IS665102 DATA ANALYTICS FOR INFORMATION SYSTEMS

Under the Guidance of Dr. Lin Lin

ABSTRACT

Data Mining of House Pricing Dataset using Linear Regression Model & Decision Tree Model, comparing the two models and finding the suitable one.

Group 6 - Team Members

Akash Asli
Atish Nayak
Likhit Jain
Pooja Suvarna
Shubhangi Rakhe

Spring 2018

Contents

1. Background

While buying a house it is important to know what factors affect the most, from a buyer's perspective. Does getting a place in a neighbour affects the decision to own a house? Does the setting of house affect the decision of buyer? Or does the area of the house make person go for the house? There are number of vast different factors like number of floors, waterfront, condition of the house, area of the basement per sq. Ft. and construction design. In this project we will be trying to answer some of the above questions.

We will be trying to predict the price of a house based on certain factors. We will accomplish this by analysing the data set to build a price predictor model, which will tell us the price for a house. This predictor can also be used by buyers to decide on which type of house to own given a house setting.

2. Problem Statement

We have chosen 'House Pricing Dataset' for modelling of a Price Predictor. The model construction will have conducted using Regression and Decision Tree data mining algorithms which will predict the price of house based on dimensions like area sq. ft., no, of bedrooms, no. of bathrooms, no. of offers, presence of bricks in house construction and the type of neighbourhood. At the end we will be comparing the performance of both the algorithms to find out which is better.

3. Dataset

The 'House Pricing Dataset' was taken from Kaggle.com. It consists of 128 observations and 7 unique descriptors.

Dataset has been attached below

Microsoft Excel Worksheet

House Pricing Dataset.xlsx

They are outlined as follows:

Sr. No.	Variable	Data type	IP/OP	Description
1	Price	Numeric	Output	These variables tell us the selling price of a house
2	SqFt	Numeric	Input	This describes the total sq. ft. surface area of the house
3	Bedrooms	Numeric	Input	This variable depicts the number of bedrooms whose values fall in the range of 2-5
4	Bathrooms	Numeric	Input	Specify the number of bathrooms, either 2 or 3
5	Offers	Numeric	Input	Gives us the number of offers for the house ranging in between 1 and 6
6	Brick	Binomial	Input	Specify whether the house is built either by building bricks or no.
7	Neighbourhood	Textual	Input	This descriptor mentions neighbourhood around the house in four directions.

4. Experiment Design

4.1. Regression

Regression analysis is a set of statistical processes for estimating the relationships among variables. The focus of regression is finding out the relationship between dependent variable and independent variable. It is a supervised learning method, widely used for predicting and forecasting. Regression analysis is performed by comparing the observations of dependent variables to the observations of each of the independent variables by plotting it on the graph. The tread line through all the points, then summarizes the relationship between dependent and independent variable.

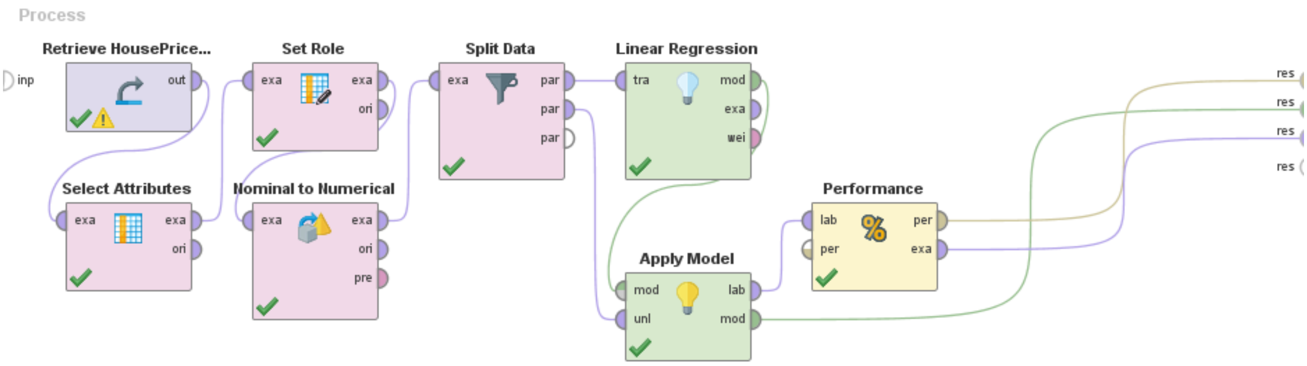
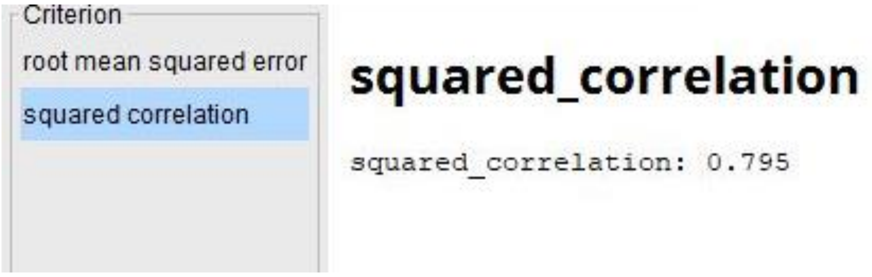


Figure 1: Regression Model Design

As shown in the Figure 1, the setup for Regression Mining is constructed in such a way. After retrieving data from the file, we selected the independent variables which had most impact on the dependent attribute Price. After setting the appropriate roles, the data type of certain variables like brick need to be changed to binomial. Then we split the data in to training dataset and testing data set. After performing linear regression on the data, we applied it on the testing dataset and used the Performance block to get the performance metrics.



