ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

House Hold Price Statistical Analysis through Linear Regression

Swathi Rathi

EAS SAP Analytics, Cognizant Technologies, Bangalore, India

ABSTRACT: Linear regression includes finding the best-fitting line through the core interests. The best-fitting line is understood as a regression curve. We took a dataset containing various attributes having certain values. In our study, we have implemented the Linear Regression on the dataset. In which we will analyze the dataset and find out the statistical information of each and every attribute for the final decision making process. Here we'll find some important statistical measures such as Coefficients, Standard Errors, and Standard Coefficients, tolerance, t-state and p-values of each attribute. These measures would be very helpful to find the characteristics of the dataset and its relevant variables. And due to this we can have a better understanding about the dataset and predicting variables.

KEYWORDS: Regression, Dataset, Coefficients, p-values, tolerance.

I. INTRODUCTION

Linear regression contains of finding the best-fitting line through the focuses. In direct statistical regression, we have a tendency to envision scores on one variable from the scores on a second variable. The variable we have a tendency to predicting is understood because the commonplace variable and is implicit as Y[1-5]. The variable we have a tendency to collection our gauges as for is understood because the marker variable and is recommended as X. Right once there's solely a solitary marker variable, the figure technique is termed clear regression. In basic statistical regression, the figures of Y once aforethought as a part of X structure a line. we are able to see that there's a positive association among X and Y[6-8]. just in case we might foresee Y from X, the upper the estimation of X, the upper your conjecture of Y[9-12].

Linear regression includes finding the best-fitting line through the core interests. The best-fitting line is understood as a regression curve. The corner to corner line is that the regression curve and involves the foretold score on Y for every potential estimation of X. The vertical lines from the concentrations to the regression curve address the missteps of need. The red purpose is astonishingly close to the regression line; its error of need is little. The other purpose is way on top of the regression curve and per se its bumble of estimate is big[13-14].

Linear Access that shows the connection betwixt scalier reaction and at least single illustrative factors. In the same, the direct indicator capacity is utilized to anticipate the outcome and it is additionally utilized for demonstrating the connection betwixt factors. Also, these exemplary are knows as Linear relapse.

II. LITERATURE SURVEY

Implementation of linear regression can be done in the predicting the house cost prediction relates to the main simple theories of identifying the relations between the features which are considering.

Regression is the methodology which will be used for the statistical analysis for certain information. The methodologies which are accepting the estimation of the relation among the different variables can perform the regression analysis. For suppose consider any continuous variables and we need to identify the relation between the certain variables or features we are considering those as the regression methodologies. We need to identify the dependent variable based on the different number of independent variables. This concept is very much useful for the predictive analysis based on domains or domain independence[15-17].

DOI: 10.15680/IJIRCCE.2019. 0705105

$$E(Y \mid X) = f(X, \beta)$$

It is an easy access methodology to implement the predictive analysis.

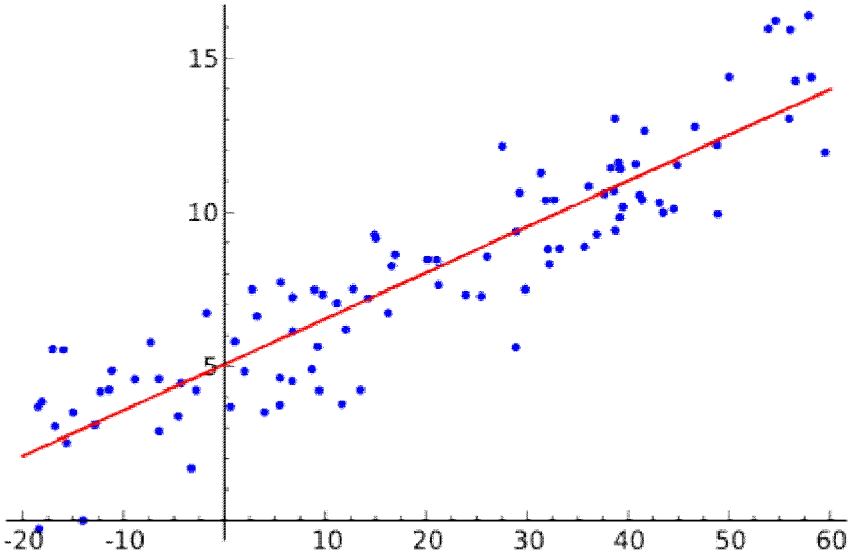


Figure 1: Scatterplot of the variables.

$$Y = f(X) = \beta 0 + \beta 1 * X$$

$\beta 0$ is the intercept of the line
$\beta 1$ is the slope of the line
Linear regression is the process of understanding the predictive analysis with the variable which are linearly seperable in nature. We have a two dimentional and multi dimensional space of the implementation and the requirement is to identify the outline of the connections. The linear models are supported with each other in the interactions among the variables which are in action with the prediction model. The prediction can be a supervised or unsupervised. But the main criteria is to understand the importance of regression models in this instance.

