Predictive Data Model on Ames Housing Data

Maliha Gangat · Follow

Published in Analytics Vidhya · 7 min read · Oct 31, 2020

Introduction

Understanding the customer needs and predicting customer's purchase intents form the core of success for any business. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this project dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. For the vastly diversified realty market, with prices of properties increasing exponentially, it becomes essential to study the factors which affect directly or indirectly when a customer decides to buy a house and to predict the market trend. In general, for any purchase, a potential customer makes the decision based on the value for the money. **Here I need to predict the price of a house in Ames, US.** Also, this problem statement allowed me to study the Advanced regression techniques and their implementation in machine learning language like R.

Understanding the theory and implementing it into practice was a challenge for us.

Data Set

I had two different data sets, namely train and test. Both contained numerous variables in terms of features which were describing a house. Training data set contained 1460 observations for which the sale price of the homes was provided. Based on this data, a prediction model was to be built. Test data set contained 1459 observations for which the sale price was to be predicted.

In total, 80 variables focus on the quality and quantity of many physical attributes of the property. Most of the variables are precisely the type of information that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?).

Data set contained 23 nominal variables and 23 ordinal variables. Nominal Includes variables like-the weather condition and material used for construction. For the nominal and ordinal variables, the levels were in the range of 2 to 28. Total of 14 discrete variables comprises the number of kitchens, washrooms, and bedrooms. This also includes the garage capacity and construction or re-modelling dates; 20 continuous variables describe the area dimension of each observation. Lot size and total dwelling square footage are standard home listing available online. Area measurements on the basement, porches and main living area are further classified into respective classes based on quality and type.

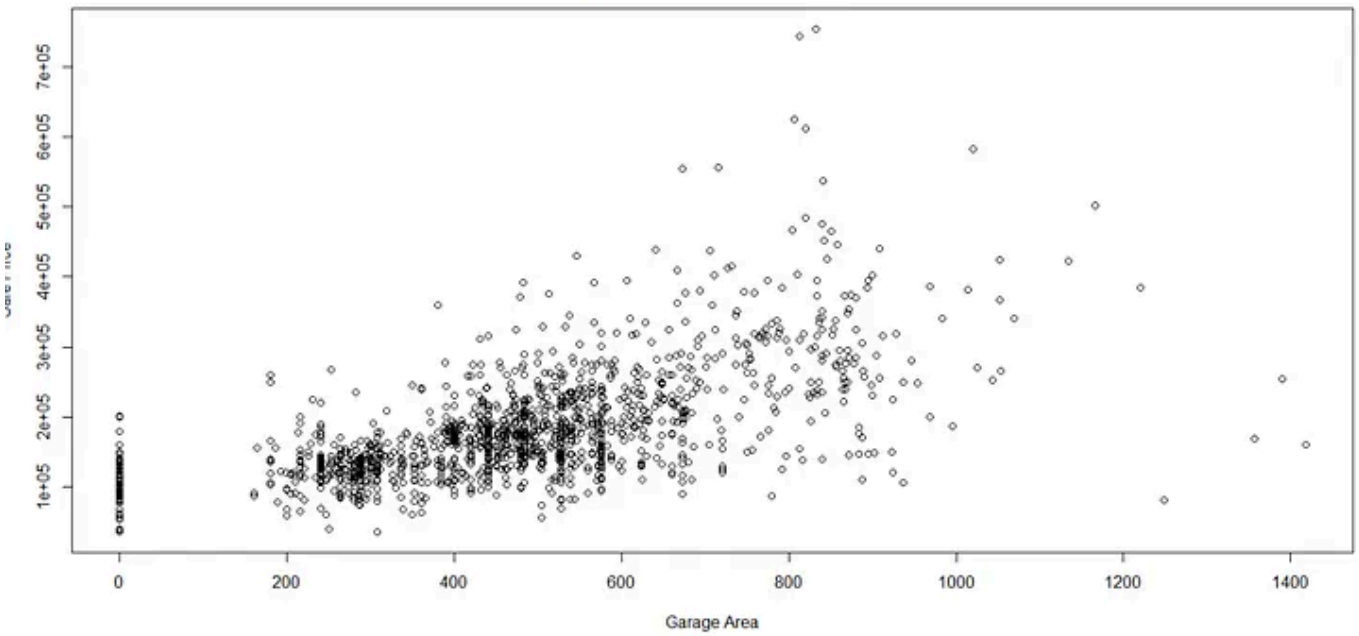
2. Steps Involved in Building Model

> *Ø Load the data train and test data set*
>
> *Ø Check for missing values. Replace missing values with the mean of numeric data and with the mode of categorical data.*
>
> *Ø Check for Outliers if any with respect to the target variable*
>
> *Ø Build the Regression Model — Target variable against independent variables*
>
> *Ø Predict the testing data set on the built model*

