# PRONÓSTICO DE VALOR DE COMPRAVENTA PARA VIVIENDAS MEDIANTE METODOLOGÍAS DE ML

Alejandro Cristancho, Cristian Castaño, Diego Agudelo
*Facultad de Ingeniería, Universidad de Antioquia*
*Medellín, Colombia*

***Abstract*—**

The sale of housing and the demand for it increases every year due to several factors, including even population growth. Forecasting the value of prices is relevant, since by not having a base measure you can lose money by buying or selling lower prices. This article intends to describe how the use of forecasting techniques are suitable to solve this forecasting problem, such as multiple regression, KNN, artificial neural networks, random forests and support vector machines (SVM), as well as linear kernel and RBF. It should be noted that the data set used by Kaggle.com contains 79 properties between 1872 and 2010, in which the reasons why one house has a higher value than another are described. Similarly, the methodology to determine the best model to use will be based on error metrics such as MAPE, RMSE and MSE.

***Resumen*—**

La compraventa de vivienda y la demanda de la misma aumentan cada año debido a varios factores, incluyendo incluso el crecimiento demográfico. El pronosticar el valor de los predios resulta relevante, ya que al no tener una medida base se puede perder dinero al comprar o vender a precios menores. Este artículo pretende describir cómo el uso de técnicas de pronóstico son adecuadas para resolver este problema de predicción, como regresión múltiple, KNN, redes neuronales artificiales, bosques aleatorios y máquinas de vectores de soporte (SVM), así como kernel lineal y RBF. Cabe destacar que el conjunto de datos usado por Kaggle.com contiene 79 propiedades entre 1872 y 2010, en estas se describen las razones del porqué una casa tiene mayor valor que otra. De igual manera, la metodología para determinar el mejor modelo a emplear se basará en métricas como el error porcentual absoluto medio y el error cuadrático medio.

***Palabras claves* — Machine learning, modelo, error, viviendas, EDA**

Precios de la vivienda: técnicas de regresión avanzadas, en el mundo inmobiliario es sumamente importante el precio de las viviendas, conocer cuales son los precios actuales y sobre todo saber cómo serán en un futuro. Es sumamente relevante poder predecir el precio de venta de una propiedad, esto teniendo en cuenta las características esenciales que elevan el valor comercial y la aceptación de una población específica (tamaño de lote, zonificación, etc).

Claramente, dando un enfoque en un campo de acción asociado al problema con la venta de bienes raíces en alguna empresa. Este problema corresponde a un modelo de regresión debido a que, teniendo en cuenta estas características de interés de los usuarios se pretende predecir el valor de la vivienda.

## II. PRE-ANÁLISIS y PRE-PROCESAMIENTO

El estudio realizado ha logrado consolidar 80 variables explicativas recolectadas entre 1872 y 2010, estas pretenden describir la mayoría de aspectos para domicilios residenciales en Ames, Iowa, Estado Unidos. De igual manera, cabe tener en cuenta que se encuentran catalogadas entre categóricas y numéricas de la siguiente manera.

### 2.1.1 Variables categóricas

Entiéndase como variables categóricas, aquellas que, por lo general, resultan ser variables que toman como valores cualidades o categorías.

| Nombre de la variable | Descripción | Valores |
|---|---|---|
| 'MSZoning' | Identifica la clasificación general de zonificación de la venta | A, C, FV, I, RH, RL, RP, RM |
| 'Street' | Tipo de acceso por carretera a la propiedad | Grvl, Pave |
| 'Alley' | Tipo de callejón de acceso a la propiedad | Grvl, Pave, NA |
| 'LotShape' | Forma general de propiedad | Reg, IR1, IR2, IR3 |
| 'LandContour' | Planitud del inmueble | Lvl, Bnk, HLS, Low |