Before implementing any prediction model we need to understand whether there is any underlying relationship between the variables which are being considered. The relation between the variables will be identified using scatterplot. The graphical representations of the data will be useful for understanding the importance of each model. The relationship will indicate based on their existence of the features. For example if the instance is existed then the importance will be increased in the model. If the existence is not there, then the prediction model accuracy can be increased[18-19].

Mean Squared error and Root Mean Squared Error are the most used features which are used to calculate the error rate of the models. If the error rate is more then we need to manipulate the model and we need to remodel that.

$$\text{RSS} = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

RMSE *(Square root of MSE) = √ (**MSE**)*

The additional number of variables will add more dimension to the model.

$$Y = f(X) = \beta 0 + \beta 1 * X1 + \beta 1 * X2 + \beta 1 * X3$$

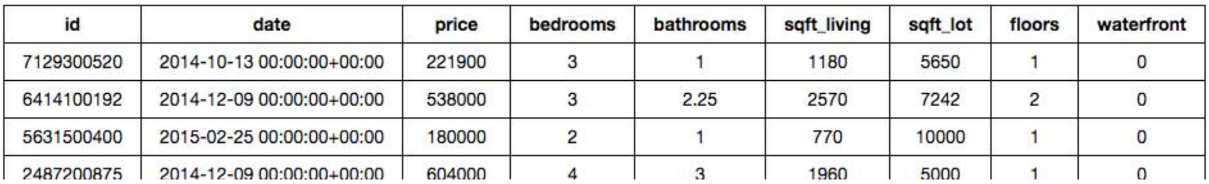
Environment Setup

1. Python Programming
2. Graphlab lybrary
3. S Frame (similar to Pandas Data Frame)

DATA LOADING

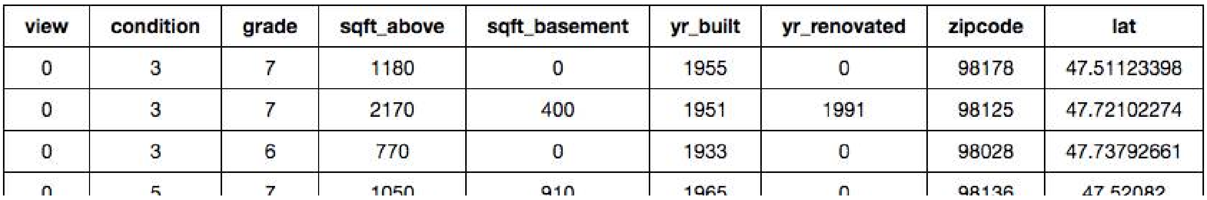
Here we are loading the house cost prediction dataset which consists of the following features which are mentioned following. The data which we are considering are the important for the prediction model design and implementation.

Table 1: Dataset Structure



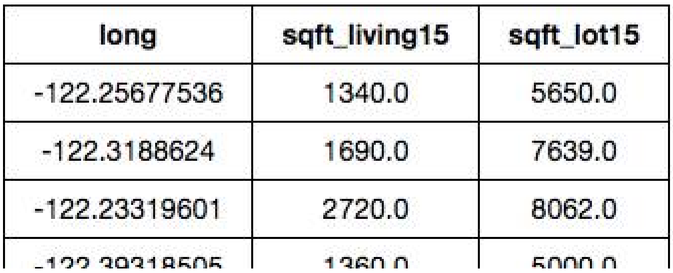
id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
7129300520	2014-10-13 00:00:00+00:00	221900	3	1	1180	5650	1	0
6414100192	2014-12-09 00:00:00+00:00	538000	3	2.25	2570	7242	2	0
5631500400	2015-02-25 00:00:00+00:00	180000	2	1	770	10000	1	0
2487200875	2014-12-09 00:00:00+00:00	604000	4	3	1960	5000	1	0

Table 2: Dataset and feature values



view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat
0	3	7	1180	0	1955	0	98178	47.51123398
0	3	7	2170	400	1951	1991	98125	47.72102274
0	3	6	770	0	1933	0	98028	47.73792661
0	5	7	1050	910	1965	0	98136	47.52082

Table 3: Dataset and Values



long	sqft_living15	sqft_lot15
-122.25677536	1340.0	5650.0
-122.3188624	1690.0	7639.0
-122.23319601	2720.0	8062.0
-122.39318505	1360.0	5000.0

Lest make the relationships between the variables which are required most.