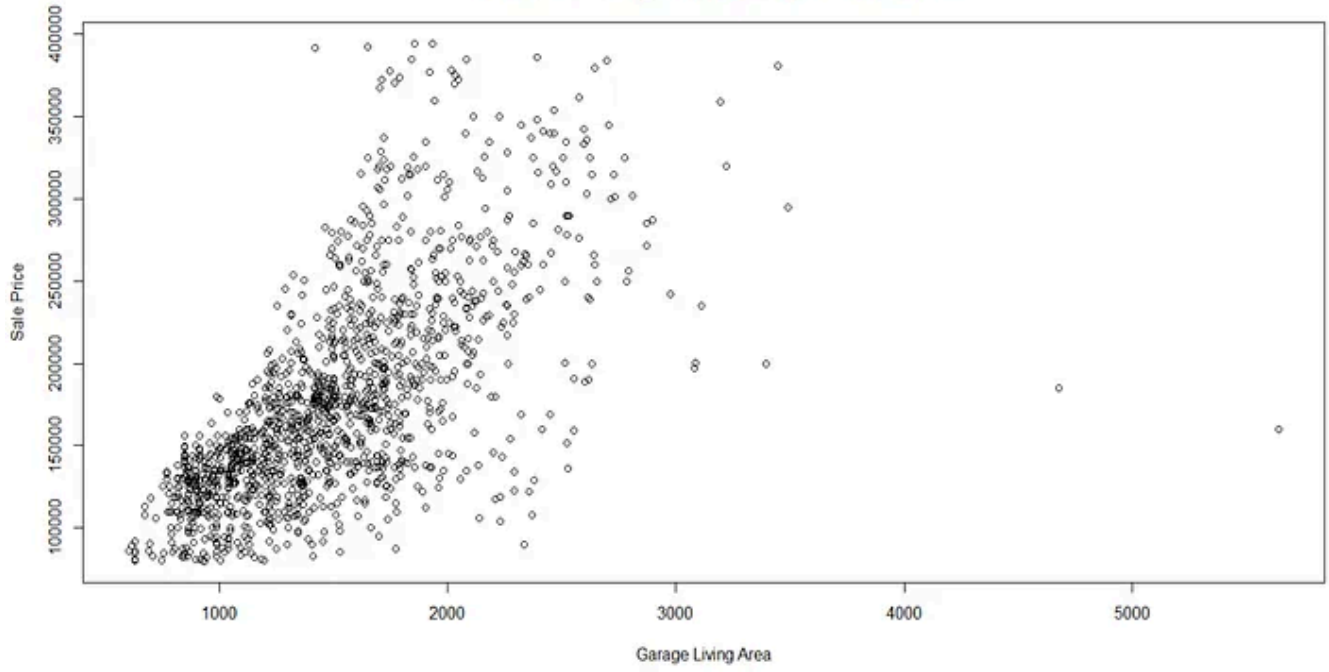
DataVisualization

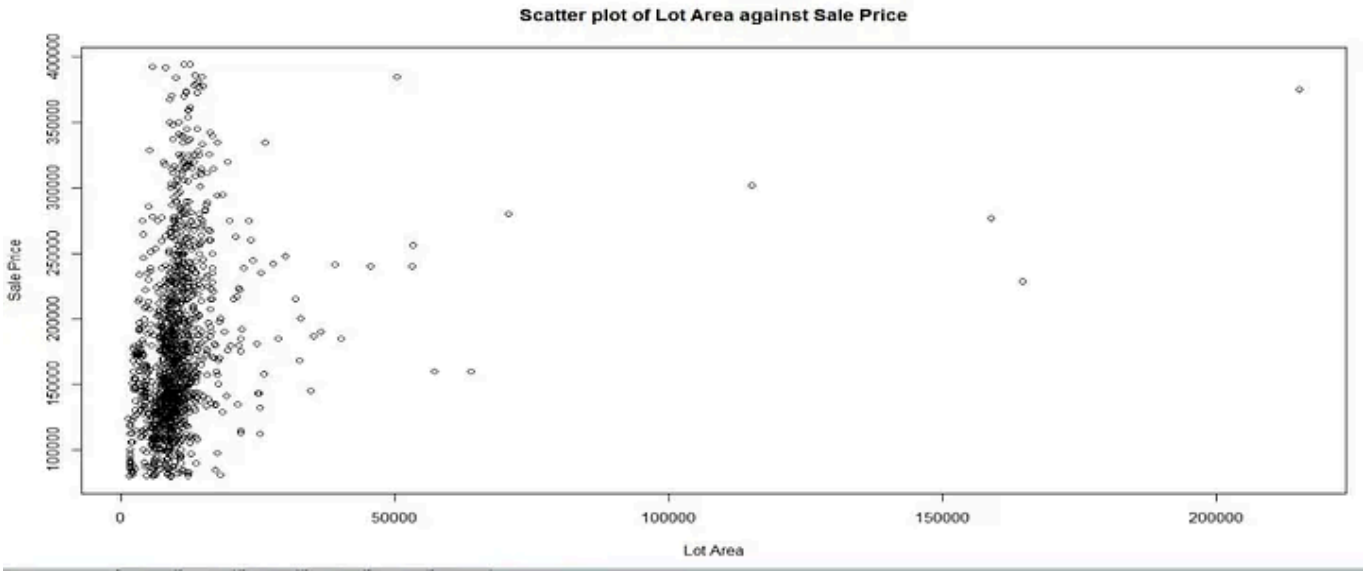
For every data analysis, initial data cleaning is required. Otherwise, outliers in the data set produce wrong prediction models, and that can seriously affect the model accuracy. Initially, I decided to analyse our data through visualisation. I compared the response variables (sale price) concerning a few important quantitative variables like ‘Garage living area’, ‘Lot Area’ etc. Scatter plots were plotted, and few outliers were observed right away.

