



Big data analytics predicting real estate prices

Archana Singh¹ · Apoorva Sharma¹ · Gaurav Dubey²

Received: 25 July 2019 / Revised: 14 December 2019

© The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2020

Abstract The enormous data generated on daily basis amounts to big data technologies. This large amounts of data have knowledge and hidden patterns. Real estate turning out to be another biggest application in big data. The emphasis of this paper is to map the process involved in taking large amounts of data to predict the price of a house in real estate. The real estate sounds to be a long-term investment. In this paper, the housing Sale Data from Ames, Iowa is considered for the timeframe 2006–2010 with a view to construct relevant models to estimate the final sale price of a house. Due to high number of explanatory variables several models such as linear regression, random forest and gradient boosting models have been used as tools for feature selection to determine the statistically significant characteristics that influence the final sale price of a house. It has been observed that out of all the models, the gradient boosting model returned the efficient results.

Keywords Big data · Real estate · Random forest model · Gradient boosting model · Linear regression · LASSO

1 Introduction

Big data is a complex term that showcases a voluminous measure of organized, semi-structured and unstructured information that can be mined for data. The tremendous data has extended the need for information affiliation bosses to such a degree, to the point that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than \$15 billion on programming firms growing. In 2010, this industry was worth more than \$100 billion and was making at basically 10% a year that is about twice as more than the previous years. There are 4.6 billion remote selections around the globe, and between 1 billion and 2 billion people getting to the web. “Big data is making the commercial real estate industry more transparent,” commented by reputed CEO. It has also become the hub of the real estate market with brokers, lenders, investors and owners-all a part of it.

Big Data has literally changed the entire face of Information Technology. It is quite prevalent in most of the fields we come across daily, be it the management of Road Traffic where an Intelligent Transportation System from several devices, like road sensors, camera on traffic light, or vehicle sensors can provide more efficient data to the users or Big Visual Data Analysis in Security Space (Battle 2013; Brown et al. 2011). Big Data has also started doing wonders in the field of medicine, where data is now available in the form of Big Data in a box (Rahman et al. 2016; Dachuan and Baoshan 2012). There has also been a vast increase of the use of Big data in the science sectors of the government to be able to collect and manage huge amounts of data (Tene and Polonetsky 2012). As we have learnt already how to use Web-based collaborative big data analytics on big data as a service platform, we need to

✉ Archana Singh
asingh27@amity.edu

Apoorva Sharma
apoorvasharma.888@gmail.com

Gaurav Dubey
gdubey1977@gmail.com

¹ Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

² ABES Engineering College, Ghaziabad, India

introduce it to sectors where Big data is unheard of. Presently, using big data framework (Singh and Ranjan 2016), business is enabled to identify hidden attributes for immediate decisions. (Sonka and IFAMR 2014). Working faster—and staying agile—gives organizations a competitive edge they didn't know they had before.

Such high level of analytics gives technical systems even though the data is way too huge. For example, we can get judicious insights to take better decisions about dynamic opportunities, get explicit solutions to technical problems and discover new growth opportunities—all at the same time. Not only with the essential information, but Big data is also helping real estate with newer and interesting techniques in planning a strategy for society building. In order to make a residential society, a proper planning has to be done. In this part, big data is helping a lot, by doing research and letting property developers know about the quality of the environment, a balanced atmosphere, and much more which is a must for the construction of a residential society (Huang and Yin 2014; Juan 2013). So the big data analytics provide the real estate builders with information regarding health care and energy competence needed for a proper construction development. Data even plays a big role in the engineering concepts (Singh 2017) with the help of which, the civil engineers get better information to engineer those buildings with stronger basements. Presently, the companies using big data technologies are able to build the perfect performance insights of businesses. It provides to its users, people are too aware of which properties they should buy and which they should not. Even big data has given the banking sectors full rights to view the real estate insights, as a result of which, before investing any money, the investing sectors make sure that every rupee is worthy for that particular property (Wang and Kang 2014). Without the involvement of big data technologies, real estate businesses used unprocessed data which never really gave them completely accurate results. The application of big data has turned real estate into a completely clear business and along with such transparency, business analytics also gave them a new turn (Singh et al. 2015). The processing of the unprocessed data and other relations are advanced by using the big data analytics. Managing of the information extracted from the outputs of censuses, the consequences of consumer buying capacity, catalogues of homes for sale and lease, data of the geographic information systems etc. are all done by data analytics today (Qiuming and Yunfeng 2011). The big data analytic tools study about the patterns that customers use while buying properties, and these data helps in improving the real estate sales and also helps them appreciate home-value fashions within a scrupulous environment. From the information generated by big data about the real estates, people come to know easily about the different buildings

and houses available for rent as well as for sale within a location. Even if you hire a real estate agent, you can always confirm with the prices by comparing it with similar properties available on the internet. A quick survey in the World Wide Web will allow you to come to a better conclusion protecting you from the huge loss you might have gone through. A real-estate agent holds the complete knowledge of all the buyer incentives, exclusive offers going on in the market and all other insights. Hiring a good real estate agent is really helpful but do not forget to make use of a little of the internet to have a cross check of what the agent is taking you through.

In this paper, several modelling techniques have been used, such as the LASSO (Least Absolute Shrinkage and Selection Operator), a regression model that regularizes the selected predictor variable (Regularization (mathematics) 2019). A hybrid variable selection/classification approach is considered that is based on linear combinations of the gene expression profiles that maximize an accuracy measure summarized using the receiver operating characteristic curve. Under a specific probability model, this leads to the consideration of linear discriminant functions. An automated variable selection approach using LASSO to simulated data as well as data from a recently published prostate cancer study has been incorporated (Ghosh and Chinnaiyan 2005). Random forest has also been used in this paper which uses several random trees to give the result of one. The classification and regression using random forest in R has been explained in detail in a recent study. Another technique, gradient boosting works on the principle that the next best possible model, when added to the former models, reduces the overall prediction error.

1.1 LASSO: variable selection

LASSO (Least Absolute Shrinkage and Selection Operator) is generally used to deal with the absolute size of the coefficients of regression. It reduces the variability and improves the more correctness of linear regression models. LASSO uses penalty function to bring out more accuracy in variable selection, as it causes some parameter estimates to converge to zero. The assumptions in this regression are similar to the least squared regression other than the fact that normality is not to be simulated. It now reduces the coefficients to zero (exactly zero) that, then contributes to the feature selection. This uses a regularization method which is a process of adding data to be able to solve a vaguely illustrated problem it is also used to avoid over-fitting. Regularization also enforces to objective functions in ill-posed optimization problems. If a set of predictors are highly correlated, LASSO chooses only one of them and reduces the rest to zero.