| | | |
|---|---|---|
| ’Utilities’ | Tipo de servicios disponibles. | AllPub, NoSewr, NoSeWa, ELO |
| ’LotConfig’: | Configuración del lote. | Inside, Corner, CulDSac, FR2, FR3 |
| ’LandSlope’ | Pendiente de la propiedad. | Gtl, Mod, Sev |
| ’Neighborhood’: | Ubicaciones físicas dentro de los límites de la ciudad de Ames. | Blmngtn, Blueste, BrDale, Brk Side, ClearCr, CollgCr, Crawfor, Edwards, Gilbert, IDOTRR, MeadowV, Mitchel, Names, NoRidge, NPkVill, NridgHt, NWAmes, OldTown, SWISU, Sawyer, SawyerW, Somerst, StoneBr, Timber, Veenker) |
| ’Condition1’ | Proximidad a varias condiciones. | Artery, Feedr, Norm, RRNn, RRAn, PosN, PosA, RRNe, RRAe |
| ’Condition2’ | Proximidad a varias condiciones (Si hay más de una). | Artery, Feedr, Norm, RRNn, RRAn, PosN, PosA, RRNe, RRAe |
| ’BldgType’ | Tipo de vivienda. | 1Fam, 2FmCon, Duplx, TwnhsE, TwnhsI |
| ’HouseStyle’ | Estilo de vivienda. | 1Story, 1.5Fin, 1.5Unf, 2Story, 2.5Fin, 2.5Unf, SFoyer, SLvl) |

| | | |
|---|---|---|
| ’RoofStyle’ | Tipo del techo | Flat, Gable, Gambrel, Hip, Mansard, Shed |
| ’RoofMatl’ | Material del techo. | ClyTile, CompShg, Membran, Metal, Roll, Tar&Grv, Wd Shake, WdShngl |
| ’Exterior1st’ | Revestimiento exterior en casa. | AsbShng, AsphShn, BrkComm, BrkFace, CBlock, CemntBd, HdBoard, ImStucc, MetalSd, Other, Plywood, PreCast, Stone, Stucco, VinylSd, Wd Sdng, WdShing |
| ’Exterior2nd’ | Revestimiento exterior en casa (Si hay más de uno). | AsbShng, AsphShn, BrkComm, BrkFace, CBlock, CemntBd, HdBoard, ImStucc, MetalSd, Other, Plywood, PreCast, Stone, Stucco, VinylSd, Wd Sdng, WdShing |
| ’MasVnrType’ | Tipo de chapa de albañilería. | BrkCmn, BrkFace, CBlock, None, Stone |
| ’ExterQual’ | Evalúa la calidad del material en el exterior. | Ex, Gd, TA, Fa, Po |
| ’ExterCond’ | Evalúa la condición actual del material en el exterior. | Ex, Gd, TA, Fa, Po |
| ’Foundation’ | Tipo de base. | BrkTil, CBlock, PConc, Slab, Stone, Wood |

| | | |
|---|---|---|
| ’BsmtQual’ | Evalúa la altura del sótano. | Ex, Gd, TA, Fa, Po, NA |
| ’BsmtCond’ | Evalúa el estado general del sótano. | Ex, Gd, TA, Fa, Po, NA |
| ’BsmtExposure’ | Se refiere a huelga o paredes a nivel del jardín. | Gd, Av, Mn, No, NA |
| ’BsmtFinType1’ | Calificación del área terminada del sótano. | GLQ, ALQ, BLQ, Rec, LwQ, Unf, NA |
| ’BsmtFinType2’ | Calificación del área terminada del sótano (Si hay múltiples tipos). | GLQ, ALQ, BLQ, Rec, LwQ, Unf, NA |
| ’Heating’ | Tipo de calefacción. | Floor, GasA, GasW, Grav, OthW, Wall |
| ’HeatingQC’ | Calidad y estado de la calefacción. | Ex, Gd, TA, Fa, Po |
| ’CentralAir’ | Aire acondicionado central. | N, Y |
| ’Electrical’ | Sistema eléctrico. | SBrkr, FuseA, FuseF, FuseP, Mix |
| ’KitchenQual’ | Calidad de la cocina. | Ex, Gd, TA, Fa, Po |
| ’Functional’ | Funcionalidad del hogar (se supone que es típica a menos que las deducciones estén garantizadas). | Typ, Min1, Min2, Mod, Maj1, Maj2, Sev, Sal) |
| ’FireplaceQu’ | Calidad de la chimenea. | Ex, Gd, TA, Fa, Po, NA |
| ’GarageType’ | Localidad del garaje. | 2Types, Attchd, Basment, BuiltIn, CarPort, Detchd, NA |
| ’GarageFinish’ | Acabado interior del garaje. | Fin, RFn, Unf, NA |
| ’GarageQual’ | Calidad del garaje. | Ex, Gd, TA, Fa, Po, NA |
| ’GarageCond’ | Condición del garaje. | Ex, Gd, TA, Fa, Po, NA |
| ’PavedDrive’ | Camino pavimentado | Y, P, N |
| ’PoolQC’ | Calidad de la piscina. | Ex, Gd, TA, Fa, NA |
| ’Fence’ | Calidad de la cerca. | GdPrv, MnPrv, GdWo, MnWw, NA |
| ’MiscFeature’ | Caracterısticas varias no cubiertas en otras categorías. | Elev, Gar2, Othr, Shed, TenC, NA |
| ’SaleType’ | Tipo de venta. | WD, CWD, VWD, New, COD, Con, ConLw, ConLI, ConLD, Oth |
| ’SaleCondition’ | Condición de venta. | Normal, Abnorml, AdjLand, Alloca, Family, Partial |