Scatter plot of Garage Area against Sale Price

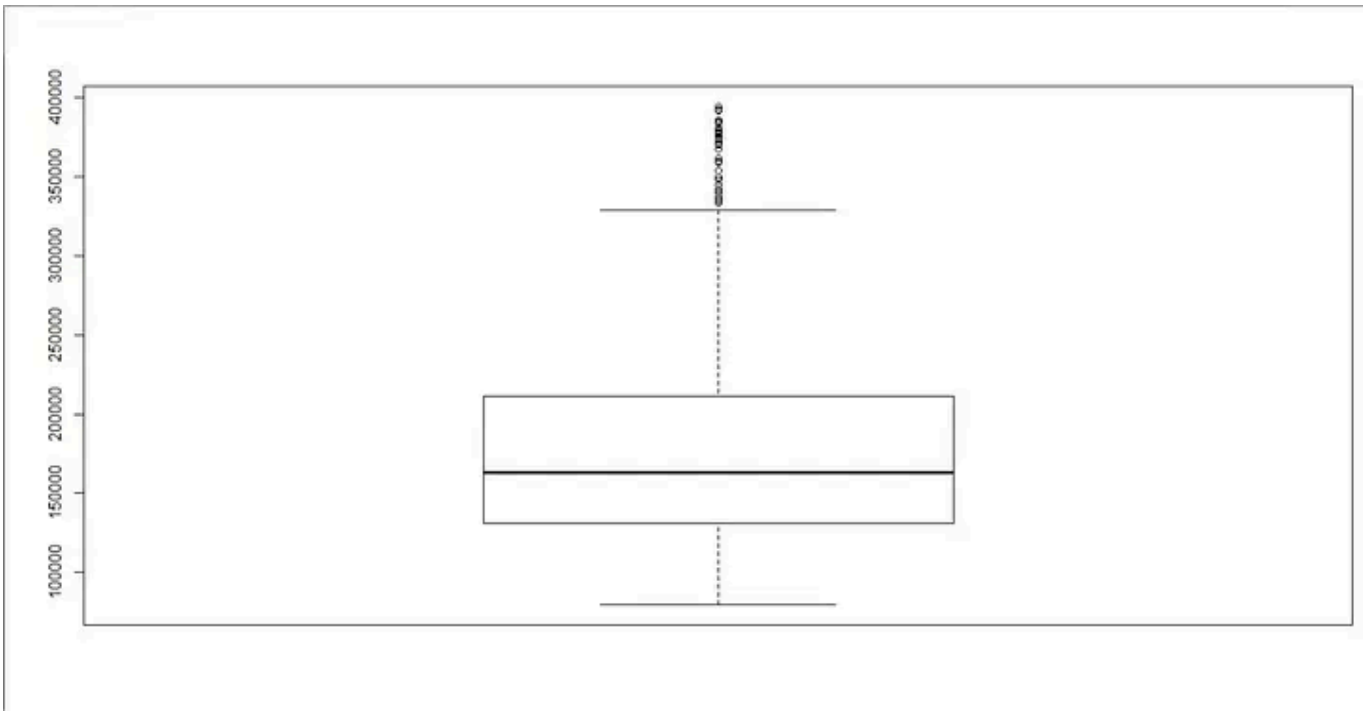


Scatter plot of Garage Living Area against Sale Price



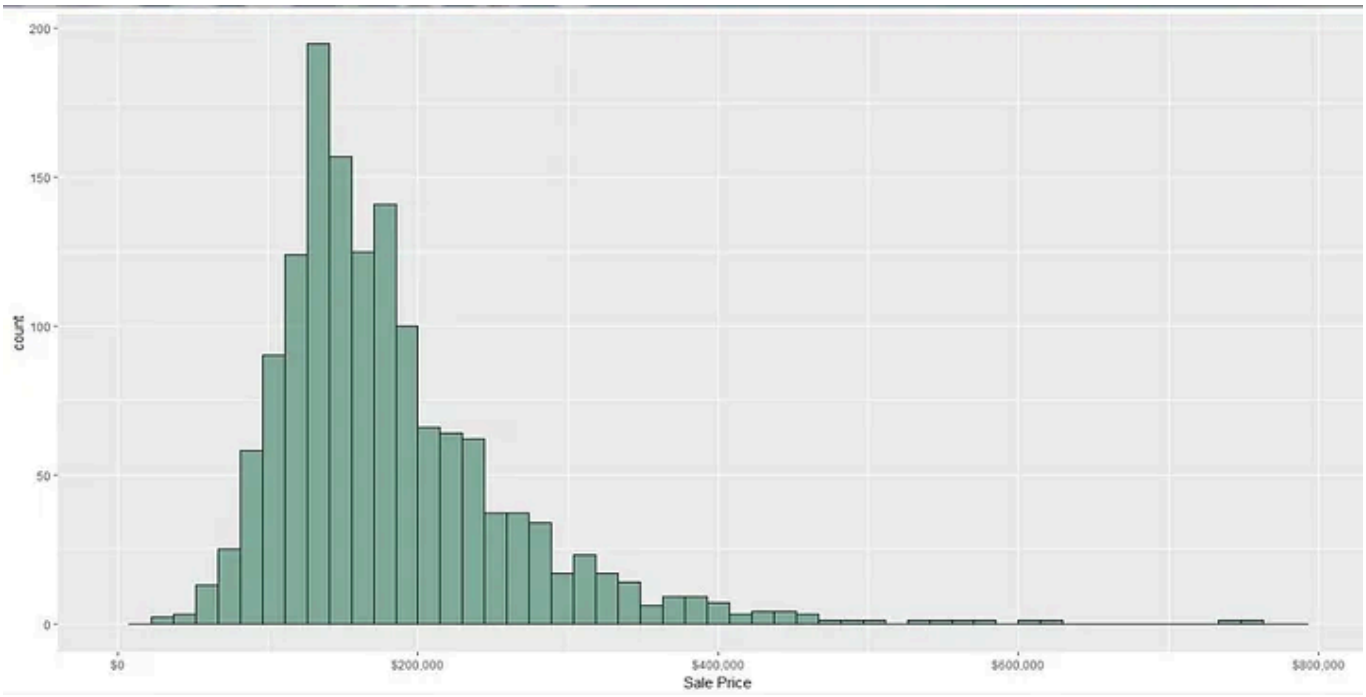


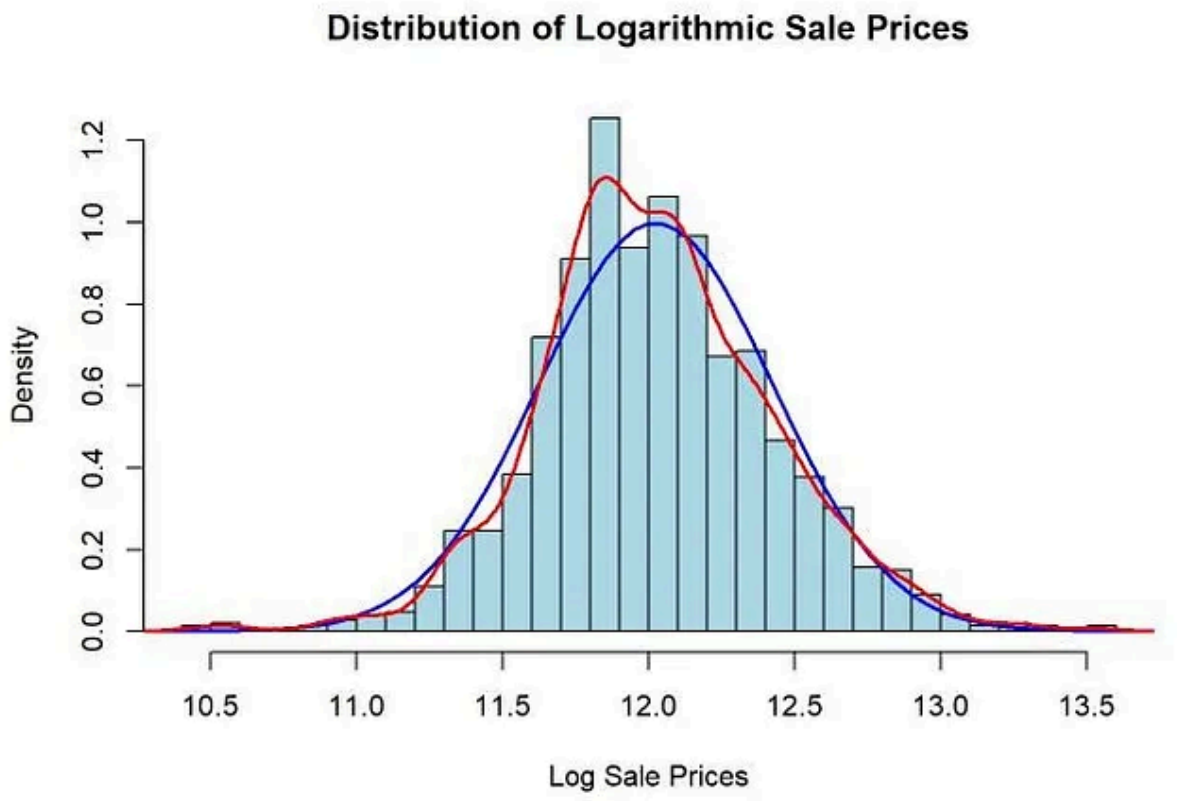
Also, the box plot helped us to visualise the range of house prices were dealing with. Most of the sales price is below $500k. All the value points above $500k (total =10) are excluded from the prediction model. Thus, outliers were removed through the initial screening. To confirm the same outliers, normality check was done in the next step.



Normality Check

All statistical methods rely on the initial assumption of data being normally distributed. Hence, before applying any statistical method, it should be confirmed whether the data is normally distributed or not. Histogram was plotted to determine the normality trend of data. The histogram was skewed as the outliers were present at the higher price range. Data points representing sale price more than $ 500K were excluded from the analysis.