1.2 Predictive analysis

The correctness of a predictive model can be boosted in two ways: Either by following feature engineering or by implementing boosting algorithms right away.

A. Gradient boosting

The gradient boosting uses the greedy approach in constructing the trees for prediction. Weak learner to predict the values. In calculating gain uses, greedy approach, maximum number of nodes, a smaller number of layers and depth (Pantazopoulos and Maragoudakis 2018). The gradient descent helps in minimizing the loss while adding of trees (Natekin and Knoll 2011).

Thus, we use gradient boosting machines, an ensemble algorithm that can learn with different loss functions providing the ability to work efficiently with high dimensional data as shown in Fig. 1.

B. Random forest

It is supervised algorithm; random forest can be applied in both the cases of classification and regression. It creates an ensemble of decision trees, using the bagging method. Bagging method outcomes, the best combination of machine learning models and enhances the overall results as shown in Fig. 2.

Some key facts to consider before running any random forest model are as follows:

- With a high number of predictors present, the eligible predictor set will be quite different from node to node.
- The greater the inter-tree correlation, the greater the random forest error rate, so one pressure on the model is to have the trees as uncorrelated as possible.

C. Linear regression

Linear regression is used to predict the value of an outcome variable Y based on one or more input predictor variables X . The aim is to establish a linear relationship (a mathematical formula) between the predictor variable(s) and the response variable, so that, we can use this

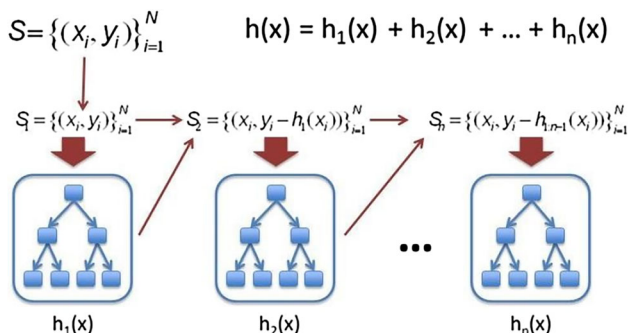


Fig. 1 Gradient boosting. Source: <http://statweb.statford.edu/>

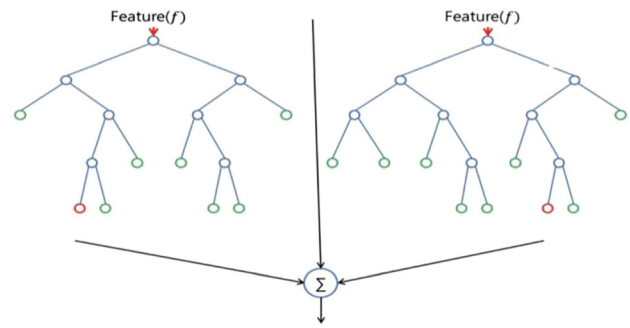


Fig. 2 Random forest. (Source: internet)

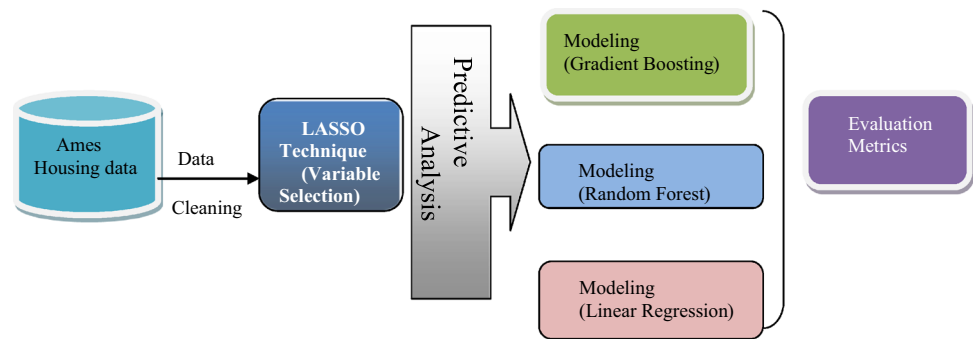
formula to estimate the value of the response Y , when only the predictors (X s) values are known. The aim of linear regression is to model a continuous variable Y as a mathematical function of one or more X variable(s), so that we can use this regression model to predict the Y when only the X is known (Wheeler and Tiefelsdorf 2005).

2 Proposed framework

The proposed framework uses the feature selection to determine the statistically significant characteristics that influence the final sale price of a house. In this paper, LASSO is the pre-processing technique used for the cleaning of data. There may exist missing values that need to be replaced or ignored on a case by case basis which cannot be ignored. Looking at the complexity of datasets, this paper attempted to propose a guideline on how to go about data processing in the most efficient manner. The dataset considered is Ames Housing data with 2930 observations over 81 explanatory variables. The timeframe for these sales is 2006–2010 and we may have to factor in the economic downturn and stagnation of housing prices in the United States during the 2008 recession. Sale Price significant component to be predicted using data mining technique Predictive modelling in presence of a large number of exploratory variables requires use of methods that support feature selection. The fundamental approach is to break down the train dataset into separate data frames based on variable type viz. nominal, ordinal, discrete and continuous. Also, to predict the top 5 variables that determine the price of the house using LASSO, random forest, linear regression and gradient boosting models as shown in Fig. 3.

3 Research methodology

To employ these learning algorithms such as random forest and generalized boosting models as tools for feature selection to identify the most significant variables in model

Fig. 3 Proposed model

fitting of real estate sale price. The hypothesis for this paper is based on the findings of a survey which has categorized the variables that influence the decision of home buyers as follows. The predictive analysis of price was made on the basis on the significant features which were considered in this paper are Property Layout, Property Size, Location, Parking, and Garden. The less significant features considered were (Fireplace, Interior Design, Kitchen and Bath Finish). The price prediction does not include the feature No garden, No heating system, No driveway. A potential home owner may consult this list of “most significant” variables in conjunction with their own preferences to make a decision to buy or not buy based on evidence. The pre-processing of data is an unspoken aspect of big data analytics that has no set methodology. It is data specific and is driven by the business objectives of the firm. The relationships may exist between the variables that require human interpretation. Further, missing values that need to be replaced or ignored on a case-by-case basis, which cannot be generalized. Given a diversity and cumbersome nature of the data cleaning process, this paper attempts to propose a guideline on how to go about data processing in the most efficient manner.