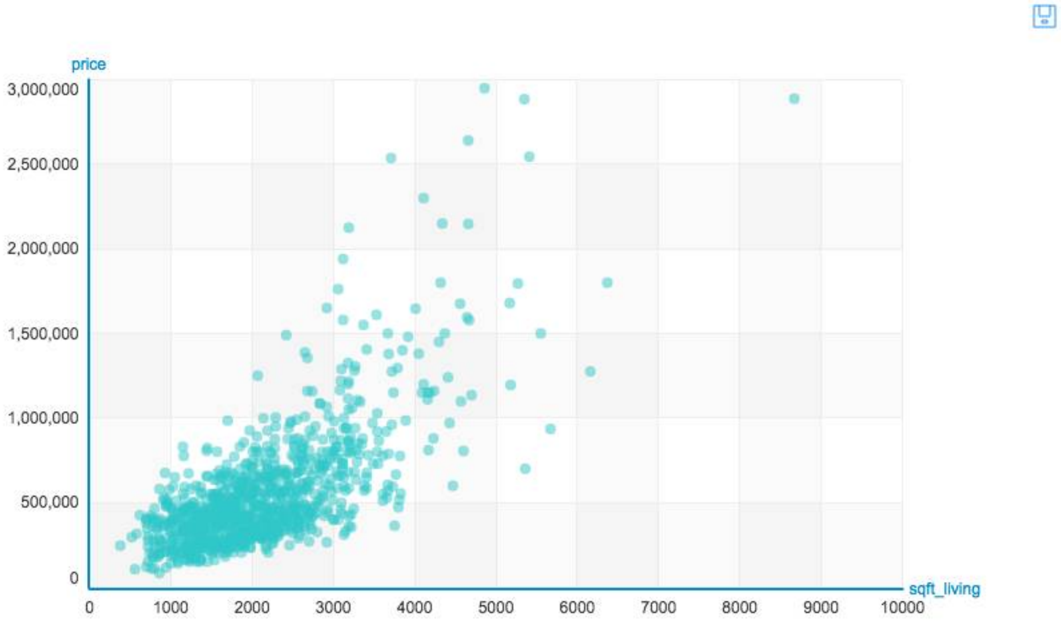


Figure 2: Scatterplot for the relationship identification among features

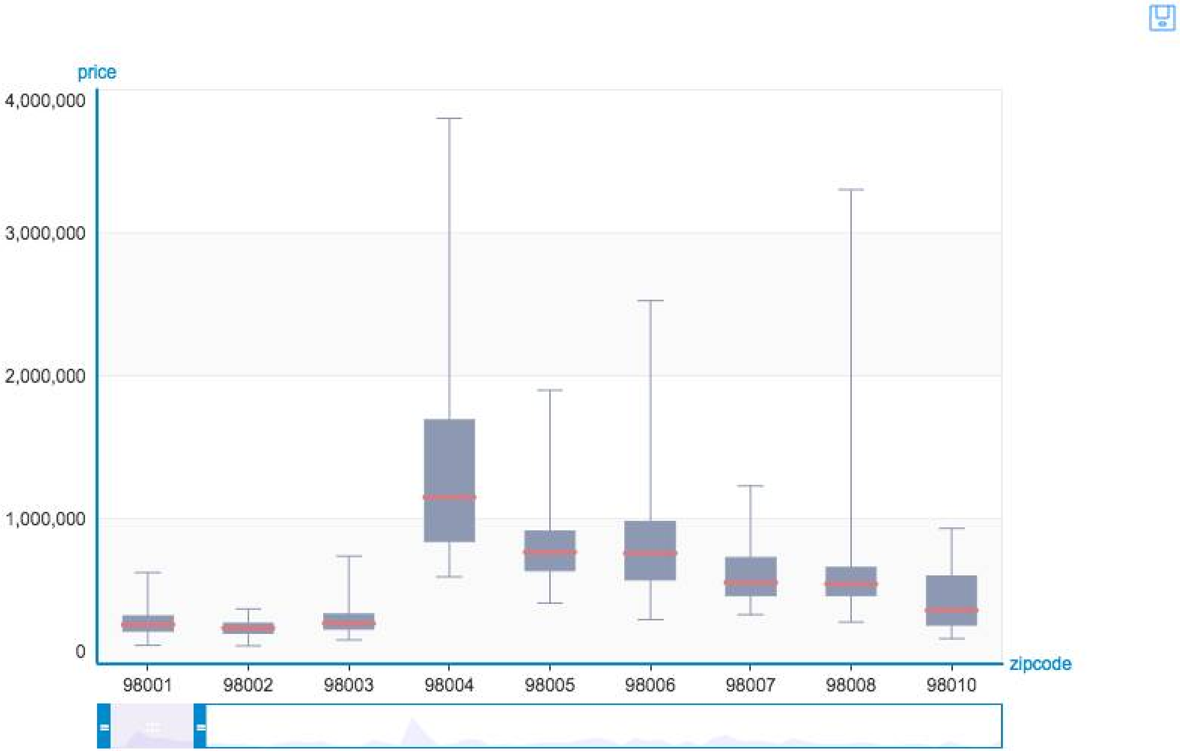
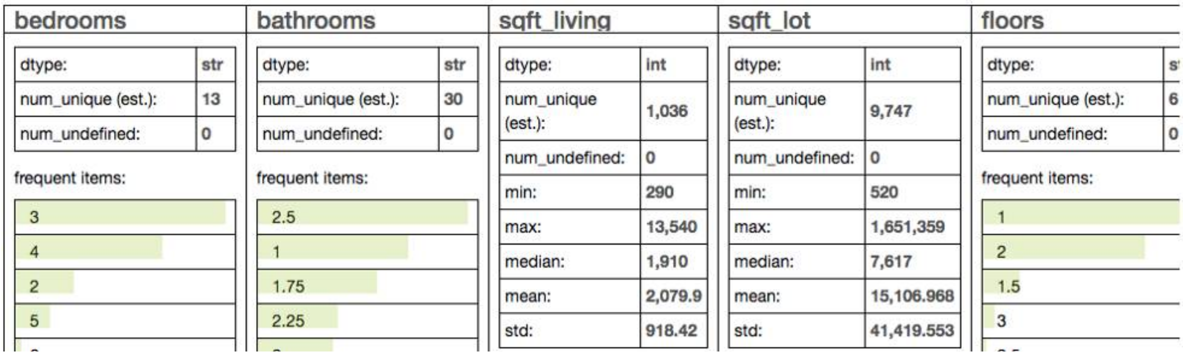


Figure 3: Indicates the range of pricing with respect to the Zipcode

We need to plot the variables based on the features and the considered features will give the following result.

Table 4: Outcome of the prediction



bedrooms		bathrooms		sqft_living		sqft_lot		floors	
dtype:	str	dtype:	str	dtype:	int	dtype:	int	dtype:	s
num_unique (est.):	13	num_unique (est.):	30	num_unique (est.):	1,036	num_unique (est.):	9,747	num_unique (est.):	6
num_undefined:	0	num_undefined:	0	num_undefined:	0	num_undefined:	0	num_undefined:	0
frequent items:		frequent items:		min:	290	min:	520	frequent items:	
3		2.5		max:	13,540	max:	1,651,359	1	
4		1		median:	1,910	median:	7,617	2	
2		1.75		mean:	2,079.9	mean:	15,106.968	1.5	
5		2.25		std:	918.42	std:	41,419.553	3	

IV. PREDICTIVE ANALYTICS

Split the test and train dataset in the form of 80% and 20%. The training set will be 80% the remaining will be test set.
Fit the linear regression mode into the variable considered.
Plot the variables with the conditions we have with the features.
Blue dots indicate the price of the houses based on the conditions we have and the green line indicates the predicted value of the conditions. The dependant variables are considered as the green line. That is the hyper plane which divides the models[20].