Removing Outliers

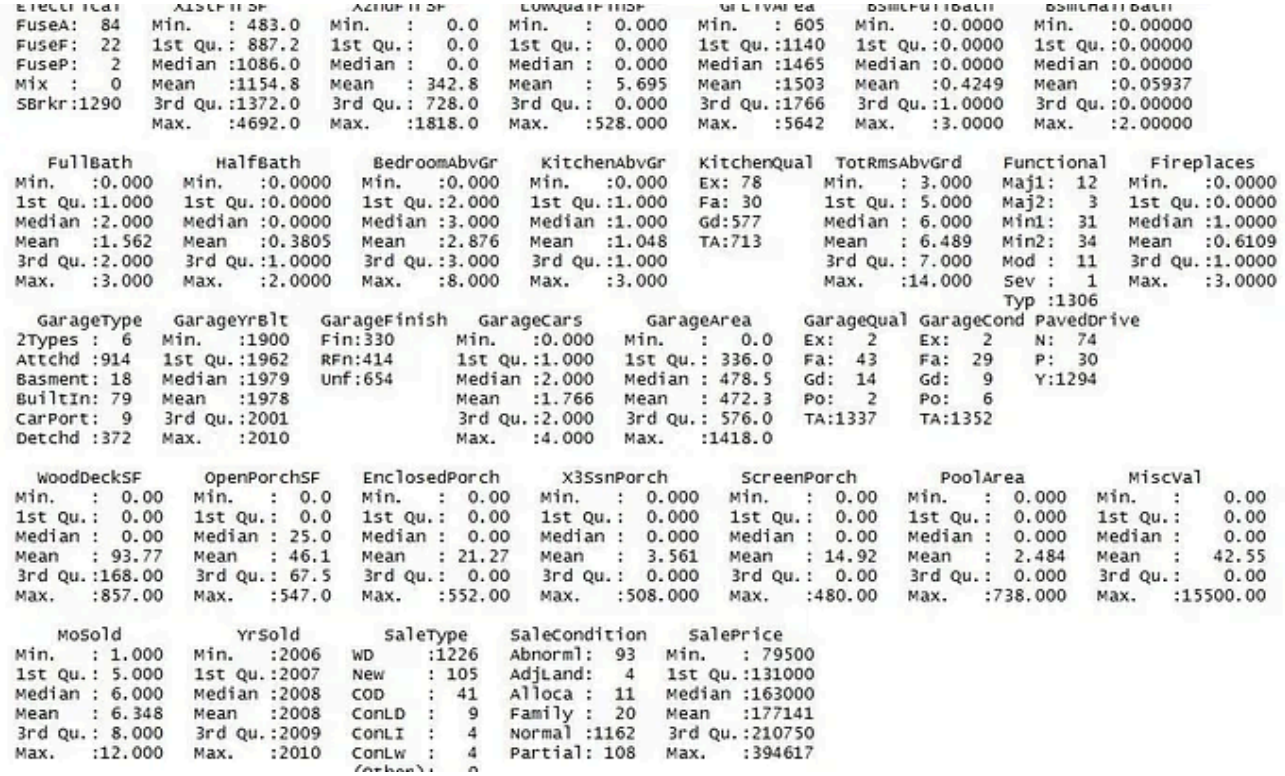
An outlier is defined as an observation which stands far away from most of the other observations. Often an outlier is present due to the measurements error. Therefore, one of the most essential tasks in data analysis is to identify and to remove the outliers as it will degrade our statistical results. I have removed outliers concerning the target variable, i.e. SalePrice and below is the summary.

```
FuseP:   3   Median :1087   Median :   0   Median :  0.000   Median :1464   Median :0.0000
Mix   :   1   Mean   :1163   Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
SBrkr:1335   3rd Qu.:1391   3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
             Max.   :4692   Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
```

```
 BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd
Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex:100       Min.   : 2.000
1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 39       1st Qu.: 5.000
Median :0.00000   Median :2.000   Median :0.0000   Median :3.000   Median :1.000   Gd:586       Median : 6.000
Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866   Mean   :1.047   TA:735       Mean   : 6.518
3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000                3rd Qu.: 7.000
Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000   Max.   :3.000                Max.   :14.000
```

```
Functional    Fireplaces      GarageType   GarageYrBlt   GarageFinish   GarageCars      GarageArea      GarageQual
Maj1:  14   Min.   :0.000   2Types :  6   Min.   :1900   Fin:352      Min.   :0.000   Min.   :   0.0   Ex:   3
Maj2:   5   1st Qu.:0.000   Attchd :951   1st Qu.:1962   RFn:422      1st Qu.:1.000   1st Qu.: 334.5   Fa:  48
Min1:  31   Median :1.000   Basment: 19   Median :1979   Unf:686      Median :2.000   Median : 480.0   Gd:  14
Min2:  34   Mean   :0.613   BuiltIn: 88   Mean   :1979                Mean   :1.767   Mean   : 473.0   Po:   3
Mod :  15   3rd Qu.:1.000   CarPort:  9   3rd Qu.:2001                3rd Qu.:2.000   3rd Qu.: 576.0   TA:1392
Sev :   1   Max.   :3.000   Detchd :387   Max.   :2010                Max.   :4.000   Max.   :1418.0
Typ :1360
GarageCond PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch     X3SsnPorch      ScreenPorch
Ex:   2    N:  90     Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
Fa:  35    P:  30     1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00
Gd:   9    Y:1340     Median :  0.00   Median : 25.00   Median :  0.00   Median :  0.00   Median :  0.00
Po:   7               Mean   : 94.24   Mean   : 46.66   Mean   : 21.95   Mean   :  3.41   Mean   : 15.06
TA:1407               3rd Qu.:168.00   3rd Qu.: 68.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00
                      Max.   :857.00   Max.   :547.00   Max.   :552.00   Max.   :508.00   Max.   :480.00
```

```
    PoolArea          MiscVal            MoSold          YrSold       SaleType    SaleCondition    SalePrice
Min.   :  0.000   Min.   :    0.00   Min.   : 1.000   Min.   :2006   WD     :1267   Abnorml: 101   Min.   : 34900
1st Qu.:  0.000   1st Qu.:    0.00   1st Qu.: 5.000   1st Qu.:2007   New    : 122   AdjLand:   4   1st Qu.:129975
Median :  0.000   Median :    0.00   Median : 6.000   Median :2008   COD    :  43   Alloca :  12   Median :163000
Mean   :  2.759   Mean   :   43.49   Mean   : 6.322   Mean   :2008   ConLD  :   9   Family :  20   Mean   :180921
3rd Qu.:  0.000   3rd Qu.:    0.00   3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   5   Normal :1198   3rd Qu.:214000
Max.   :738.000   Max.   :15500.00   Max.   :12.000   Max.   :2010   ConLw  :   5   Partial: 125   Max.   :755000
                                                                     (Other):   9