The dataset considered is Ames Housing data with 2930 observations over 81 explanatory variables. The timeframe for these sales is 2006–2010 and we may have to factor in the economic downturn and stagnation of housing prices in the United States during the 2008 recession. The dataset is partitioned into a train set and a test set with a view to build the model based on the train set and test its efficacy on the test dataset. The train dataset has 1460 observations with 81 variables while the test dataset has 1459 observations over 80 variables minus Sale Price which is to be predicted. The predictive modelling is used for a large number of exploratory variables requires use of methods that support feature selection. The fundamental approach is to break down the train dataset into separate data frames based on variable type viz. nominal, ordinal, discrete and continuous. The methodology to deal with the missing values for each variable type differs and this paper intends to discuss the steps involved in cleaning data including checks for

linearity, the problem of multi collinearity and finally imputing the missing data to further build models. The eighty variables target the quality and quantity of various physical characteristics of the houses. Mostly, the variables given are typically the kind of information that a house-buyer shall like to enquire about a potential house (For instance, When and Where was it constructed? How much parking space does it provide? How much area of living space is in the house? Is the house well-furnished? How many washrooms are present?). Usually, the twenty continuous variables refer to many dimensions of area provided for each observation. Additional to the general lot size and total property square footage on most usual house listings, various other distinct variables are quantified in the dataset. The area measurements of the basement area, central living space, and even porches are categorized into distinct categories depending upon their quality and type. The 14 discrete variables usually quantify the total number of variables present in the house. Most of them usually focus on the number of kitchens, bedrooms, and bathrooms. Furthermore, the garage space and re-construction/renovation dates are also catalogued.

There is a huge number of variables (23 nominal, 23 ordinal) present in this data set. It starts from 2 to 28 classes with the lowest, as STREET (gravel or paved) and the biggest, as NEIGHBOURHOOD (areas within the Ames city limits). The nominal variables usually come under the different types of houses, garages, materials, and other external factors while the ordinal variables usually rate the variables within the house.

3.1 Importing packages and datasets on R

The most vital part of a data science project is the right usage of packages. The packages are identified and thus, installed onto RStudio one by one. The packages used in this data science work are: ‘gdata’ (which is a data manipulation tool), ‘tidyverse’ (which lists all the conflicts between packages, ‘stringr’, which is used for string values in R, ‘lubridate’, used for the parsing process in R, ‘scales’, which is used for visualization, graphics, which, as the

name suggests is used for graphical representation and lastly, 'caret', which is used for Classification and Regression Training as shown in Fig. 4.

3.2 Cleaning the data

The ordinal variables shall be given some numeric values so that comparison between ordinal variable becomes easier. For instance, a variable like 'LotShape' that has four levels such as Regular, slightly irregular, moderately irregular, irregular etc. shall be recoded as 4, 3, 2, 1 respectively. To be able to clean the nominal variables, a graphical analysis is vital to make the comparison intelligible. For instance, 'Street'—the type of road access to property, which also has a significant impact on the final sale price of the house.

The two types of streets are: Gravel or Paved, as shown in Fig. 5.

Another important Nominal Variable that widely affects the Sale Price of the house is the type of Neighbourhood. The neighbourhood in the area of Ames, Iowa are given in Table 1.

(A) Checking the Continuous and Discrete Variables

These variables are numeric and do not require any form of recoding. During further analysis while model fitting using random forest, the numeric variables will have to be dealt as factors. The main challenge is to deal with the missing values in these variables for which a R package in called MICE (Multivariate Imputation Using Chained Equations) is used to impute the missing data. R offers a "mice" package full of classes to create imputations for multivariate data. This algorithm can take variables of continuous, binary, unordered categorical and ordered. In Additionally, Multivariate Imputation Using Chained

Equations can include continuous two-level data, and maintain consistency between entries by means of passive imputation. Several diagnostic rules are carried out to check the quality of the imputations. However, the method is used to impute missing values for variables that had less than 5% missing values, all the others were dropped from further analysis as shown in Fig. 6.

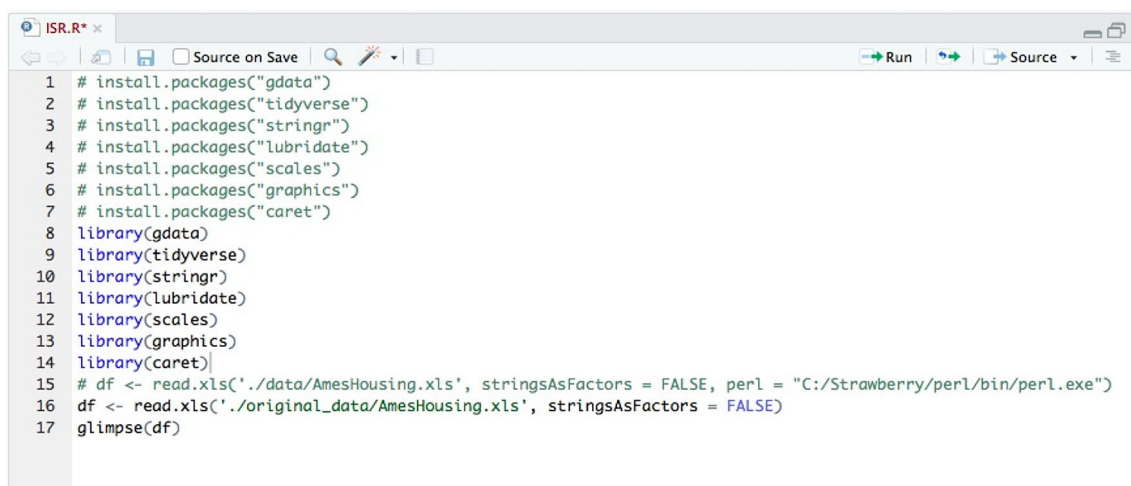
The variables with the highest correlation were extracted and a Correlation Plot was shown in Fig. 7.

3.3 Checking for multi collinearity

In order to identify the variable exhibiting high multi collinearity a Heat Map was used to plot correlation between all 81 explanatory variables as shown in Fig. 7. To remove the problem of multi collinearity in the data the heuristic approach is used that looks at the two way predictor correlation as shown in Fig. 8. Therefore, the algorithm to be followed is as followed (Wheeler and Tiefelsdorf 2005).