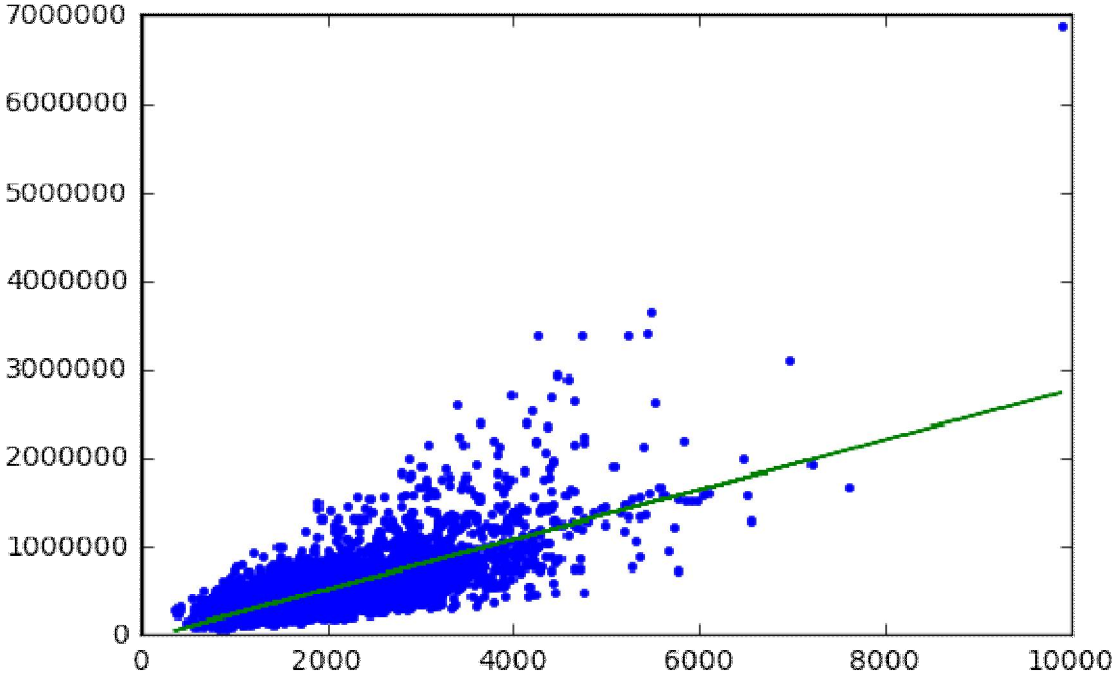
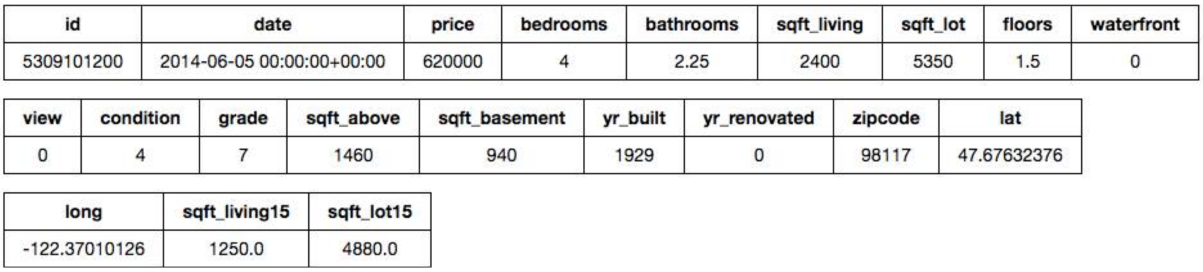


Figure 4: Scatterplot with error rectification

Now we shall select a house and try to predict the value using “sqft_model”.

Table 5: Features without error rate



id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
5309101200	2014-06-05 00:00:00+00:00	620000	4	2.25	2400	5350	1.5	0

view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat
0	4	7	1460	940	1929	0	98117	47.67632376

long	sqft_living15	sqft_lot15
-122.37010126	1250.0	4880.0

V. PROPOSED WORK

In our study, we have implemented the Linear Regression on the dataset. In which we will analyze the dataset and find out the statistical information of each and every attribute for the final decision making process. Here we’ll find some important statistical measures such as Coefficients, Standard Errors, Standard Coefficients, tolerance, t-state and p-values of each attribute[21-22]. These statistical measures would be very helpful to identify the characteristics of dataset and its attributes. In previous studies there is lot of work in this domain to compare different algorithms or finding predictions etc. But in our study we’ll achieve two major objectives one is that which predictive variable performs its best to predict the outcome and another one is that the set of predictor attribute performs its best to find the best outcome or dependent variable. To do so, we have applied the Linear Regression Model on dataset using linear regression operator. This would reflect the major statistical measures mentioned before to *achieve the objectives discussed above.*

Linear Regression

(i) Process:

