ISSN 2347 - 3983
Volume 8. No. 4, April 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter70842020.pdf
https://doi.org/10.30534/ijeter/2020/70842020

House Sale Price Prediction using Advanced Regression Techniques and AutoML (TPOTRegressor)

V. S. B. Tejaswi[1], K. V. V. Satyanarayana[2]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Andhra Pradesh, India, siriteja97@gmail.com

[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Andhra Pradesh, India, kopparti@kluniversity.in

ABSTRACT

The increasing and decreasing occurrence of house prices changes from time to time. There are more reasons that effect the fluctuation of house prices. Some are build year, location, physical amenities, size of house etc. Predicting the value of house (Sale price) helps the customers to take right choice of buying the house. Machine learning is being adapted for various fields that could build prediction model and estimate the outcomes. In this paper, we are contemplating the issue of rise and fall of house rates as a regression problem. Regression is a process that aims to predict the correlation between target dependent feature and a sequence of other changing independent features. In our experimental analysis, we are using Decision Tree Regression, Linear Regression, Ada Boost Regression, Gradient Boost Regression, Random Forest Regression techniques. In addition, we are also used AutoML to predict the House sale price. AutoML is a system that takes labelled trained data as input and automatically build a suitable optimized model that the dataset fits.

Key words: House price prediction, Machine learning Decision Tree Regression, Random Forest Regression, Gradient Boost Regression, Ada Boost Regression, AutoML.

1. INTRODUCTION

Day by day, the tendency of people liking to improve their living standards has been increasing. This is heading more of demand for houses. But, the problem is that customers may not know how much worth exactly a house to purchase. It may leads to take wrong decisions. Predicting the price or rate of a house is called house sale price prediction. It helps buyers and sellers. The price of the house varies now and then. The reason behind the changing value of house is based on many features. Some features are location, size of house, bedroom units, and count of storeys, living area, bathrooms, garage size, house age, type of roof and other utilities.

These are considered as independent features in which no feature has relation with other feature. The target feature to predict is Sale price. This is considered as dependent feature in which its value effects in changing the values of independent features.

Machine learning allows the machines to learn and to perform operations by themselves instead of inputting instructions explicitly. The machine learning project life-cycle workflows is as shown in below figure 1.

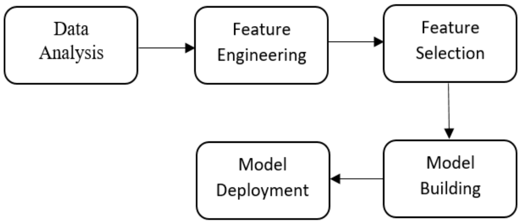


Figure 1: Machine learning project life-cycle

The first phase is Data Analysis. It is the process of visualizing data in means of graphs, finding missing values and observing correlations between features. Second phase is Feature Engineering. It is an activity of converting raw datasets into features which helps in improving the performance of machine learning techniques. Next one is feature selection. It is defined as a way of picking the most important features that effect more to predict the resultant outcome. Model building is defined as using machine learning techniques that enables model to learn from data without giving instructions. Model Deployment is the process of integrating built ML model to dynamic environment to make predictions from historical data.

Prediction in machine learning can be defined as generating an output from dataset input that is applied to a model. The model that best fits to a dataset implies an accurate prediction. We observed and implemented this problem using Regression analysis. Regression is a machine learning technique in which it can be performed whenever we need to model numerous independent features and to predict continuous dependent feature. Since predicting house price that is based on many independent changing features, we implemented regression analysis. We used Decision Tree Regressor, Random Forest Regressor, Linear Regressor, AdaBoost Regressor, Gradient Boost Regressor technique and AutoML.

We organised this paper into five sections. First Section is Introduction. Second section demonstrates Data Pre-processing. Third section explains Algorithms, Forth section describes the model evaluation. Five and six sections are Conclusion and References.

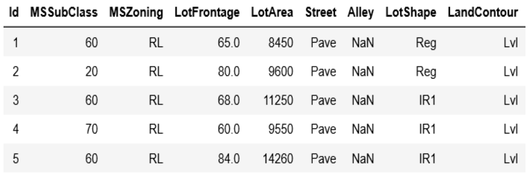
2. RELATED WORK

Sifei Lu et al.[1] developedhybrid Lasso and Gradient boosting regression technique to predict the price of an individual house. Ayush Varma et al.[2] developed a system which predicts accurate house price and also used neural networks. Adyan Nur et al.[3] developed a swarn optimisation along with regression analysis to give optimum outcomes. G. Naga Satish et al.[4] developed a system which facilitates more functionalities to the vendor and used lasso regression. Nehal N Ghosalkar et al.[5] developed linear regression model and contemplated customer requirements and built according to it. D. Banerjee et al.[6] developed a system that utilised classification machine learning techniques. Bruno Klausde et al.[7] developed Recurrent neural networks, random forest and stated that this prediction helps house sellers and buyers.T. D. Phan et al.[8] developed SVM and neural networks and described that can be used for housing market. Sayan Putatunda et al.[9] developed random forests and gradient boosting which considers price as target outcome and all other features are independent features.