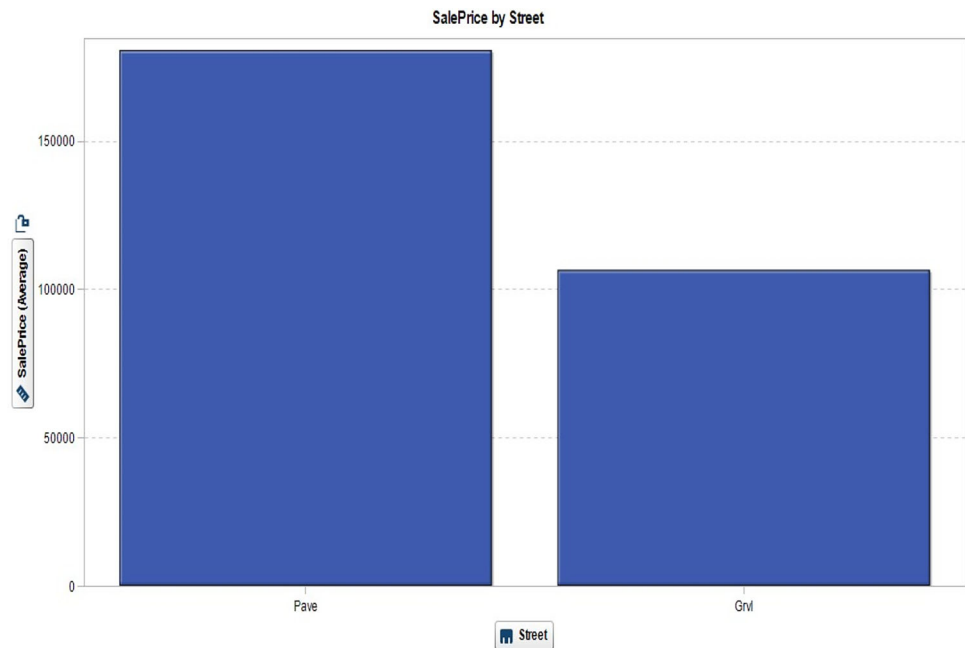
```
Step 1: Compute the correlation matrix of all independent variables (predictors)
Step 2: Identify two significant predictors for largest absolute pairwise correlation
(assuming predictors X and Y)
Step 3: Compute the average correlation X and the other variables.
Step 4: Repeat the same for the variable Y.
Step 5: Compare X and Y correlation value, remove the variable with larger value.
Step 6: Repeat Steps 2–5 until correlation value > threshold value
```

Based on the results of the above algorithm, even after attempting to resolve the issue of multi collinearity in the data, the following variables are to be dropped from further analysis: Garage Quality, Pool Quality, Total Basement Area, and Basement Finished Area.



```
1 # install.packages("gdata")
2 # install.packages("tidyverse")
3 # install.packages("stringr")
4 # install.packages("lubridate")
5 # install.packages("scales")
6 # install.packages("graphics")
7 # install.packages("caret")
8 library(gdata)
9 library(tidyverse)
10 library(stringr)
11 library(lubridate)
12 library(scales)
13 library(graphics)
14 library(caret)
15 # df <- read.xls('./data/AmesHousing.xls', stringsAsFactors = FALSE, perl = "C:/Strawberry/perl/bin/perl.exe")
16 df <- read.xls('./original_data/AmesHousing.xls', stringsAsFactors = FALSE)
17 glimpse(df)
```

Fig. 4 Packages used in the analysis

Fig. 5 Types of streets**Table 1** Type of neighbourhood (variable) areas of Ames, Iowa

Blmngtn	Bloomington heights	Mitchel	Mitchell
Blueste	Bluestem	Names	North Ames
BrDale	Briardale	NoRidge	Northridge
BrkSide	Brookside	NPkVill	Northpark Villa
ClearCr	Clear Creek	NridgHt	Northridge Heights
CollgCr	College Creek	NWAmes	Northwest Ames
Crawfor	Crawford	OldTown	Old Town
Edwards	Edwards	SWISU	South and West of Iowa State University
Gilbert	Gilbert	Sawyer	Sawyer
Greens	Greens	SawyerW	Sawyer West
GrnHill	Green Hills	Somerst	Somerset
IDOTRR	Iowa DOT and Rail Road	StoneBr	Stone Brook
Landmrk	Landmark	Timber	Timberland
MeadowV	Meadow Village	Veenker	Veenker

4 Experiments and results: model fitting

The aim of this paper is to identify the statistically significant predictor variables to build an accurate model to estimate the final Sale Price. Based on the Exploratory Data Analysis performed thus far, the variables that may be considered significant are listed as follows:

Ordinal Variables Overall Quality, Exterior Quality, Basement Quality, Heating Quality, Electrical System.