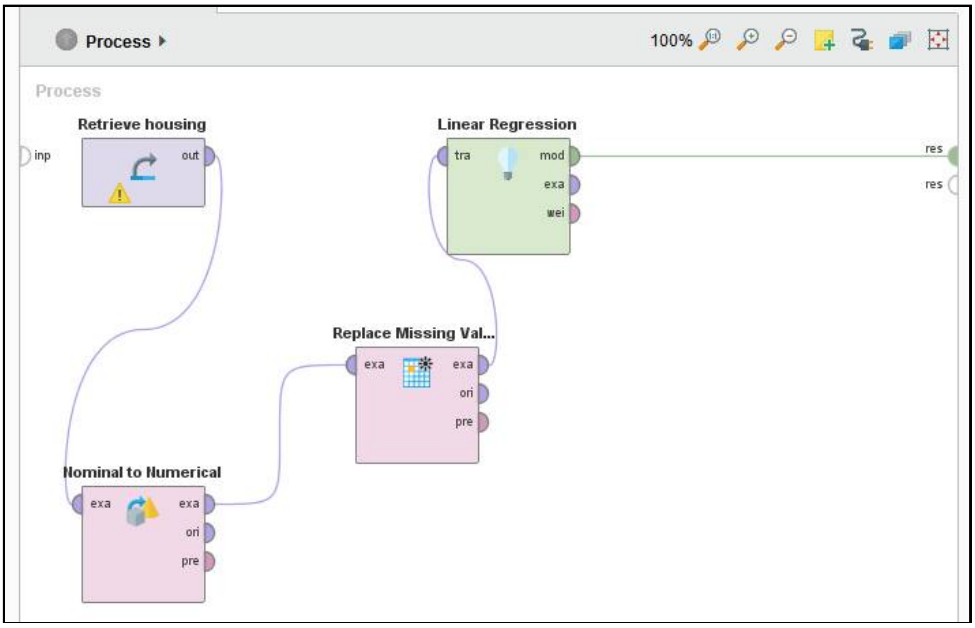


Figure 5 Simulator implementation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

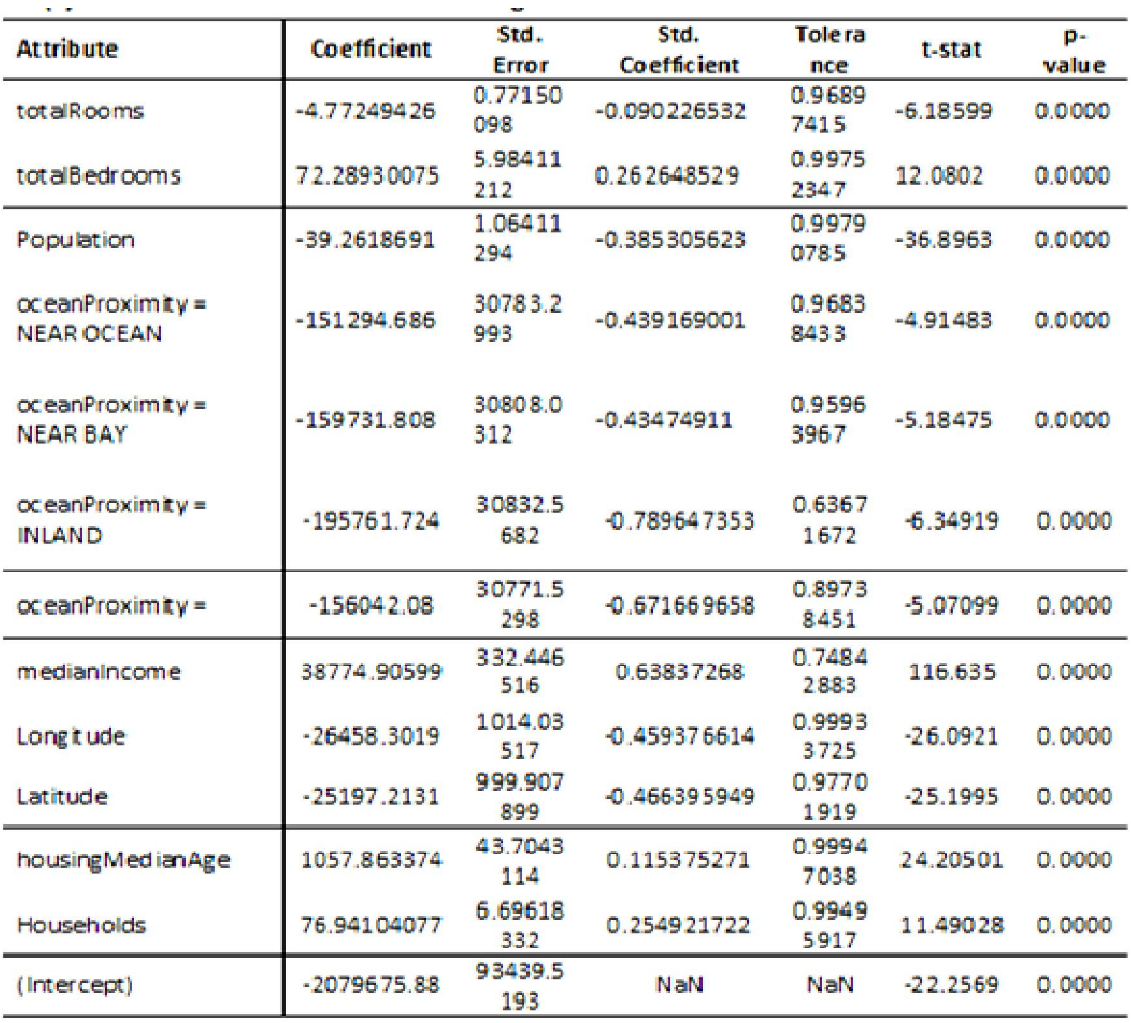
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Raised visualization demonstrates the Process Exemplary of Linear Regression. This procedure exemplary incorporates the database, Linear Regression Technique as fundamental segments. Here we have 2 more aiding Techniques are likewise admitted for example Supplant Lost quality Technique and Nominal to Statistical Technique. These Techniques are utilized on the grounds that our database has lost qualities as we examined in the EDA ("all out rooms =207") and a polemical Variable "Sea Proximity". We also have various component determination techniques alternatives of Linear Regression, for example, M5prime, avaricious, T-Examination and Iterative T-Examination. We connected M5 prime component determination technique on the database through Linear Regression.
For our situation we achieved the Linear Regression on the database. Because of that, we accomplished some fundamental objectives i.e., we can foresee the result, we would have a thought of mistakes for blunder decrease since it can fit a prescient exemplary to the compassionate estimations of database of qualities and different informative Variables or factors. As should be obvious in the given beneath table.