2.1.2 Variables numéricas

Cada una de las variables son del tipo ’int16’, ’int32’, ’int64’, ’float16’, ’float32’ y ’float64’. Una de las de mayor relevancia es la variable a predecir,’SalePice’.

| Nombre de variable | Descripción |
|---|---|
| ’MSSubClass’ | Identifica el tipo de vivienda involucrada en la venta. |
| ’LotFrontage’ | Pies lineales de calle conectados a la propiedad |
| ’LotArea’ | Tamaño del lote en pies cuadrados. |
| ’OverallQual’ | Califica el material general y el acabado de la casa. |
| ’OverallCond’ | Califica el estado general de la casa. |

| | |
|---|---|
| ’YearBuilt’ | Fecha de construcción original. |
| ’YearRemodAdd’ | Fecha de remodelación (igual que la fecha de construcción si no hay remodelaciones o adiciones) |
| ’MasVnrArea’ | Área de chapa de albañilería en pies cuadrados. |
| ’BsmtFinSF1’ | Tipo 1 terminado pies cuadrados. |
| ’BsmtFinSF2’ | Tipo 2 terminado pies cuadrados. |
| ’BsmtUnfSF’ | Pies cuadrados inacabados del área de sótano. |
| TotalBsmtSF’ | Total de pies cuadrados de área de sótano. |
| ’’1stFlrSF’ | Primer piso pies cuadrados. |
| ’2ndFlrSF’ | Segundo piso pies cuadrados |
| ’LowQualFinSF’ | Pies cuadrados terminados de baja calidad (todos los pisos). |
| ’GrLivArea’ | Superficie habitable sobre el nivel del suelo (pies cuadrados). |
| ’BsmtFullBath’ | Sótano y baños completos. |
| ’BsmtHalfBath’ | Sótano a medias baños. |
| ’FullBath’ | Baños completos sobre rasante. |
| ’HalfBath’ | Baños a medias por encima del grado. |
| ’BedroomAbvGr’: | Dormitorios sobre rasante (NO incluye dormitorios en el sótano). |
| ’KitchenAbvGr’ | Cocinas por encima del grado. |
| ’TotRmsAbvGrd’ | Total de habitaciones por encima del grado (no incluye baños). |
| ’Fireplaces’ | Número de chimeneas. |
| ’GarageYrBlt’ | Año de construcción del garaje. |
| ’GarageCars’ | Tamaño del garaje en capacidad para automóviles. |
| ’GarageArea’ | Tamaño del garaje en pies cuadrados |
| ’WoodDeckSF’ | Área de cubierta de madera en pies cuadrados |
| ’OpenPorchSF’ | Porche abierto en pies cuadrados. |
| ’EnclosedPorch’ | Porche cerrado en pies cuadrados. |
| ’3SsnPorch’ | Área de porche de tres estaciones en pies cuadrados. |
| ’ScreenPorch’ | Pantalla del porche en pies cuadrados. |
| ’PoolArea’ | Área de piscina en pies cuadrados. |
| ’MiscVal’ | Valor de la característica miscelánea |
| ’MoSold’ | Mes vendido (MM). |
| ’YrSold’ | Año vendido (AAAA). |
| ’SalePrice’ | Variable a predecir. Precio de venta. |