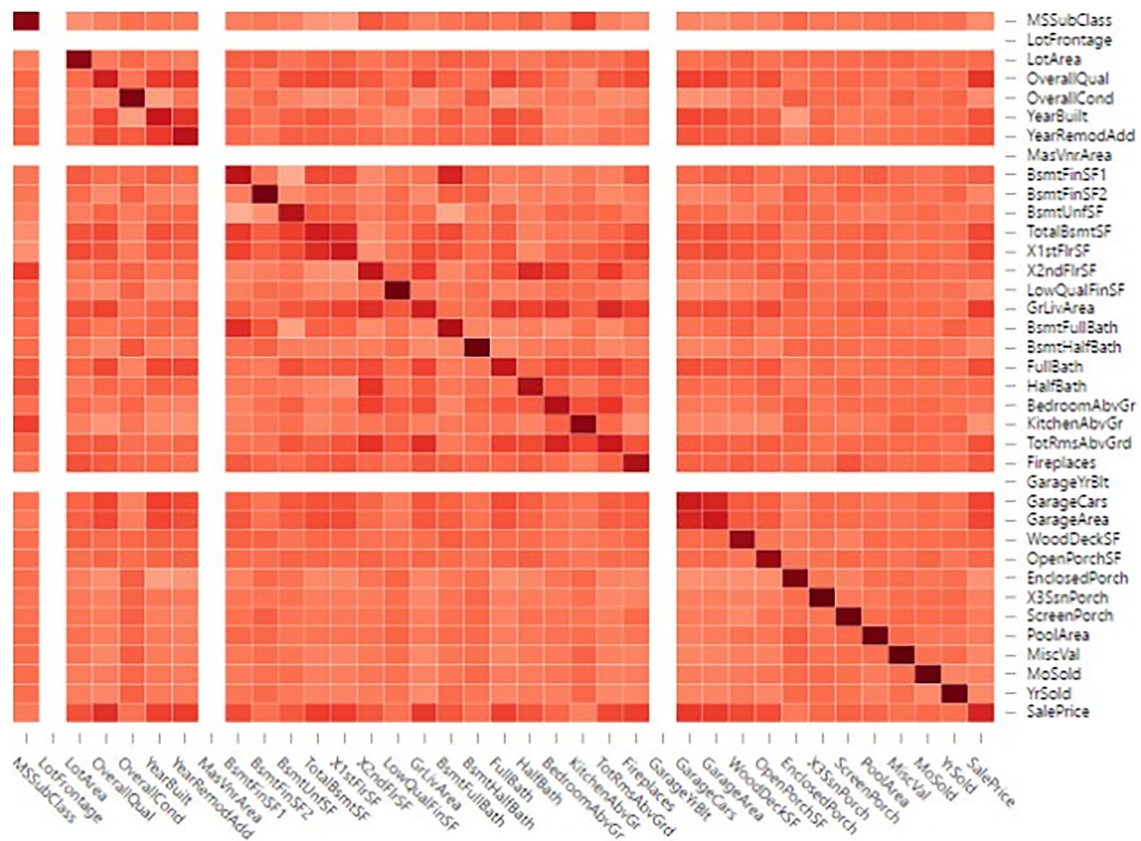
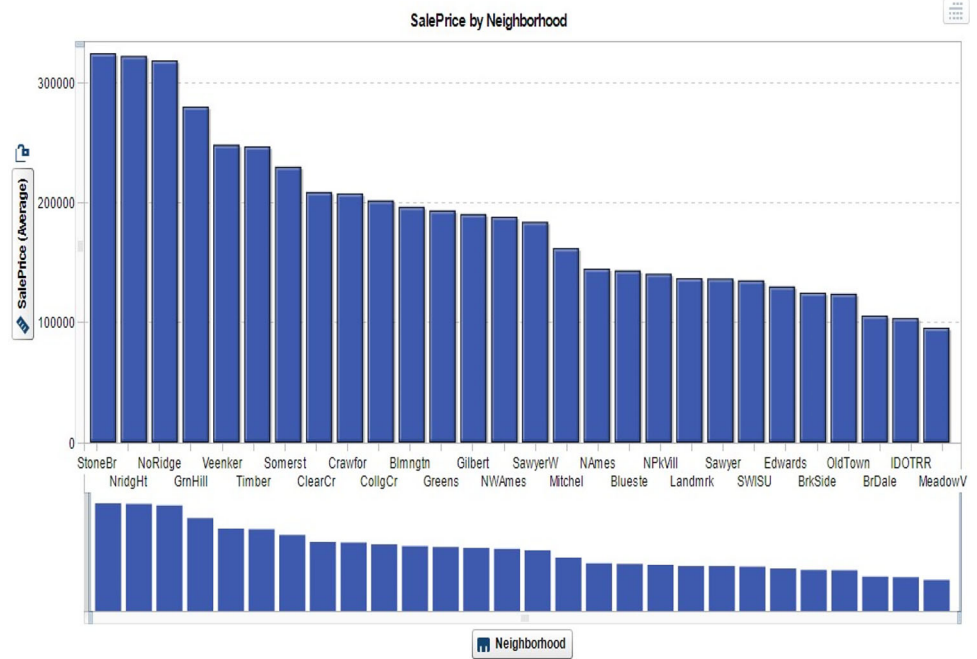
Nominal Variables Sub Class of Building, Zoning Class of Building, Driveway Access to Property, Neighbourhood and Condition 1 and 2: Describes the nearby facilities. In case of the discrete and continuous variables, it would be inappropriate to make a claim

based on the nature of EDA performed in this paper so we shall proceed to discuss the various models that may be used to predict Sale Price. After data pre-processing and cleaning, the cleaned and relevant data is analysed using linear regression, random forest and gradient boosting respectively.

4.1 Model fitting

(A) Linear regression model

A kitchen sink model was built using all the significant variables produced after the pre-processing of data. The initial output for this model is as follows:

Fig. 6 Analysis of saleprice by neighbourhood**Fig. 7** HeatMap to detect correlations between 81 variables

Note that the adjusted R Squared value for the kitchen sink model is: 92.06% indicative of a good data cleaning process as shown in Figs. 9, 10.

Note that the scatter appears to be random thus not violating any assumptions behind linear regression modelling. A backward stepwise elimination method was used

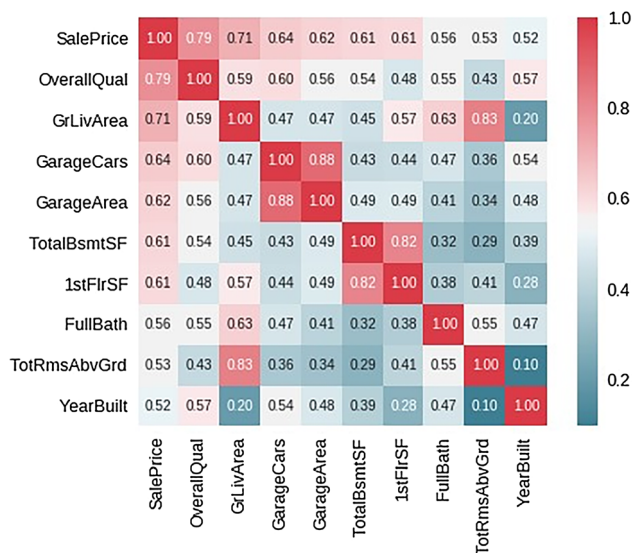


Fig. 8 Correlation plot

with each least significant variable basis p value removed and the model refitted (Prabhakaran 2019). The RMSE remained quite higher than the rest and the most significant predictor variables are found to be as follows, shown in Fig. 11.