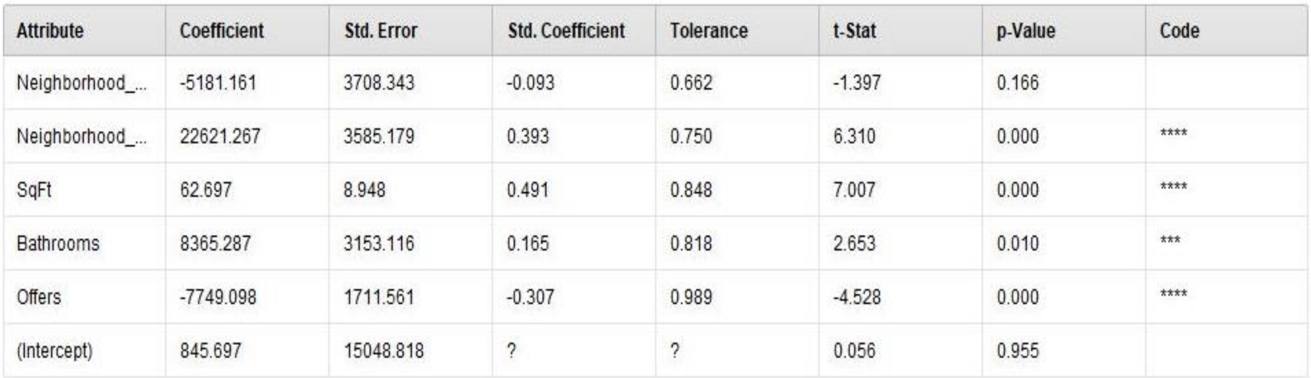
Criterion
root mean squared error
squared correlation

squared_correlation

```
squared_correlation: 0.795
```

Figure 2. Regression R-squared

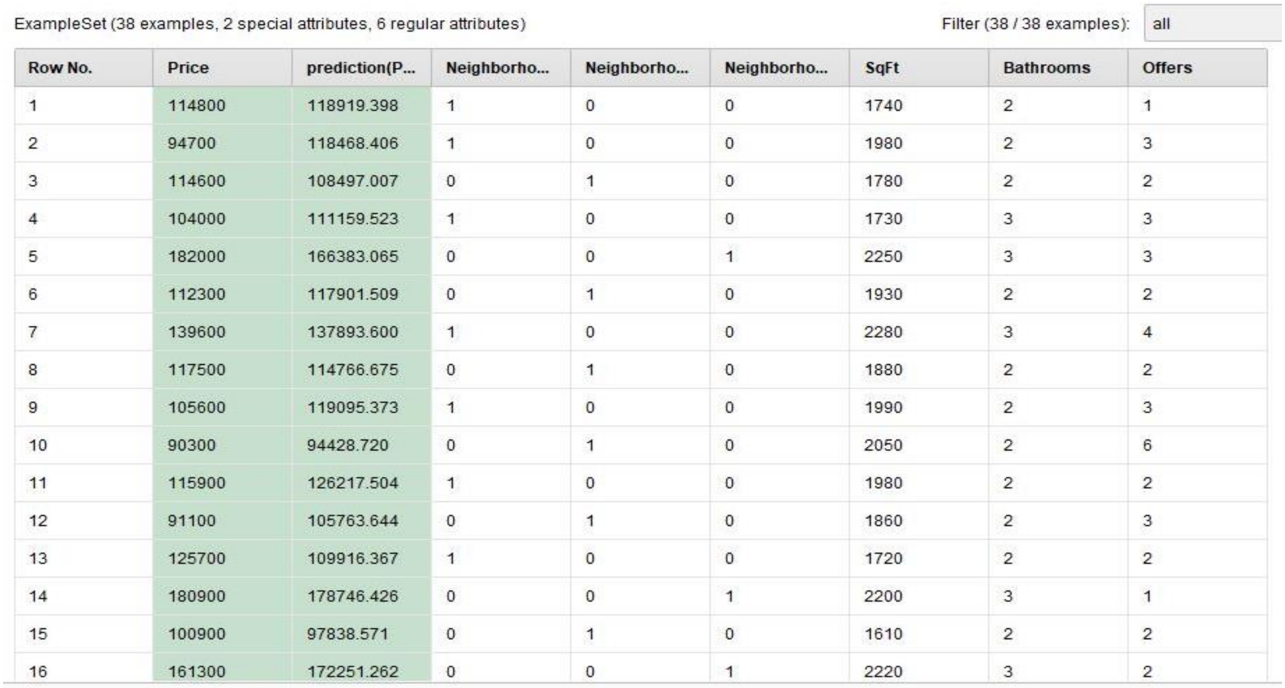
The Value for the R^2 as shown in Figure 2 is 0.79 that means its 79%, which is a fairly good value. From this we can conclude that our model is good model.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Neighborhood_...	-5181.161	3708.343	-0.093	0.662	-1.397	0.166	
Neighborhood_...	22621.267	3585.179	0.393	0.750	6.310	0.000	****
SqFt	62.697	8.948	0.491	0.848	7.007	0.000	****
Bathrooms	8365.287	3153.116	0.165	0.818	2.653	0.010	***
Offers	-7749.098	1711.561	-0.307	0.989	-4.528	0.000	****
(Intercept)	845.697	15048.818	?	?	0.056	0.955	