## III. ARTÍCULOS RELACIONADOS

Los siguientes artículos sirven de guías para implementar y comparar los resultados obtenidos.

| Artículo | Referencia |
|---|---|
| Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning | [4] |
| A Hybrid Regression Technique for House Prices Prediction | [5] |
| Applied Machine Learning Project 4 Prediction of real property prices in Montreal | [6] |
| Machine Learning based Predicting House Prices using Regression Techniques | [7] |

**Técnicas de aprendizaje usada en estos artículos**

- Lasso Regression
- Elastic Net Regression
- Ridge Regression
- Multilayer Perceptron Regressor
- Regression Trees
- Random Forest
- Gradient Boosting
- Linear Regression
- SVR - Support Vector Regression -
- XGBoost Regression Model

**Metodología de validación usada**

Se usa principalmente Cross Validation, sin embargo, en algunos artículos se evidencia variaciones y estrategias distintas de su implementación.

**Resultados obtenidos en cada uno de los artículos**

1. Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning [4]

| Algorithm | CV score | CV score standard deviation |
|---|---|---|
| Lasso | 0.11139 | 0.0106 |
| XGBoost | 0.13058 | 0.0108 |
| XGBoost with logit transform | 0.12986 | 0.0107 |
| ElasticNet | 0.11203 | 0.0107 |
| Neural network | 0.11787 | 0.0095 |
| Neural network machine | 0.11695 | 0.0096 |
| Ensemble of lasso, XGBoost, ElasticNet and Neural network machine | 0.11125 | 0.0109 |
| Ensemble + residual regresor | 0.11108 | 0.0120 |
| Ensemble + ensemble of residual regressors | 0.11093 | 0.0120 |
| Full solution without polynomial features | 0.11127 | 0.0114 |

2. A Hybrid Regression Technique for House Prices Prediction [5]

- Ridge Regression: usando selección de características.

| No of Features | Alpha | RMSE | Score |
|---|---|---|---|
| 160 | 13 | 0.11222276 | 0.11638 |
| 230 | 18 | 0.113627 | 0.11558 |
| 280 | 20 | 0.114547 | 0.11583 |

- Lasso Regretion: usando selección de características.

| No of features | Alpha | RMSE | Score |
|---|---|---|---|
| 160 | 1.55E-04 | 0.113838 | 0.11706 |
| 230 | 3.70E-04 | 0.114974 | 0.11499 |
| 280 | 5.40E-04 | 0.115464 | 0.11675 |

- Gradient: usando selección de características.

| No of Features | Subsample | Score |
|---|---|---|
| 160 | 0.6 | 0.12032 |
| 23 | 0.5 | 0.11876 |
| 230 | 0.6 | 0.11843 |
| 280 | 0.6 | 0.12057 |

3. Applied Machine Learning Project 4 Prediction of real property prices in Montreal [6]

| | LR | SVR | KNN | Random Forest | Ensemble |
|---|---|---|---|---|---|
| Error | 0.1725 | 0.1604 | 0.1103 | 0.1135 | 0.0985 |

4. Machine Learning based Predicting House Prices using Regression Techniques [7]

- Linear Regression

| Metric | Train set | Test set |
|---|---|---|
| R-square | 0.418 | -2.12 |
| RMSE | 0.0912 | 0.2077 |

| | | |
|---|---|---|
| RMSLE | 0.02755 | 0.03493 |

- Ridge Regression

| Metric | Train set | Test set |
|---|---|---|
| R-square | 0.4345 | 0.4358 |
| RMSE | 0.5415 | 0.5224 |
| RMSLE | 0.0410 | 0.040701 |

- Lasso Regression

| Metric | Train set | Test Set |
|---|---|---|
| R-square | 0.799 | 0.06630 |
| RMSLE | 0.0256 | 0.0317 |