In extension to the non-linear behaviour amidst the independent and dependent variables, the regression output above stated that complete assumption of rank is not conferred with—as shown by the ‘NA’ values for the regression coefficients. Four of the above mentioned assumptions are blatantly violated, although, it looks to be the case that ergogeneity of the independent variables still works as the mean of the residuals is mostly equal to 0. Therefore, as several of the assumptions for linear regression models are

Fig. 9 Adjusted R Square value 92.06%

```

KitchenAbvGr3    -2.787e+03    4.132e+04    -0.067    0.946496
KitchenAbvGr4    -1.804e+04    4.491e+04    -0.402    0.687968
KitchenQual      4.970e+03    1.573e+03     3.160    0.001617 **
TotRmsAbvGrd2    -2.927e+04    2.542e+04    -1.151    0.249805
TotRmsAbvGrd3    -3.074e+04    2.504e+04    -1.228    0.219873
TotRmsAbvGrd4    -3.221e+04    2.510e+04    -1.283    0.199631
TotRmsAbvGrd5    -3.022e+04    2.524e+04    -1.197    0.231504
TotRmsAbvGrd6    -3.232e+04    2.537e+04    -1.274    0.202959
TotRmsAbvGrd7    -2.757e+04    2.550e+04    -1.081    0.279826
TotRmsAbvGrd8    -2.859e+04    2.571e+04    -1.112    0.266368
TotRmsAbvGrd9    -1.120e+04    2.610e+04    -0.429    0.667849
TotRmsAbvGrd10   -5.139e+03    2.674e+04    -0.192    0.847646
TotRmsAbvGrd11   2.403e+04    2.805e+04     0.857    0.391767
TotRmsAbvGrd12   NA                NA                NA                NA
Fireplaces2      -2.417e+02    4.568e+03    -0.053    0.957810
Fireplaces3      3.723e+03    5.171e+03     0.720    0.471730
Fireplaces4      1.284e+04    1.239e+04     1.037    0.299961
FireplaceQu      2.984e+02    1.291e+03     0.231    0.817232
GarageType2      2.192e+04    1.074e+04     2.041    0.041484 *
GarageType3      2.347e+04    1.246e+04     1.884    0.059793 .
GarageType4      2.363e+04    1.125e+04     2.100    0.035937 *
GarageType5      2.566e+04    1.383e+04     1.855    0.063832 .
GarageType6      2.561e+04    1.071e+04     2.391    0.016976 *
GarageType7      1.893e+04    1.837e+04     1.030    0.303061
GarageYrBlt      3.867e+00    5.645e+01     0.068    0.945407
[ reached getoption("max.print") -- omitted 54 rows ]
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21610 on 1209 degrees of freedom
Multiple R-squared:  0.934,    Adjusted R-squared:  0.9206
F-statistic: 69.6 on 246 and 1209 DF,  p-value: < 2.2e-16

```

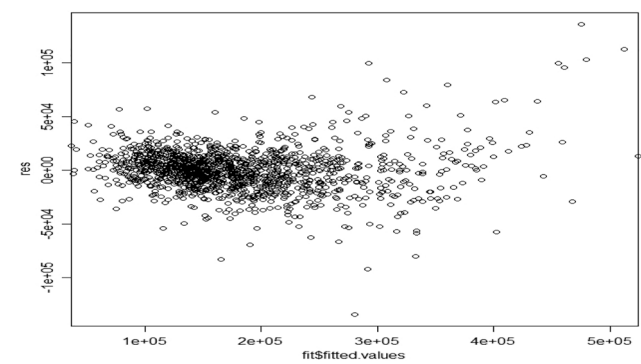


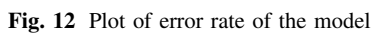
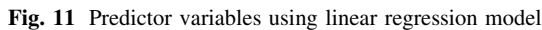
Fig. 10 Plot of residuals

not followed, one must not use these results for further analysis.

(B) Random forest

As soon as a new input is entered into the system, it is run down all of the trees. The result is either an average or weighted average of all the terminal nodes that are reached, or, in the case of categorical variables, a voting majority. In order to determine the most appropriate number of trees, a plot of error rate of the model using all 79 variables to predict Sale Price is as follows in Fig. 12.

It can be observed from the figure above that the error rate becomes flat after about 150 trees. This indicates that increasing the number of trees beyond the default value of 500, is unlikely to have a significant impact on the model accuracy. Therefore, for this study the default value for number of trees is kept constant at 500. The relative importance plot for Sale Price after fitting random forest is as in Fig. 13. Feature Selection based on the above plot shown in Fig. 13. It may be concluded that the most



(C) Gradient boosting

The function above runs Boost on R and this method was rerun using stepwise backward elimination by removing the insignificant variables. The least R MSE recorded was: \$ 8009.23 while the model containing least number of predictor variables recorded an RMSE of \$ 8022.17. The relative importance plot based on gradient boosting is shown in Fig. 15.

Lot Area, Above Ground Area, Overall Area, Total
Basement Area, Year Built, Year Remodelled

In order to check the accuracy of the built models, RMSE (Random Mean Squared Error) was computed for each model and compared to the values on the Kaggle website (source of dataset). Random forest algorithms are relatively fast, and used for majorly unbalanced and missing data. This algorithm is limited to predict to a particular range of training dataset, may produce noisy and over fit models when used in regression. However, in this case gradient boosting models returned the best results as compared to the linear regression and random forest.

RMSE, interprets the standard deviation of the unexplained variance. Since, the lower value of RMSE indicates better fit of the model. So, from the above Table 1, gradient boosting looks to be the best fit.