Figure 3. Regression T-stats and P-value

The Figure 3 shows the T-Stats Values for the variables that we selected as attributes while building the model. For a good performance it is necessary that the T-Stat variables be <-2 and >2The values for variables Neighbourhood, SqFt, Bathrooms and Offers are good in the required range. We will not be considering Neighbourhood because its T-stats is -1.4 and hence it will not be considered. The p-value shows us that it >0.005, hence we can reject the null Hypothesis. The p-value for Neighbourhood is >0.005 which is weak evidence against null hypothesis, hence for this variable it cannot be rejected.



ExampleSet (38 examples, 2 special attributes, 6 regular attributes)

Filter (38 / 38 examples): all

Row No.	Price	prediction(P...	Neighborho...	Neighborho...	Neighborho...	SqFt	Bathrooms	Offers
1	114800	118919.398	1	0	0	1740	2	1
2	94700	118468.406	1	0	0	1980	2	3
3	114600	108497.007	0	1	0	1780	2	2
4	104000	111159.523	1	0	0	1730	3	3
5	182000	166383.065	0	0	1	2250	3	3
6	112300	117901.509	0	1	0	1930	2	2
7	139600	137893.600	1	0	0	2280	3	4
8	117500	114766.675	0	1	0	1880	2	2
9	105600	119095.373	1	0	0	1990	2	3
10	90300	94428.720	0	1	0	2050	2	6
11	115900	126217.504	1	0	0	1980	2	2
12	91100	105763.644	0	1	0	1860	2	3
13	125700	109916.367	1	0	0	1720	2	2
14	180900	178746.426	0	0	1	2200	3	1
15	100900	97838.571	0	1	0	1610	2	2
16	161300	172251.262	0	0	1	2220	3	2

Figure 4. Regression Prediction

After training the model with training dataset, to check its working efficiency, we applied the model on the Testing Dataset. The result we obtained is shown in Figure 4.

4.2. Decision Tree

Decision Trees are a nonparametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision tree breaks the data set into smaller parts and at the same time an associated decision tree is incrementally developed. Decision trees are an excellent tool because it provides a highly effective structure which you can lay out options and investigate the possible outcomes of choosing those options.

Usage of Decision Tree would help in the analysis of our dataset in the following manner:

- Decision Tree makes explicit all possible alternatives and traces each alternative to its conclusion in a single view, allowing for easy comparison among the various alternatives.
- Reduces ambiguity in decision making while choosing one with good monetary value.

Figure 5 represents the Decision Tree Process built as follows

- Selected Houce Pricing Dataset
- Chose the attributes needed for the analysis
- Set 'Price' as label for our analysis
- Convert the Nominal attributes to Numerical
- Split the data into 0.7 for training and 0.3 for Testing Dataset
- Apply Decision Tree Model on the data.

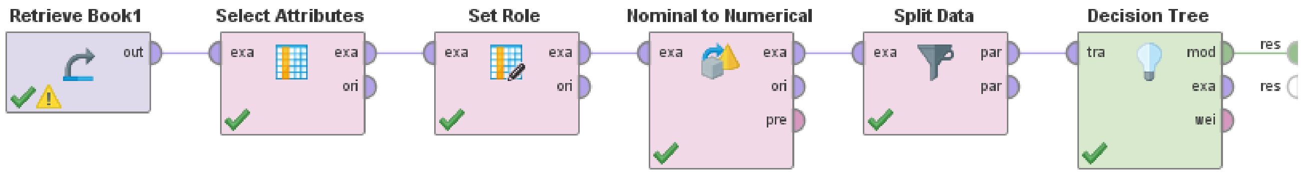


Figure 5. Decision Tree Model Using Rapid Miner

After executing the Decision Tree Process, the following Decision Tree was generated as shown in Figure 6.