- SVR

| Metric | Train Set | Test Set |
|---|---|---|
| R-square | 0.799 | 0.6630 |
| RMSLE | 0.0256 | 0.0317 |

- XGBoost Regression Model

| Metric | Train set | Test set |
|---|---|---|
| R-square | 0.7868 | 0.7584 |
| RMSE | 0.3309 | 0.3462 |
| RMSLE | 0.0256 | 0.0317 |

## III. EXPERIMENTOS

La base de datos utilizada para este problema es de Kaggle.com, “House Prices: Advanced Regression Techniques”, que contiene 80 variables de entrada, 43 de las cuales son variables categóricas que describen características o categorías y 37 son variables numéricas que incluye la previsión, en este caso los precios de la vivienda.

Como metodología de validación, se hace uso de bootstrapping, se emplea el 80% de los datos para el entrenamiento del modelo y el otro 20% para la validación del modelo de regresión.

**Bootstrapping**:

Es una técnica empleada para validar la efectividad de un modelo predictivo, los métodos de conjunto, la estimación del sesgo y la varianza del modelo, para este fin se puede utilizar la función “grid search cv” que implementa el método de “fit” y arroja un score, también implementa métodos com “predict”, predict proba”, “decision function”, “transform” y “inverse transform”.

Es una técnica utilizada para validar la efectividad de un modelo de pronóstico, métodos de conjunto y estimar el sesgo y la varianza del modelo. Para ello, se utiliza la función *"Grid Search CV"*, que implementa el método *"Fit"* y otorga una puntuación o *score*, así como la "predict", "predict proba", "decision function", "transform" y "inverse transform".

Para evaluar el rendimiento de los modelos se utilizaron; **MAE, MAPE, RMSE Y R2**

- **MAE:** Promedio de todos los errores absolutos.

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

En este se miden las diferencias entre dos variables continuas.

- **MAPE:** Error porcentual absoluto medio

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|A_t - F_t|}{|A_t|}}{n}$$

Es una medida de precisión de predicción de un método de pronóstico, se usa comúnmente como una función de pérdida para problemas de regresión y en la evaluación de modelos.

- **RMSE**: error cuadrático medio

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(real_i - estimado_i)^2}$$

Es de fácil interpretación y utiliza valores absolutos pequeños que facilitan los cálculos informáticos.

- **R2:** R cuadrado, se calcula usando la siguiente fórmula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Donde SS res es la suma residual de cuadrados y SS tot es la suma total de cuadrados, entre más cercano a 1 el valor de r-cuadrado, mejor es el modelo.

Para la elección de los mejores parámetros de los modelos se utilizó la función *GridSearchCV* de sklearn para los que aplicaban, recibe un conjunto de parámetros y entrega cual es la combinación que da mejores resultados. Los modelos entrenados con su respectivo resultado se describen a continuación, los algoritmos utilizados están disponibles en el siguiente enlace HousePrices

### A. *Regresión Múltiple*

Este modelo trata de simular el comportamiento de un conjunto de datos con una función, no hubo necesidad de configurar ningún parámetro, ya que no aplica. Comparado con otros se obtuvieron muy buenos resultados, un factor importante puede ser el ajuste inicial de la variable de salida para tener una distribución más uniforme. Los resultados de las medidas de evaluación se describen a continuación en la Tabla 1.

TABLA 1
RESULTADOS REGRESIÓN MÚLTIPLE

| MAE | MAPE | RMSE | R2 |
|---|---|---|---|
| 0.0916 | 0.764 | 0.121 | 0.894 |

### B. *K vecinos más cercanos*

Este modelo asume que las variables de salida se comportan en función de la similitud y el comportamiento de sus muestras "vecinas". Los parámetros "*n_neighbors"* variaron de 1 a 10 y *"algorithm"* varió entre *"ball_tree"* o *"brute"*, con 5 vecinos , ya que este resultó ser el mejor parámetro, los resultados para esta ejecución se ven reflejados en la Tabla 2.