Table 6: Result of LR Model



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-stat	p-value
totalRooms	-4.77249426	0.77150098	-0.090226532	0.96897415	-6.18599	0.0000
totalBedrooms	72.28930075	5.98411212	0.262648529	0.99752347	12.0802	0.0000
Population	-39.2618691	1.06411294	-0.385305623	0.99790785	-36.8963	0.0000
oceanProximity = NEAR OCEAN	-151294.686	30783.2993	-0.439169001	0.96838433	-4.91483	0.0000
oceanProximity = NEAR BAY	-159731.808	30808.0312	-0.43474911	0.95963967	-5.18475	0.0000
oceanProximity = INLAND	-195761.724	30832.5682	-0.789647353	0.63671672	-6.34919	0.0000
oceanProximity =	-156042.08	30771.5298	-0.671669658	0.89738451	-5.07099	0.0000
medianIncome	38774.90599	332.446516	0.63837268	0.74842883	116.635	0.0000
Longitude	-26458.3019	1014.03517	-0.459376614	0.99933725	-26.0921	0.0000
Latitude	-25197.2131	999.907899	-0.466395949	0.97701919	-25.1995	0.0000
housingMedianAge	1057.863374	43.7043114	0.115375271	0.99947038	24.20501	0.0000
Households	76.94104077	6.69618332	0.254921722	0.99495917	11.49028	0.0000
(Intercept)	-2079675.88	93439.5193	NaN	NaN	-22.2569	0.0000

In raised table we have some factual data which demonstrates the reliance and relationship among the factors or Variables as measurements, for example, Coadjuvants and sexually transmitted disease. Co- adjuvant which is a statistical steady for portraying the relationship among factors. The positive estimation of coadjuvants demonstrates the great Examination conditions among Variables. For example, housingmedianage that have the estimation of 1057.863374, median income has estimation of 38774.0599, etc.

Next is about blunders, we have a few mistakes data in our outcome during investigation. Like sexually transmitted disease. Mistakes. It is std. Dev. estimation. As should be obvious contained into outcome, the minimum mistake Variable is a complete room having 0.772 qualities. Furthermore, greaExamination mistake Variable named "Ocean proximity"=ISLAND is having 30832.568.

The resistance esteem characterizes the one free Variable on all needy variable or Variables. Each Variable having the estimation of 0.637 to 0.999 (min to max). This demonstrates the reliance of Variable on others.

T-detail is a T-measurements is co adjuvant. It was determined when it is partitioned by its standard mistake. The p-esteem means the relationship of factors. As should be obvious in our outcome the p-estimation of the considerable number of factors is 0.000, it implies the likelihood of finding the relationships is outrageous. So we should dismiss the invalid theory for the elective speculation. It implies all Variables are might be theoretical for forecast of outcome.