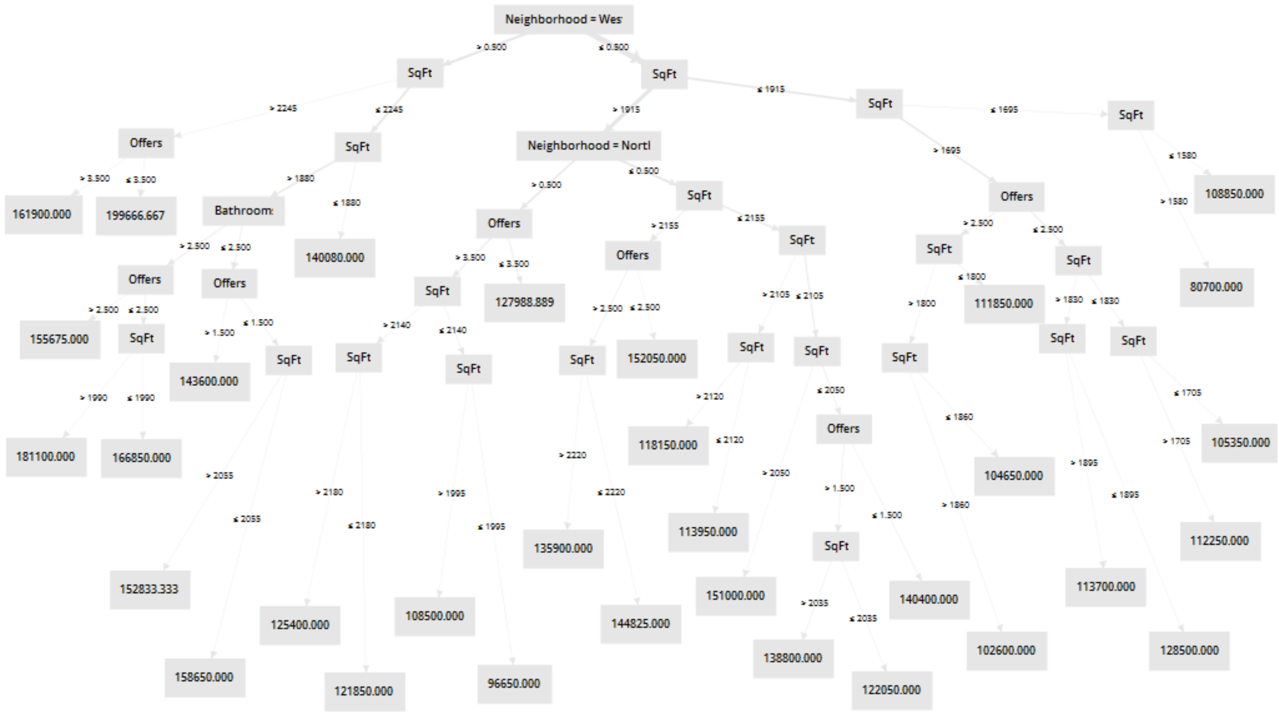
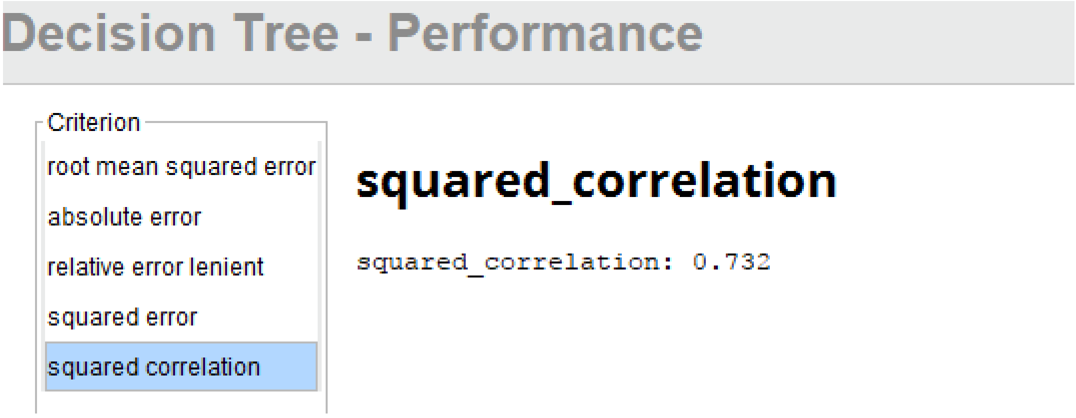


Figure 6. Decision Tree Process Results

The performance of the model can be explained using Squared Correlation (Figure 7), F-measure (Figure 13), Precision (Figure 10), Recall (Figure 11) and Lift (Figure 15).

