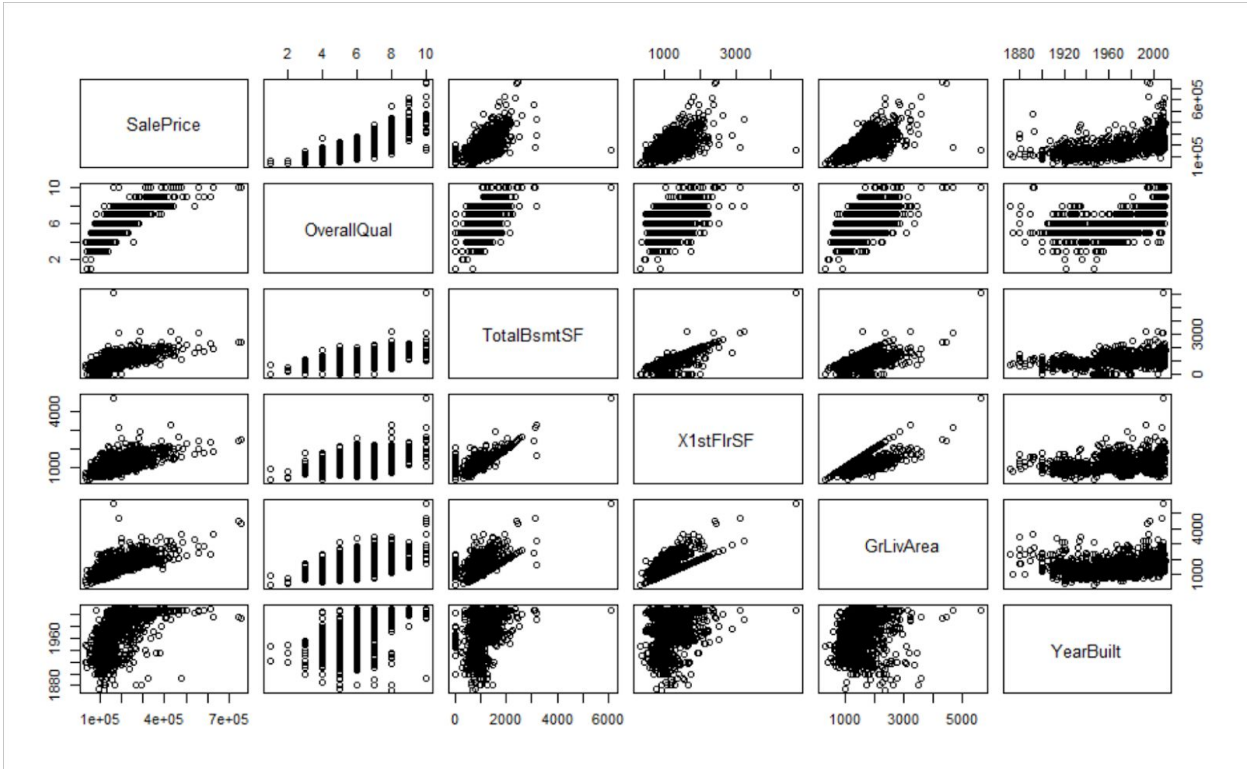
Las pruebas de normalidad sobre los datos produjeron las siguientes gráficas. Primero, se observa que las variables son normales.