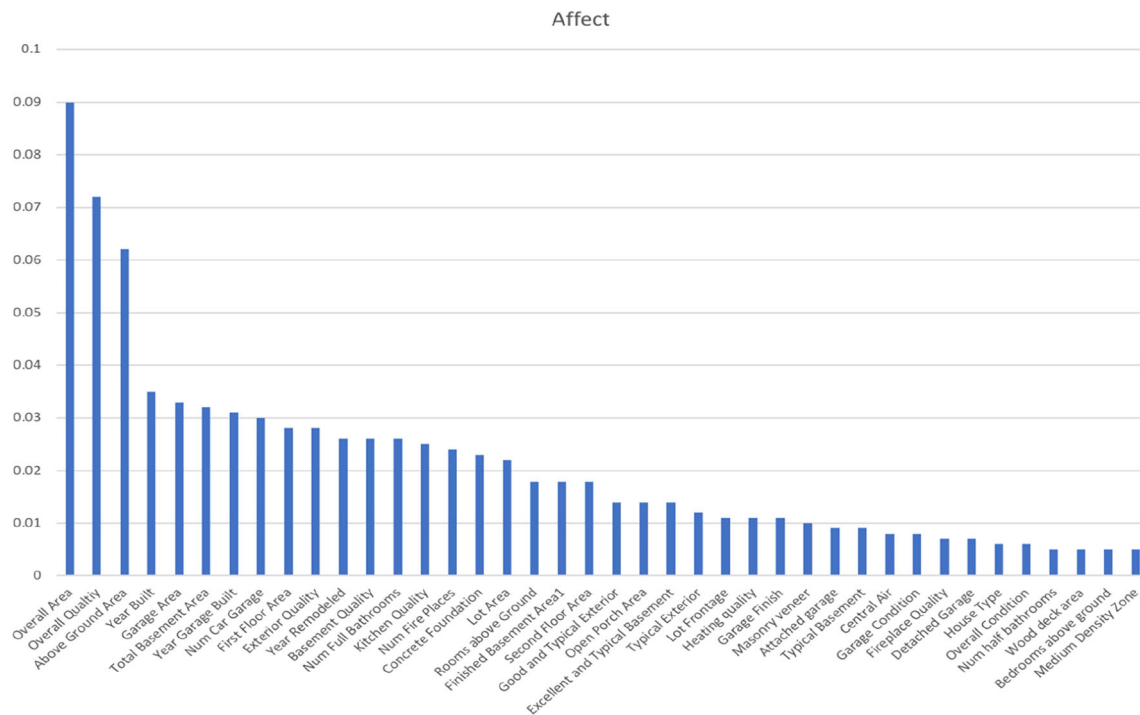


Fig. 13 Feature selection using random forest

```
runBoost <- function(trees, depth, learn, trainset, testset)
{
  set.seed(1)
  boost.price <- gbm(SalePrice~.,
    data = trainset,
    distribution = "gaussian",
    n.trees = trees,
    interaction.depth = depth,
    shrinkage = learn)
  yhat.boost <- predict(boost.price, newdata = testset,
    n.trees = trees)

  rmse.boost <- sqrt(mean((new.Train$SalePrice -predict(boost.price, newdata = trainset,
    n.trees = trees))^2))

  print(summary(boost.price))
  print(paste("Training RMSE : ", rmse.boost))
}

## Running the function with different values:
runBoost(trees = 20000, depth = 10, learn = 0.001, trainset = new.Train , testset = new.Test)
runBoost(trees = 200, depth = 10, learn = 0.1, trainset = new.Train , testset = new.Test)

## Based on the results we shall drop the following columns:
drop <- c("PoolQC", "PoolArea",
  "Street", "Condition2", "MiscVal", "MiscFeature", "BldgType", "Heating", "Alley", "RoofMatl", "GarageCond", "PavedDrive",
  "LandSlope")
train.gbm <- new.Train[,!names(new.Train) %in% drop]
test.gbm <- new.Test[,!names(new.Test) %in% drop]

## Running the code after dropping the above columns as follows:
runBoost(trees = 20000, depth = 10, learn = 0.001, trainset = train.gbm , testset = test.gbm)
runBoost(trees = 200, depth = 10, learn = 0.1, trainset = train.gbm , testset = test.gbm)
write.csv(data.frame(Id = c(1461:2919), SalePrice = yhat.boost), row.names = FALSE)
```

Fig. 14 Function used to implement gradient boosting

5 Conclusion and future scope

Big data analytics is empowering real estate with technical models used in planning and marketing for society building. In order to make a residential society an elite project, the big data analytics provide the real estate builders with

information regarding environment and energy competence needed for a proper construction development. The Big Data even has a big role in the engineering concepts, with the help of which, the civil engineers get better information to construct buildings with stronger basements. This leads to lot of investment of money, manpower and time. In

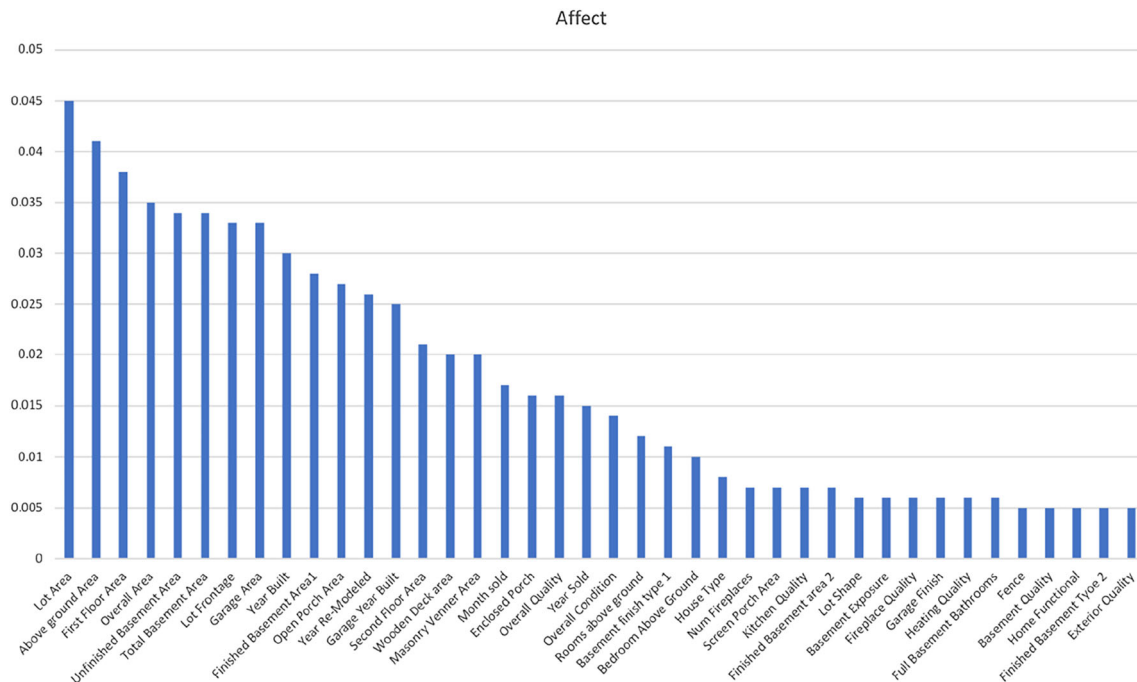


Fig. 15 Feature selection using gradient boosting

Table 2 RMSE scores

Model type	RMSE score
Linear regression	0.5188
Random forest	0.1499
Gradient boosting	0.1128

order to decide the price of housing society, it becomes a challenge with the competitive edge. In this paper, three modelling techniques have been used, such as the LASSO (Least Absolute Shrinkage and Selection Operator), a regression model that regularizes the selected predictor variable helps to decide the Sale price of the housing project. An automated variable selection approach using LASSO to process the relevant data have been used. Random forest, gradient boosting and linear regression modelling techniques are used to detect the predictive

feature selection to estimate the sale price in real estate by taking the Housing data set of Ames, Iowa. The classification and regression techniques using random forest in R has been implemented. Another technique, gradient boosting relies on the estimation that the next best possible model, when added to the former models, reduces the overall prediction error. The models were validated by using the accuracy measure RMSE (Random Mean Squared Error). It was computed for each model and compared to the values on the Kaggle website (internet). The experiment and results predicts that the out of all the models, the gradient boosting model returned the efficient results in depicting the predictor variables of sale price.