We have classified the data into two sections as follows,

- Range 1 – from 69,100 to 140,150
- Range 2 – from 140,151 to 211,200


Decision Tree - Performance

squared_correlation

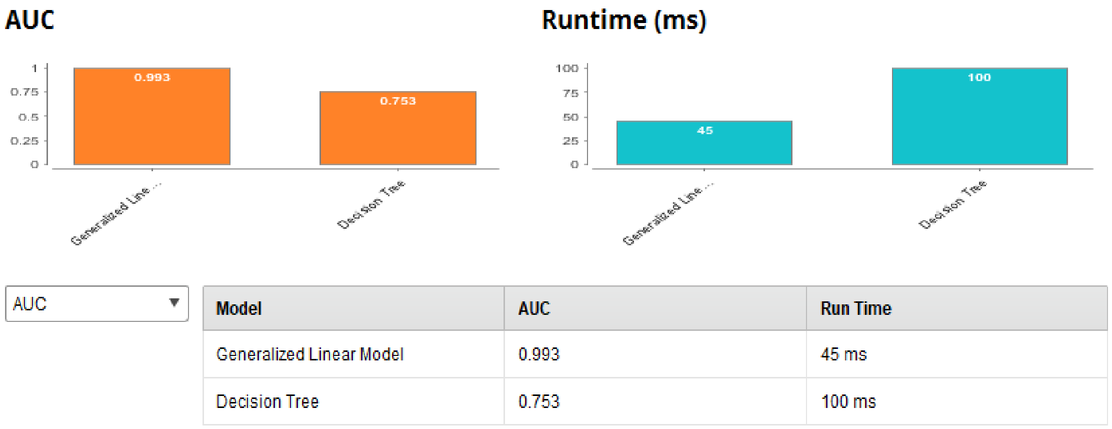
```
squared_correlation: 0.732
```

Figure 7. Squared Correlation

The R-squared is the square of the correlation r between the predicted and actual values. In Figure 7, the R-squared is 73.2% which says that the data is good enough to serve our analysis and predict desirable results.

5. Analysis

5.1. Area Under Curve(AUC)



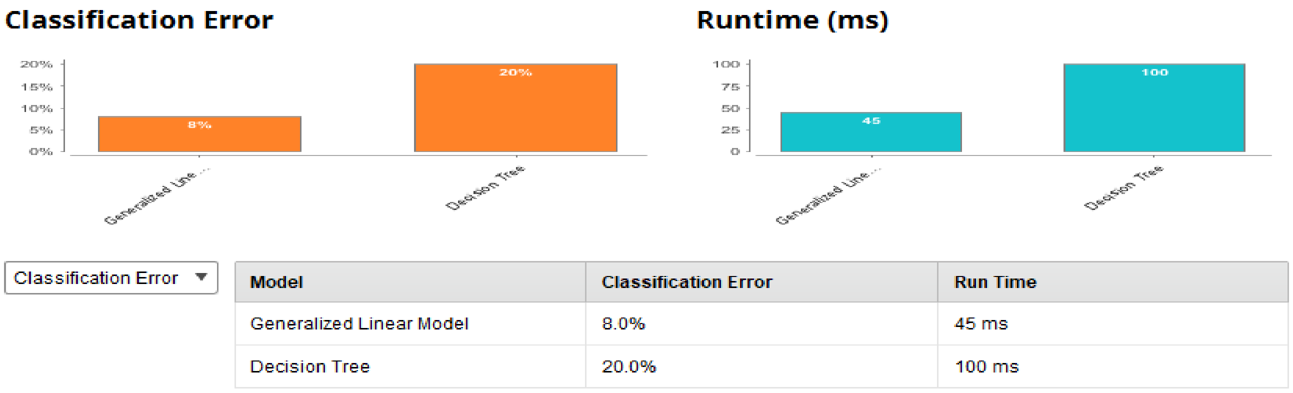
Model	AUC	Run Time
Generalized Linear Model	0.993	45 ms
Decision Tree	0.753	100 ms

Figure 8. AUC difference in Linear Regression and Decision Tree

AUC (Area Under Curve) is used in classification model to determine which of the used model predicts the classes best. Here, True Positive rates are plotted against False Positive Rates.

From Figure 8, we can state that Linear Regression Model has better AUC, i.e., it is a better predictive model for this dataset.

5.2. Classification Error



Model	Classification Error	Run Time
Generalized Linear Model	8.0%	45 ms
Decision Tree	20.0%	100 ms

Figure 9. Classification Error in Linear Regression and Decision Tree

Classification error is maximum when records are equally distributed among all classes, implying least interesting information and minimum when all records belong to one class, implying most interesting information.

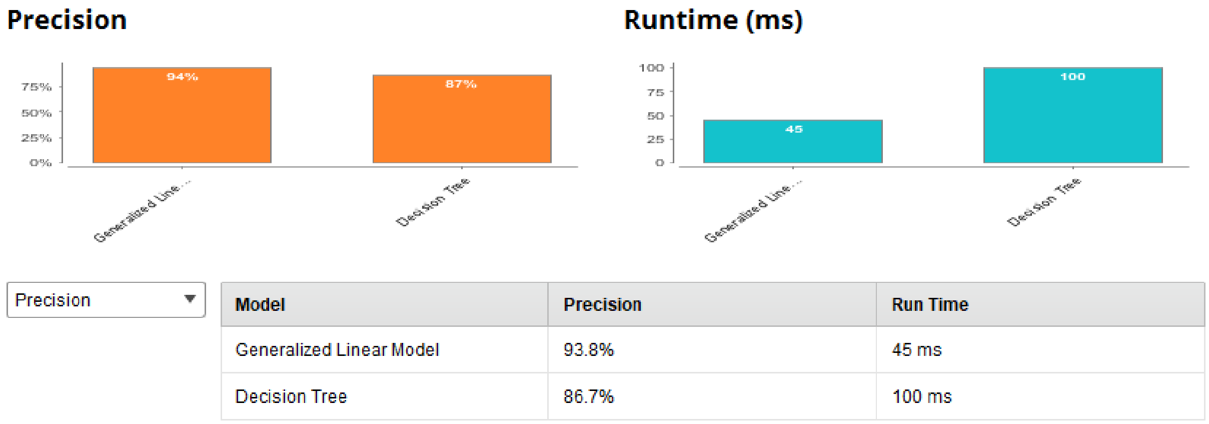
As per Figure 9, although the classification error of Decision Tree is good, Linear Regression Model can give most useful insights needed in our analysis comparatively.