TABLA 2
RESULTADOS KNN

| MAE | MAPE | RMSE | R2 |
|---|---|---|---|
| 0.161 | 1.345 | 0.217 | 0.498 |

### C. *Random Forest*

Este modelo consta de un conjunto de árboles de decisión basados en vectores aleatorios independientes. Los parámetros *'n_estimators'* variaron entre [20,50,100] y *'max_depth'* entre [15,18,20]. Se configuró con *max_depth* de 15 y *n_raters* de 100.

TABLA 3
RESULTADOS RANDOM FOREST

| MAE | MAPE | RMSE | R2 |
|---|---|---|---|
| 0.083 | 3.507 | 0.117 | 0.894 |

### D. *Gradient boosting*

En este modelo se usan árboles de decisión de forma escalonada. Los resultados de entrenar este modelo se describen en la tabla 4.

TABLA 4
RESULTADOS GRADIENT BOOSTING

| MAE | MAPE | RMSE | R2 |
|---|---|---|---|
| 0.0757 | 3.559 | 0.101 | 0.925 |

### E. *Regresión con vectores de soporte RBF*

Este método utiliza el algoritmo Support Vector Machine para predecir una variable continua. Se varía el valor de epsilon entre [0.004, 0.008, 0.0005, 0.0008], teniendo como mejor opción 'epsilon' igual a 0.008.

TABLA 5
RESULTADOS SVM RBF

| MAE | MAPE | RMSE | R2 |
|---|---|---|---|
| 0.095 | 3.463 | 0.159 | 0.792 |

TABLA 6
RESULTADOS UNIFICADOS

| MODELO | RMSE |
|---|---|
| Regresión Múltiple | 0.121 |
| KNN | 0.217 |
| Random Forest | 0.117 |
| Gradient Boosting | 0.101 |
| SVM RBF | 0.159 |

Con los resultados obtenidos, unificados en la Tabla 6, anteriormente descritos, se llega a la conclusión basada en que la medida de desempeño elegida es RMSE, por lo tanto, los tres mejores modelos son:

- Regresión múltiple
- Gradient Boosting
- Random Forest

REFERENCIAS


[1] K. GRACE-MARTIN, ”Outliers: To Drop or Not to Drop - The Analysis Factor”, The Analysis Factor. [Online]. Available: https://www.theanalysisfactor.com/outliers to-drop-or-not-to-drop/. [Accessed: 08- Jun 2020].

[2] M. Galarnyk, ”Understanding Box plots”, Medium, 2018. [Online]. Available: https://towardsdatascience.com/understanding boxplots-5e2df7bcbd51. [Accessed: 08- Jun 2020].

[3] ”Aprendizaje automático de métricas de regresión (MSE) - sitiobigdata.com”, sitiobigdata.com. [Online]. Available: https://sitiobigdata.com/2019/05/27/modelos de-machine-learning-metricas-de-regresion mse-parte-2/. [Accessed: 09- Jun- 2020].

[4] P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich and P. I. Vasilevna, ”Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning,” 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, 2018, pp. 1-5, doi: 10.1109/RPC.2018.8482191.

[5] S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, ”A hybrid regression technique for house prices prediction,” 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017, pp. 319-323, doi: 10.1109/IEEM.2017.8289904.

[6] Pow, N. ”Applied Machine Learning Project 4 Prediction of real estate property prices in Montreal,” 2014

[7] J. Manasa, R. Gupta and N. S. Narahari, ”Machine Learning based Predicting House Prices using Regression Techniques,” 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[8] Awad M., Khanna R. ”Support Vector Regression. In: Efficient Learning Machines”. 2015. Apress, Berkeley, CA

[9] ”House Prices: Advanced Regression Techniques — Kaggle”, Kaggle.com. [Online].

[10] Predicción de precios con regresiones lineales https://www.kaggle.com/jesuscarmona12/predicci-n-de-precios-con-regresiones-lineales

[11] House Prices Notebook por Kenny Van https://www.kaggle.com/kennyvan/house-prices-notebook

[12] House Price Prediction por Prosenjit123 https://www.kaggle.com/prosenjit123/house-price-prediction

[13] House Prices using GradientBoostingRegressor 90% https://www.kaggle.com/mountaga/house-prices-using-gradientboostingregressor