```



```
Electrical     1stFlrSF       2ndFlrSF      LowQualFinSF      GrLivArea     BsmtFullBath    BsmtHalfBath
FuseA:  84   Min.   : 483.0   Min.   :   0.0   Min.   :  0.000   Min.   : 605   Min.   :0.0000   Min.   :0.00000
FuseF:  22   1st Qu.: 887.2   1st Qu.:   0.0   1st Qu.:  0.000   1st Qu.:1140   1st Qu.:0.0000   1st Qu.:0.00000
FuseP:   2   Median :1086.0   Median :   0.0   Median :  0.000   Median :1465   Median :0.0000   Median :0.00000
Mix  :   0   Mean   :1154.8   Mean   : 342.8   Mean   :  5.695   Mean   :1503   Mean   :0.4249   Mean   :0.05937
SBrkr:1290   3rd Qu.:1372.0   3rd Qu.: 728.0   3rd Qu.:  0.000   3rd Qu.:1766   3rd Qu.:1.0000   3rd Qu.:0.00000
             Max.   :4692.0   Max.   :1818.0   Max.   :528.000   Max.   :5642   Max.   :3.0000   Max.   :2.00000

    FullBath       HalfBath       BedroomAbvGr    KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional    Fireplaces
Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Ex: 78       Min.   : 3.000   Maj1:  12   Min.   :0.0000
1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   Fa: 30       1st Qu.: 5.000   Maj2:   3   1st Qu.:0.0000
Median :2.000   Median :0.0000   Median :3.000   Median :1.000   Gd:577       Median : 6.000   Min1:  31   Median :1.0000
Mean   :1.562   Mean   :0.3805   Mean   :2.876   Mean   :1.048   TA:713       Mean   : 6.489   Min2:  34   Mean   :0.6109
3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000                3rd Qu.: 7.000   Mod :  11   3rd Qu.:1.0000
Max.   :3.000   Max.   :2.0000   Max.   :8.000   Max.   :3.000                Max.   :14.000   Sev :   1   Max.   :3.0000
                                                                                              Typ :1306
  GarageType   GarageYrBlt   GarageFinish   GarageCars      GarageArea     GarageQual GarageCond PavedDrive
2Types :  6   Min.   :1900   Fin:330      Min.   :0.000   Min.   :   0.0   Ex:   2    Ex:   2    N:  74
Attchd :914   1st Qu.:1962   RFn:414      1st Qu.:1.000   1st Qu.: 336.0   Fa:  43    Fa:  29    P:  30
Basment: 18   Median :1979   Unf:654      Median :2.000   Median : 478.5   Gd:  14    Gd:   9    Y:1294
BuiltIn: 79   Mean   :1978                Mean   :1.766   Mean   : 472.3   Po:   2    Po:   6
CarPort:  9   3rd Qu.:2001                3rd Qu.:2.000   3rd Qu.: 576.0   TA:1337    TA:1352
Detchd :372   Max.   :2010                Max.   :4.000   Max.   :1418.0

  WoodDeckSF      OpenPorchSF     EnclosedPorch     X3SsnPorch      ScreenPorch       PoolArea          MiscVal
Min.   :  0.00   Min.   :  0.0   Min.   :  0.00   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Min.   :    0.00
1st Qu.:  0.00   1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:    0.00
Median :  0.00   Median : 25.0   Median :  0.00   Median :  0.000   Median :  0.00   Median :  0.000   Median :    0.00
Mean   : 93.77   Mean   : 46.1   Mean   : 21.27   Mean   :  3.561   Mean   : 14.92   Mean   :  2.484   Mean   :   42.55
3rd Qu.:168.00   3rd Qu.: 67.5   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:    0.00
Max.   :857.00   Max.   :547.0   Max.   :552.00   Max.   :508.000   Max.   :480.00   Max.   :738.000   Max.   :15500.00

    MoSold          YrSold       SaleType    SaleCondition    SalePrice
Min.   : 1.000   Min.   :2006   WD     :1226   Abnorml:  93   Min.   : 79500
1st Qu.: 5.000   1st Qu.:2007   New    : 105   AdjLand:   4   1st Qu.:131000
Median : 6.000   Median :2008   COD    :  41   Alloca :  11   Median :163000
Mean   : 6.348   Mean   :2008   ConLD  :   9   Family :  20   Mean   :177141
3rd Qu.: 8.000   3rd Qu.:2009   ConLI  :   4   Normal :1162   3rd Qu.:210750
Max.   :12.000   Max.   :2010   ConLw  :   4   Partial: 108   Max.   :394617
                                (Other):   9
```