5.3. Precision and Recall

Precision also known as Positive Predicted Value is the fraction of relevant instances among the relevant instances, while Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.

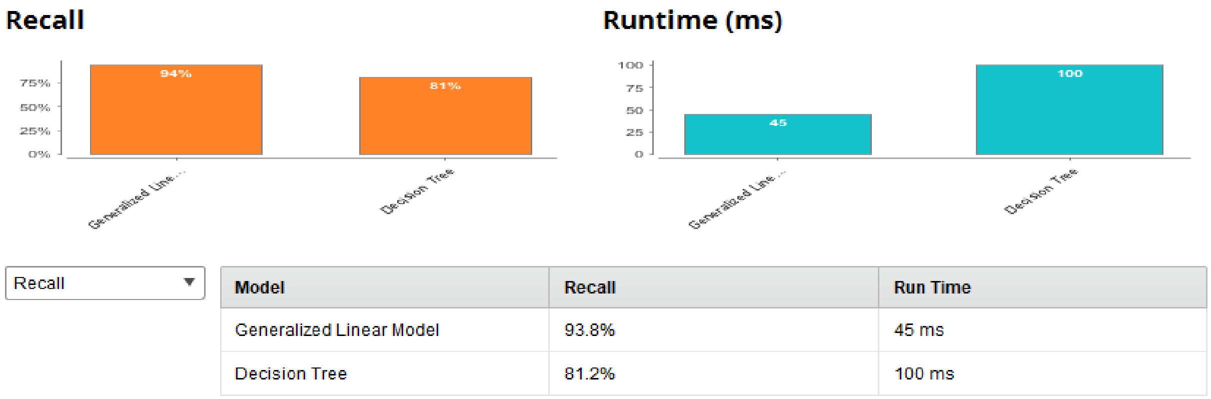
Therefore, we can say that, Precision is "How Useful the search results are", and recall is "How complete the results are".

Figure 10 and Figure 11 shows Precision and Recall of Linear Regression Model and Decision Tree respectively.



Model	Precision	Run Time
Generalized Linear Model	93.8%	45 ms
Decision Tree	86.7%	100 ms

Figure 10. Precision in Linear Regression and Decision Tree



Model	Recall	Run Time
Generalized Linear Model	93.8%	45 ms
Decision Tree	81.2%	100 ms

Figure 11. Recall in Linear Regression and Decision Tree

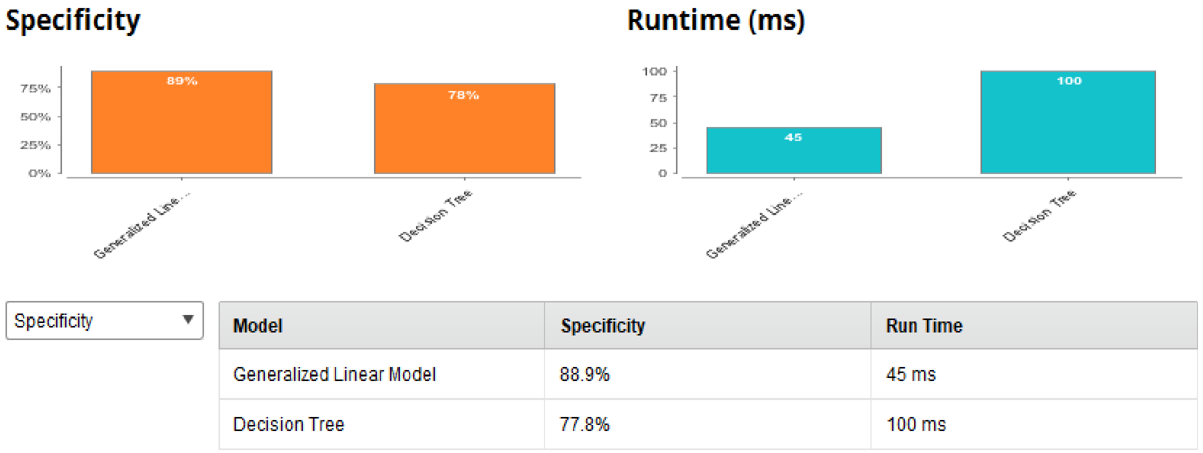
Figure 10 shows Precision of Linear Regression Model is better than Decision Tree, i.e., the measure of exactness or quality is more than Decision Tree. Therefore, Linear Regression Model will return substantially more relevant results than irrelevant ones than Decision Tree comparatively.

Figure 11 shows Recall of Linear Regression Model is better than Decision Tree, i.e., the measure of completeness or quantity is more than Decision Tree. Therefore, Linear Regression Model will return most of the relevant results than Decision Tree comparatively.

5.4. Sensitivity and Specificity

Sensitivity (True positive rate) measures the proportion of positives that are correctly identified. This is also known as Recall which is shown in Figure 11.

Specificity (True Negative Rate) measures the proportion of negatives that are correctly identified.