VI. CONCLUSION

The implementation of the regression analysis for the classification problem defined with the sample outputs of the implementation. The data consists of the basic level of the errors like missing values and other important issues. The missing values can be handled using the pre-processing methodology and the implementation of the error handling will be done by the basic architectures like CNB classifier and other sort of classifiers which are used in the real world scenarios for the major problems. The most effective model for the prediction analysis for the classification problem is CNB classifier with highest accuracy and the low error rate with the justified cost function.

REFERENCES

[1] K. B. To and L. M. Napolitano, ``Common complications in the critically ill patient,'' *Surgical Clinics North Amer.*, vol. 92, no. 6, pp. 1519_1557, 2012.

[2] C. M. Wollschlager and A. R. Conrad, ``Common complications in critically ill patients,'' *Disease-a-Month*, vol. 34, no. 5, pp. 225_293, 1988.

[3] S. V. Desai, T. J. Law, and D. M. Needham, ``Long-term complications of critical care,'' *Critical Care Med.*, vol. 39, no. 2, pp. 371_379, 2011.

[4] N. A. Halpern, S. M. Pastores, J. M. Oropello, and V. Kvetan, ``Critical care medicine in the United States: Addressing the intensivist shortage and image of the specialty,'' *Critical Care Med.*, vol. 41, no. 12, pp. 2754_2761, 2013.

[5] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, ``Machine learning and decision support in critical care,'' *Proc. IEEE*, vol. 104, no. 2, pp. 444_466, Feb. 2016.

[6] O. Badawi *et al.*, ``Making big data useful for health care: A summary of the inaugural MIT critical data conference,'' *JMIR Med. Informat.*, vol. 2, no. 2, p. e22, 2014.

[7] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics*, vol. 36. Boca Raton, FL, USA: CRC Press, 2015.

[8] D. Gotz, H. Stavropoulos, J. Sun, and F. Wang, ``ICDA: A platform for intelligent care delivery analytics,'' in *Proc. AMIA Annu. Symp.*, 2012, pp. 264_273.

[9] A. Perer and J. Sun, ``Matrix_ow: Temporal network visual analytics to track symptom evolution during disease progression,'' in *Proc. AMIA Annu. Symp.*, 2012, pp. 716_725.

[10] Y. Mao,W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, ``An integrated data mining approach to real-time clinical monitoring and deterioration warning,'' in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. 2012, pp. 1140_1148.

[11] J. Wiens, E. Horvitz, and J. V. Guttag, ``Patient risk strati_cation for hospital-associated C. Diff as a time-series classi_cation task,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 467_475.

[12] S. Saria, D. Koller, and A. Penn, ``Learning individual and population level traits from clinical temporal data,'' in *Neural Inf. Process. Syst. (NIPS), Predictive Models Personalized Med. Workshop*, 2010.

[13] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, ``Multitask Gaussian processes for multivariate physiological time-series analysis,'' *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1,pp. 314_322, Jan. 2015.

[14] M. Ghassemi *et al.*, ``Amultivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data,'' in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 446_453.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

[15] I. Batal, H. Valizadegan, G. F. Cooper, and M. Hauskrecht, ``A pattern mining approach for classifying multivariate temporal data," in *Proc. IEEE Int. Conf. Bioinformatics Biomed. (BIBM)*, 2011, pp. 358_365.
[16] T. A. Lasko, ``Ef_cient inference of Gaussian-process-modulated renewal processes with application to medical event data," in *Proc. Uncertainty Artif. Intell.*, 2014, p. 469_476.
[17] K. L. C. Barajas and R. Akella, ``Dynamically modeling patient's health state from electronic medical records: A time series approach," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015,pp. 69_78.
[18] X. Wang, D. Sontag, and F. Wang, ``Unsupervised learning of disease progression models," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 85_94.
[19] M. J. Cohen, A. D. Grossman, D. Morabito, M. M. Knudson, A. J. Butte, and G. T. Manley, ``Identi_cation of complex metabolic states in critically injured patients using bioinformatic cluster analysis," *Critical Care*, vol. 14, no. 1, p. 1, 2010.
[20] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, ``Modeling disease progression via fused sparse group lasso," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1095_1103.
[21] E. Choi, N. Du, R. Chen, L. Song, and J. Sun, ``Constructing disease network and temporal progression model via context-sensitive hawkes process," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2015, pp. 721_726.
[22] R. Pivovarov, A. J. Perotte, E. Grave, J. Angiolillo, C. H. Wiggins, and N. Elhadad, ``Learning probabilistic phenotypes from heterogeneous HER data," *J. Biomed. Informat.*, vol. 58, pp. 156_165, Dec. 2015.