Data Cleaning on Test Data set

Follow the similar steps done on the training data set and will get the cleaned data set with 0 missing values. This test data set will be ready to predict the Sale Price.

RMSE

The **root-mean-square** deviation (RMSD) or **root-mean-square error** (**RMSE**) (or sometimes root-mean-squared error) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. I used rmse to predict the best model.

4. Building Regression Model

I decided to begin our post-cleaning analysis and model development process by generating a linear model. Linear models are the first models chosen to examine as they are the easiest to fit and most intuitively interpretable. The first model I decided to create was an ordinary least squares (OLS) model containing all the predictors in the data set. The results of the model can be seen in the regression output below.

Linear Regression Model

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05185 on 1201 degrees of freedom
Multiple R-squared:  0.9318,    Adjusted R-squared:  0.9189
F-statistic:    72 on 228 and 1201 DF,  p-value: < 2.2e-16
```

The rmse value for this model is 0.04751894

I made another linear model, but this time I kept the direction of the model both. This was the summary of the code.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05187 on 1296 degrees of freedom
Multiple R-squared:  0.9264,    Adjusted R-squared:  0.9188
F-statistic: 122.6 on 133 and 1296 DF,  p-value: < 2.2e-16
```

The rmse value for this model is 0.04937933

After running two linear models, I ran an SVM model (SUPPORT VECTOR MACHINE)

Support Vector Machine (SVM)

Support vectors are those points which are near to the hyperplane. The distance of the support vector from the hyperplane should be maximum but

minimum. SVM is not affected by outliers, transformation in SVM is known as kernel transformation

```
Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial |

       cost:  1
      gamma:  0.004273504
    epsilon:  0.1

Number of Support Vectors:  840
```

The rmse value for this model is 0.04863575

After a normal SVM model, I decided to run ten-fold cross-validation or a k fold cross-validation model Here it breaks the data into 10 sections it repeats it 10 times and takes a mean of accuracy The tune() function performs the cross-validation, the cost function is very critical in building the model, the cost function should be less.

```
Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost
  0.1

- best performance: 0.005117996

- Detailed performance results:
   cost       error  dispersion
1 1e-03 0.006250854 0.003568819
2 1e-02 0.005335288 0.005079708
3 1e-01 0.005117996 0.005271890
4 1e+00 0.005366755 0.005438945
5 5e+00 0.006125590 0.005661213
6 1e+01 0.006421596 0.005758065
7 1e+02 0.012424524 0.009933301
```

The rmse value for this model is 0.06503909

Bagging and Random Forest

Bagging, aka bootstrap aggregation, is a relatively simple way to increase the power of a predictive statistical model by taking multiple random samples(with replacement) from your training data set, and using each of these samples to construct a separate model and separate predictions for your test set. These predictions are then averaged to create a more accurate final prediction value.

```
Type of random forest: regression
                      Number of trees: 500
No. of variables tried at each split: 13

          Mean of squared residuals: 788081435
                % Var explained: 87.5
```

The rmse value for this model is 0.06338549

After Bagging model, I ran a random model.

The remove the biasedness from the bagging model I choose to run a random forest model. In the random forest, the variable keeps on changing Each tree in the random forest is grown to the largest extent possible, and there is no pruning. Random-forest can handle large data easily, and it can even identify the most significant variables, so it is considered as one of the dimensionality reduction methods.

```
Call:
 randomForest(formula = SalePrice ~ ., data = train, trcontrol = fitcontrol)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 25

          Mean of squared residuals: 0.003977538
                    % Var explained: 87.99
```

The rmse value for this model is 0.0255915

KNN

K nearest neighbours is a simple algorithm that stores all the available cases and classifies new cases by a majority vote of its k neighbours. This algorithm segregates unlabeled data points into well-defined groups. Choosing the number of nearest neighbours plays a significant role in the efficiency of the model.

```
k-Nearest Neighbors

1430 samples
  75 predictor

No pre-processing
Resampling: Bootstrapped (5 reps)
Summary of sample sizes: 12, 15, 14, 20, 15
Resampling results across tuning parameters:

  kmax  RMSE       Rsquared   MAE
  5     0.1370247  0.4843059  0.10022122
  7     0.1366240  0.5004480  0.09953246
  9     0.1366240  0.5004480  0.09953246

Tuning parameter 'distance' was held constant at a value of 2
Tuning parameter 'kernel' was held constant at a value of optimal
RMSE was used to select the optimal model using  the smallest value.
The final values used for the model were kmax = 9, distance = 2 and kernel = optimal.
> |
```

The rmse value for this model is 0.05964994

5. Conclusion

After cleaning the data and handling missing values, I build the model where I predicted the sale price values on the test data. Each model prediction was different.

Model name	Rmse
LINEAR MODEL	0.04751894
LINEAR MODEL	0.04937933
SVM	0.04863575
SVM	0.06503909
BAGGING	0.06338549
RANDOM FOREST	0.0255915
KNN	0.05964994

It is seen that the random forest was the most robust model. The rmse value of random forest model is 0.0255915. I used that model to predict the SalePrice on test data.

https://github.com/MalihaG/Ames-Housing-Data

Machine Learning | Data Science | Data Visualization | Predictive Analytics | Data

Written by Maliha Gangat

11 Followers · Writer for Analytics Vidhya

"Data are like puzzles, and I enjoy solving them."

More from Maliha Gangat and Analytics Vidhya

Maliha Gangat in Analytics Vidhya

Effect of Covid-19 on Mental Health

Our Mind and Covid-19

Dec 6, 2020 45

Kia Eisinga in Analytics Vidhya

How to create a Python library

Ever wanted to create a Python library, albeit for your team at work or for some open...

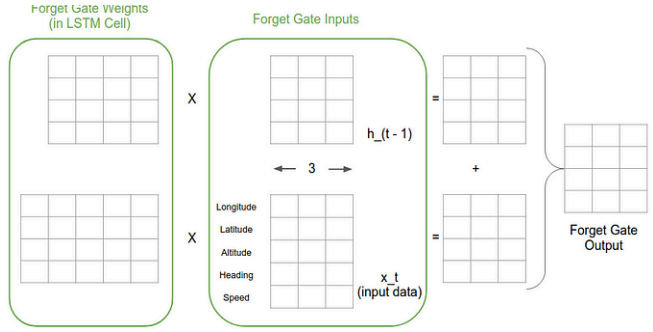
Jan 26, 2020 2.7K 28

Harikrishnan N B in Analytics Vidhya

Confusion Matrix, Accuracy, Precision, Recall, F1 Score

Binary Classification Metric

Dec 10, 2019 1.1K 6



Ryan T. J. J. in Analytics Vidhya

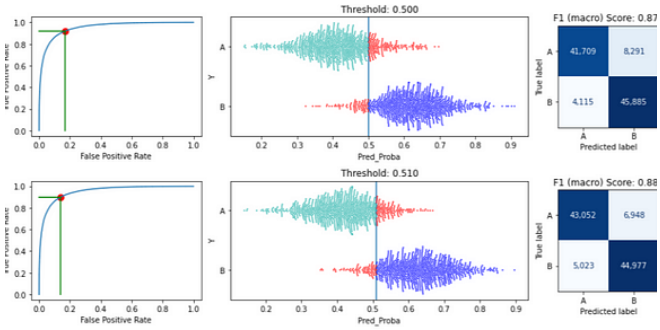
LSTMs Explained: A Complete, Technically Accurate, Conceptual...