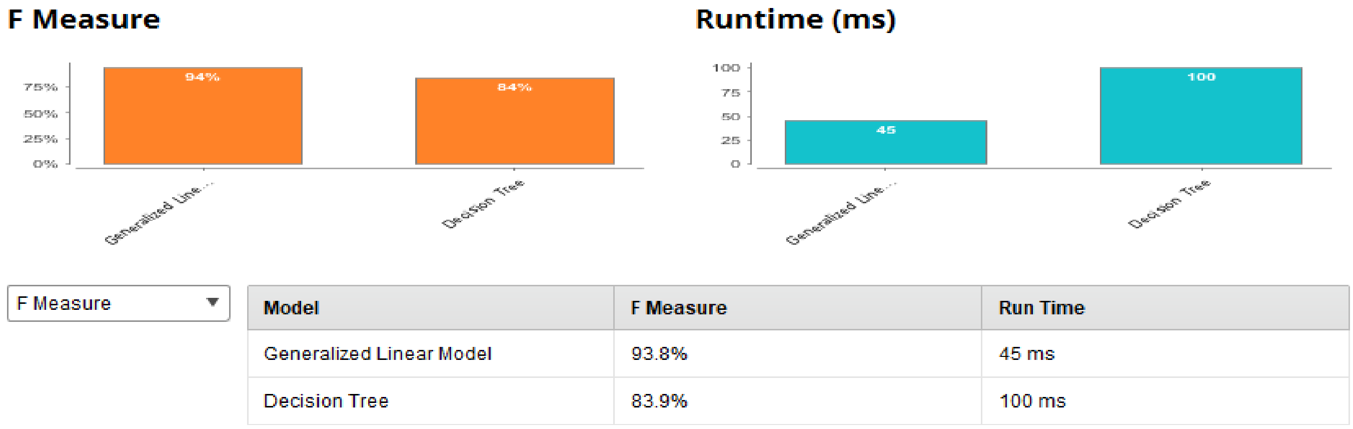


Model	Specificity	Run Time
Generalized Linear Model	88.9%	45 ms
Decision Tree	77.8%	100 ms

Figure 12. Specificity in Linear Regression and Decision Tree

Specificity is important in our analysis to identify the false values to get a better predictive model. Linear Regression Model has better Specificity than Decision Tree as shown in Figure 12. Therefore, Linear Regression Model can better identify the negatives which otherwise will impacts the accuracy of the results.

5.5. F-measure



Model	F Measure	Run Time
Generalized Linear Model	93.8%	45 ms
Decision Tree	83.9%	100 ms

Figure 13. F-Measure in Linear Regression and Decision Tree

F-measure is a harmonic average of precision and recall, where f-measure reaches its best value at 1 and worst at 0.

Therefore, we can conclude that Linear Regression Model is better than Decision Tree in terms of F-measure for this dataset as shown in Figure 13. Both Precision and Recall gives different information that can complement each other when combined. If one of them excels more than the other, F-measure will reflect it.

5.6. Lift

The Lift Chart shows the effectiveness of the model by calculating the ratio between the results obtained with the model and the results obtained without a model. Figure 14 and Figure 15 shows Lift chart of Linear Regression Model and Decision Tree respectively.