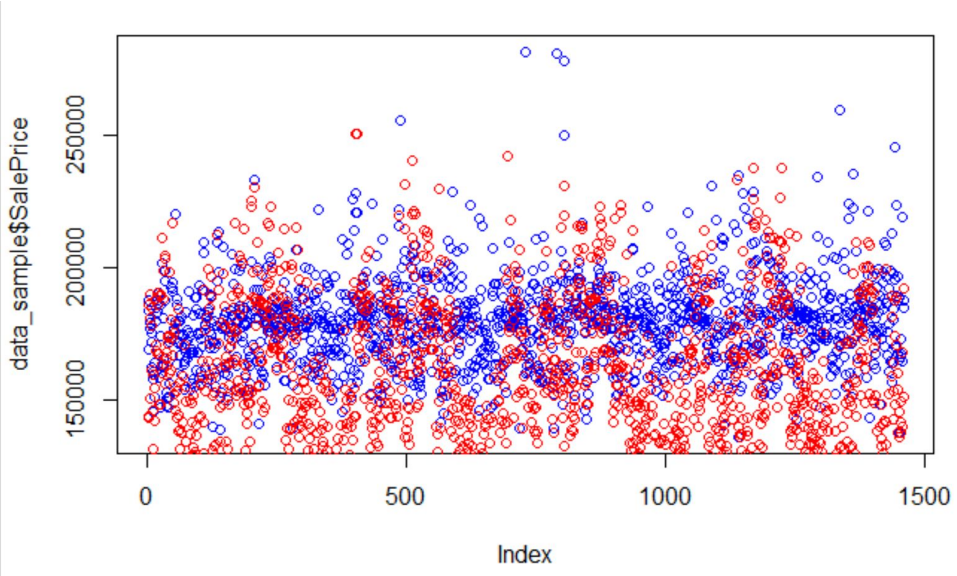
Observamos que las relaciones no presentan relación entre ellas, indicando que no afectará el modelo.

Ahora revisamos que no exista multicolinealidad entre las variables. Se observa en la gráfica siguiente que todas presentan relación lineal y que todas aportan al modelo.



Aplicación del modelo

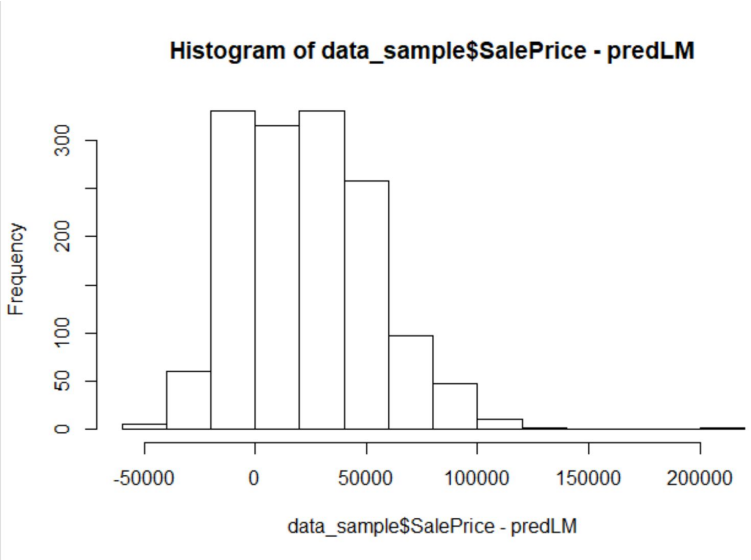
Al aplicar el modelo al conjunto de prueba, se obtuvieron los siguientes resultados. Los puntos azules son los datos reales y los puntos rojos son la predicción del modelo.



Resultados obtenidos

La gráfica anterior muestra que, aunque la predicción tiene más datos por debajo de los reales, el patrón se mantiene y sigue la misma tendencia. Además, con un histograma de la diferencia de valores, se muestra que la mayoría está cercana a 0. Esto quiere decir que la eficiencia del algoritmo/modelo fue bastante buena.

```
> summary(data_sample$SalePrice-predLM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 -49467   -1591   21922   23168   42867  201906       2
```



Comparación de métodos

Si se compara la eficiencia de ambos algoritmos, el modelo de árbol de regresión fue menos eficiente que el de regresión. El porcentaje de error en el árbol de regresión fue mayor (28%) que el error en este modelo (16%). Además, consistía de menos variables para predecir, haciéndolo más eficiente y menos complejo. En general, ambos se tardaron lo mismo en procesar, pero la cantidad de pasos es mayor en el árbol de decisión que en la regresión lineal. Se concluye que entre los dos modelos, el de regresión lineal es mejor.