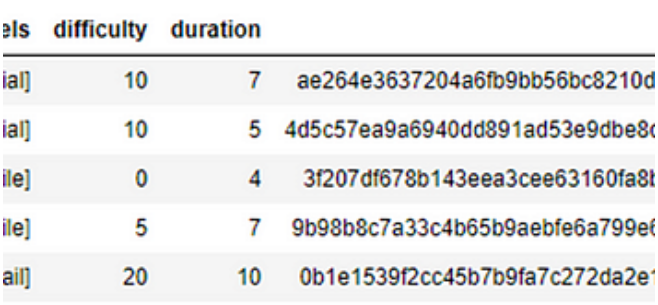
I know, I know—yet another guide on LSTMs / RNNs / Keras / whatever. There are SO many...

Sep 2, 2020 421 7

See all from Maliha Gangat | See all from Analytics Vidhya

Recommended from Medium





els	difficulty	duration	
ial]	10	7	ae264e3637204a6fb9bb56bc8210d
ial]	10	5	4d5c57ea9a6940dd891ad53e9dbe8
ile]	0	4	3f207df678b143eea3cee63160fa8
ile]	5	7	9b98b8c7a33c4b65b9aebfe6a799e
ail]	20	10	0b1e1539f2cc45b7b9fa7c272da2e

W Brett Kennedy in Towards Data Science

Achieve Better Classification Results with...

A python tool to tune and visualize the threshold choices for binary and multi-class...

5d ago 319 3

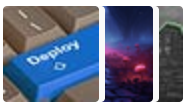
Giga85

Udacity Data Scientist Capstone Project: Starbucks dataset

Udacity Data Scientist Capstone Project: Starbucks dataset

Jun 3 2

Lists



Predictive Modeling w/ Python

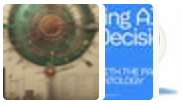
20 stories · 1509 saves

Practical Guides to Machine Learning

10 stories · 1835 saves

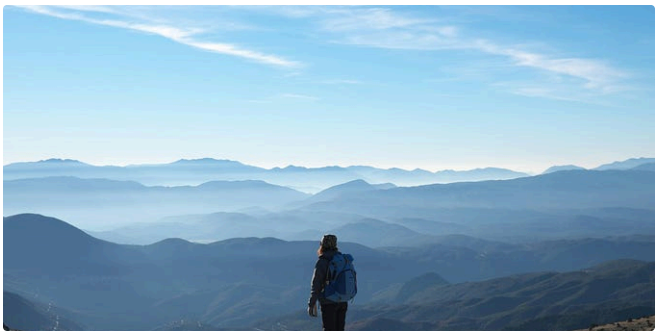
Natural Language Processing

1687 stories · 1260 saves



data science and AI

40 stories · 237 saves

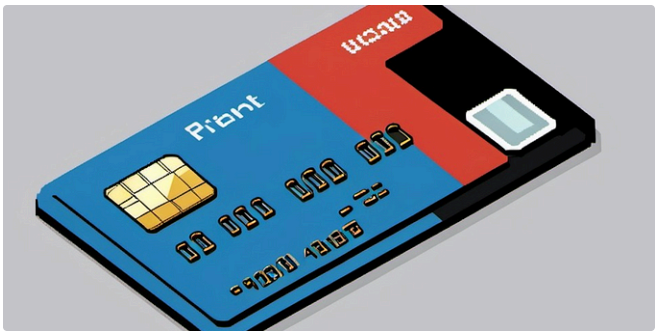


Sze Zhong LIM in Data And Beyond

Mastering Exploratory Data Analysis (EDA): Everything You...

A systematic approach to EDA your data and prep it for machine learning.

Apr 6 835 7



Saankhya Mondal in Towards AI

Data Science Case Study—Credit Default Prediction: Part 1

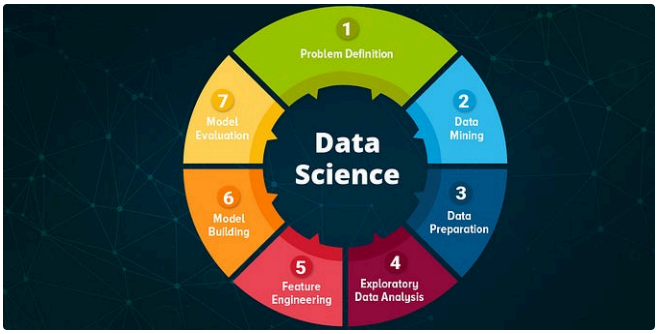
Feature Engineering, Model Training and Evaluation, and Classification Threshold...

May 2 10

Patwariraghottam in DevOps.dev

Mastering Feature Selection: Techniques to Boost Your Machin...

Aug 25 24 1



Neha Gupta

Beginner's Guide to EDA : Exploratory Data Analysis

Hey 😄 there data enthusiasts, curious minds, and fellow explorers of the data...

Apr 9 13

See more recommendations