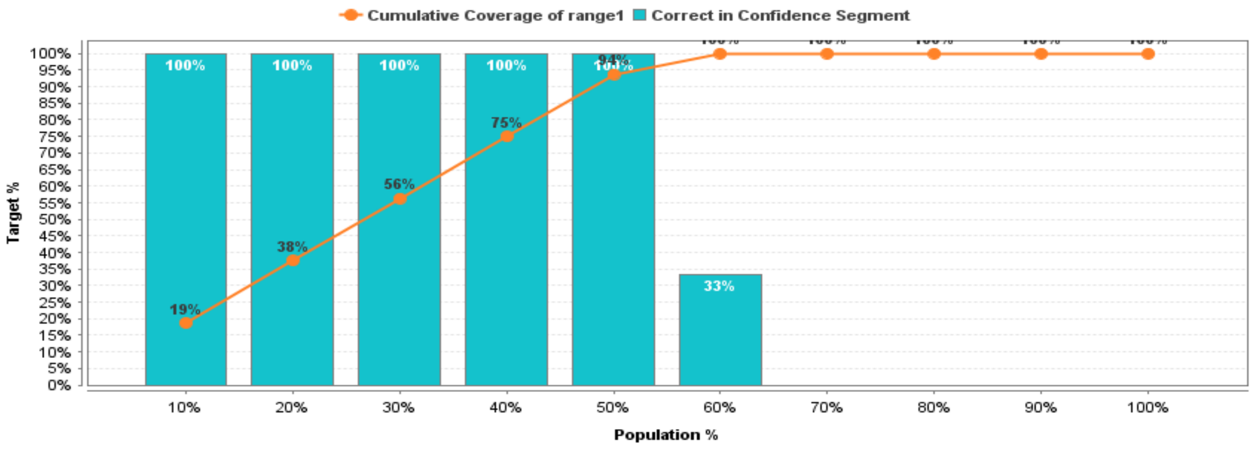


Figure 14. Lift Chart of Linear Regression Model

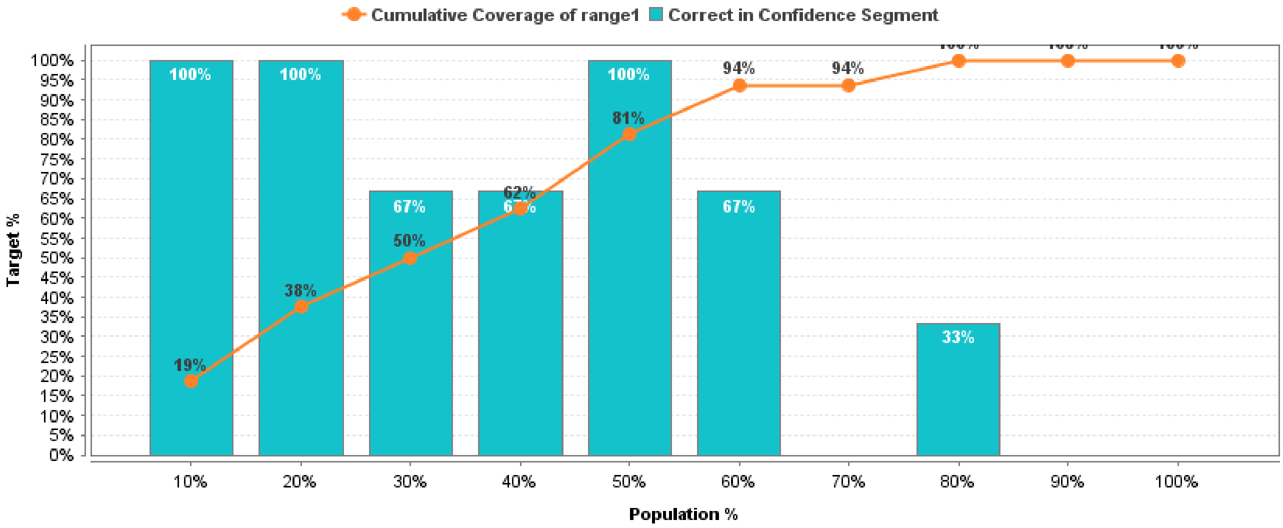
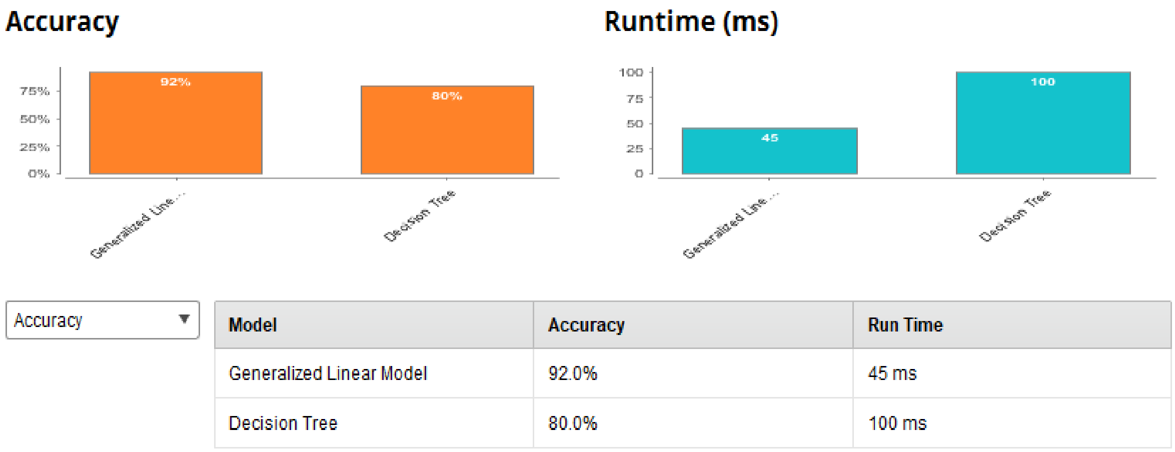


Figure 15. Lift Chart of Decision Tree

5.7. Accuracy



Model	Accuracy	Run Time
Generalized Linear Model	92.0%	45 ms
Decision Tree	80.0%	100 ms

Figure 16. Accuracy in Linear Regression and Decision Tree

Accuracy of Linear Regression Model is much better than Decision Tree. Therefore, for analysis of our dataset, Linear Regression Model is useful and effective than Decision Tree.

6. Conclusion

Based on all the above observation, we can conclude that for analysis of ‘House Pricing Dataset’, Linear Regression Model can get more optimum and accurate results in predicting the best prices based on house dimensions and features. Linear Regression Model gives better precision, recall and specificity than Decision Tree. Also, the error rate in Linear Regression Model is less as compared to Decision Tree. Although, Decision Tree is a good model for the analysis of our data, Linear Regression Model overpowers Decision Tree in all factors. Therefore, Linear Regression Model is accurate, effective and best suited for the analysis of ‘House Pricing Dataset’.

7. References

- Wikipedia
- Data Mining Classification: Basic Concepts, Decision Trees, and Model Evaluation by Tan, Steinbach, Kumar