Further, the paper can be extended by looking at the other housing datasets and finding more accuracy of predictive modelling in depicting the sales price of real estate using artificial neural networks and deep learning techniques.

Appendix

SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.

MSSubClass: The building class
 MSZoning: The general zoning classification
 LotFrontage: Linear feet of street connected to property
 LotArea: Lot size in square feet
 Street: Type of road access
 Alley: Type of alley access
 LotShape: General shape of property
 LandContour: Flatness of the property
 Utilities: Type of utilities available
 LotConfig: Lot configuration
 LandSlope: Slope of property
 Neighborhood: Physical locations within Ames city limits
 Condition1: Proximity to main road or railroad
 Condition2: Proximity to main road or railroad (if a second is present)
 BldgType: Type of dwelling
 HouseStyle: Style of dwelling
 OverallQual: Overall material and finish quality
 OverallCond: Overall condition rating
 YearBuilt: Original construction date
 YearRemodAdd: Remodel date
 RoofStyle: Type of roof
 RoofMatl: Roof material
 Exterior1st: Exterior covering on house
 Exterior2nd: Exterior covering on house (if more than one material)
 MasVnrType: Masonry veneer type
 MasVnrArea: Masonry veneer area in square feet
 ExterQual: Exterior material quality
 ExterCond: Present condition of the material on the exterior
 Foundation: Type of foundation
 BsmtQual: Height of the basement
 BsmtCond: General condition of the basement
 BsmtExposure: Walkout or garden level basement walls
 BsmtFinType1: Quality of basement finished area
 BsmtFinSF1: Type 1 finished square feet
 BsmtFinType2: Quality of second finished area (if present)
 BsmtFinSF2: Type 2 finished square feet
 BsmtUnfSF: Unfinished square feet of basement area
 TotalBsmtSF: Total square feet of basement area
 Heating: Type of heating
 HeatingQC: Heating quality and condition
 CentralAir: Central air conditioning
 Electrical: Electrical system

References

- Battle LM (2013) Interactive visualization of big data leveraging databases for scalable computation. Massachusetts Institute of Technology
- Big data area with billion fortune: four business domains in community services. http://www.ffw.com.cn/1/107/463/169906_2.html
- Brown B, Chui M, Manyika J (2011) Are you ready for the era of 'big data'. McKinsey Q 4:24–35
- Dachuan C, Baoshan Z (2012) The applications of big data in housing information system. Inf Commun Technol 5:004
- Ghosh D, Chinnaiyan AM (2005) Classification and selection of biomarkers in genomic data using LASSO. J Biomed Biotechnology 2:147–154
- <http://money.163.com/13/0520/01/8V9GEDM300253B0H.html>
- <http://www.cnfsdata.com/default/bigdata.html>
- <http://www.pcpop.com/doc/0/931/931855.shtml>
- <https://calcalist.careebiz.com/credifi>
- [https://en.wikipedia.org/wiki/Regularization_\(mathematics\)](https://en.wikipedia.org/wiki/Regularization_(mathematics))
- <https://towardsdatascience.com/the-random-forest-algorithm-d457d499fcd>
- <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>
- <https://www.quora.com/What-are-the-applications-of-big-data-and-Hadoop-in-real-estate-1>
- https://www.sas.com/en_in/insights/analytics/big-data-analytics.html
- https://www.sas.com/en_us/insights/big-data/hadoop.html
- Huang H, Yin L (2014) Creating sustainable urban built environments: an application of hedonic house price models in Wuhan, China. J Hous Built Environ 30(2):1–17
- Juan Y (2013) The precise marketing based on big data in the real estate enterprises. Market Wkly 9:66–67
- Natekin A, Knoll A (2011) Gradient boosting machines, a tutorial. Front Neurobot 7:21
- Pantazopoulos A, Maragoudakis M (2018) Sports and nutrition data science using gradient boosting machines, pp 1–7
- Prabhakaran S (2019) Linear regression. Linear regression with R. <http://r-statistics.co/Linear-Regression.html>. Accessed 26 Mar 2019
- Qiuming C, Yunfeng G (2011) The relationship between bank credit and real estate prices: empirical analysis from provincial panel data in China. Rev Invest Stud 8:009
- Rahman F, Slepian M, Mitra A (2016) A novel big-data processing framework for healthcare applications: big-data-healthcare-in-a-box. In: 2016 IEEE international conference on big data (big data), Washington, DC, 2016, pp 3548–3555
- Realty marketing has entered the era of big data in 2013. <http://gd.qq.com/a/20130309/000055.htm>
- Recognize the big data the transition of realty companies needs soft power. <http://house.china.com.cn/chongqing/view/688439.htm>
- Regularization (mathematics). Wikipedia. January 04, 2019. Accessed ch 26, 2019
- Singh A (2017) Mining of social media data of university students. Educ Inf Technol 22:1515–1526. <https://doi.org/10.1007/s10639-016-9501-1>
- Singh A, Ranjan J (2016) A framework for mobile apps in colleges and universities: data mining perspective. Educ Inf Technol 21:643–654. <https://doi.org/10.1007/s10639-014-9345-5>
- Singh A, Rana A, Ranjan J (2015) Proposed analytical customer centric analytical model for automobile industry. Int J Data Min Model Manag 7(4):314–330
- Sonka S, IFAMR I (2014) Big data and the ag sector: more than lots of numbers. Int Food Agribus Manag Rev 17(1):1
- Tene O, Polonetsky J (2012) Privacy in the age of big data: a time for big decisions. Stanf Law Rev Online 64:63
- The age of big data in real estate, hard to save the sunset indus <http://news.nb.soufun.com/2013-07->
- The press conference on big data in real estate in 2013. <http://sy.jiwwu.com/news/1301701.html>
- Wang P, Kang M (2014) An empirical analysis on the housing prices in the Pearl River Delta Economic Region of China. Int J Urban Sci 18(1):103–114
- Wheeler D, Tiefelsdorf M (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. J Geogr Syst 7(2):161–187

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.