3. DATA PRE-PROCESSING

3.1 House Price Dataset

For experimental purpose, we imported House price datasets from https://www.kaggle.com. Dataset sample is as shown in below Figure 2.

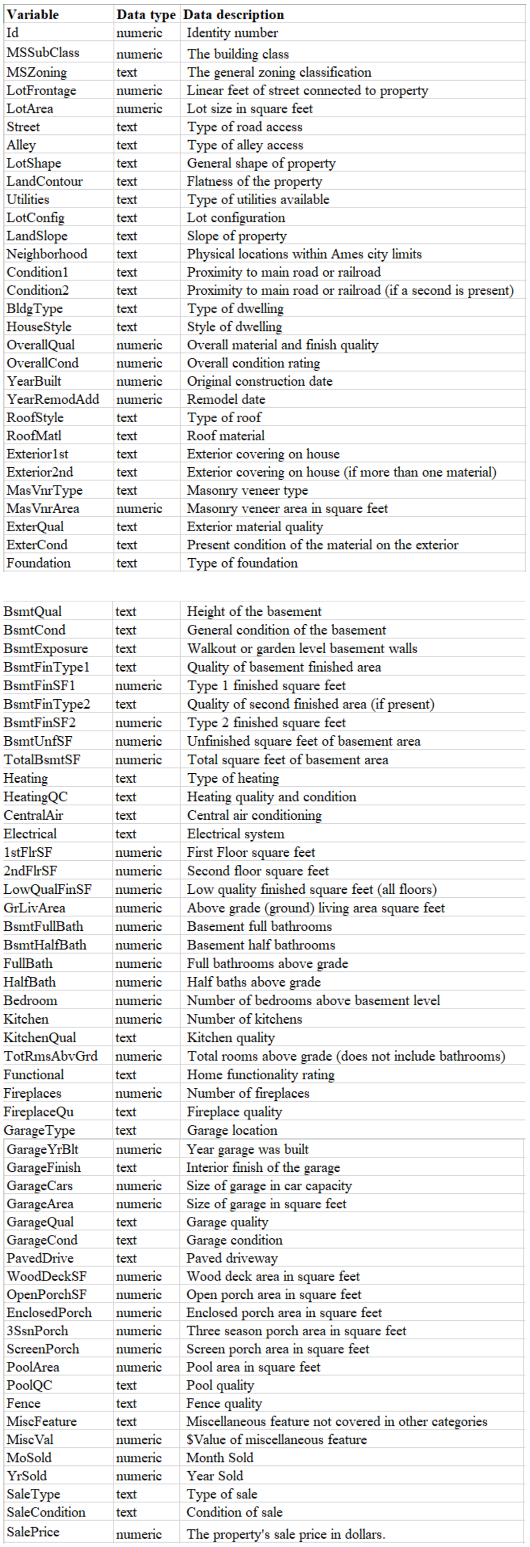


Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl
4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl
5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl

Figure 2: Dataset sample

3.2 Dataset Description

The description of dataset is representing in below Figure 3.



Variable	Data type	Data description
Id	numeric	Identity number
MSSubClass	numeric	The building class
MSZoning	text	The general zoning classification
LotFrontage	numeric	Linear feet of street connected to property
LotArea	numeric	Lot size in square feet
Street	text	Type of road access
Alley	text	Type of alley access
LotShape	text	General shape of property
LandContour	text	Flatness of the property
Utilities	text	Type of utilities available
LotConfig	text	Lot configuration
LandSlope	text	Slope of property
Neighborhood	text	Physical locations within Ames city limits
Condition1	text	Proximity to main road or railroad
Condition2	text	Proximity to main road or railroad (if a second is present)
BldgType	text	Type of dwelling
HouseStyle	text	Style of dwelling
OverallQual	numeric	Overall material and finish quality
OverallCond	numeric	Overall condition rating
YearBuilt	numeric	Original construction date
YearRemodAdd	numeric	Remodel date
RoofStyle	text	Type of roof
RoofMatl	text	Roof material
Exterior1st	text	Exterior covering on house
Exterior2nd	text	Exterior covering on house (if more than one material)
MasVnrType	text	Masonry veneer type
MasVnrArea	numeric	Masonry veneer area in square feet
ExterQual	text	Exterior material quality
ExterCond	text	Present condition of the material on the exterior
Foundation	text	Type of foundation
BsmtQual	text	Height of the basement
BsmtCond	text	General condition of the basement
BsmtExposure	text	Walkout or garden level basement walls
BsmtFinType1	text	Quality of basement finished area
BsmtFinSF1	numeric	Type 1 finished square feet
BsmtFinType2	text	Quality of second finished area (if present)
BsmtFinSF2	numeric	Type 2 finished square feet
BsmtUnfSF	numeric	Unfinished square feet of basement area
TotalBsmtSF	numeric	Total square feet of basement area
Heating	text	Type of heating
HeatingQC	text	Heating quality and condition
CentralAir	text	Central air conditioning
Electrical	text	Electrical system
1stFlrSF	numeric	First Floor square feet
2ndFlrSF	numeric	Second floor square feet
LowQualFinSF	numeric	Low quality finished square feet (all floors)
GrLivArea	numeric	Above grade (ground) living area square feet
BsmtFullBath	numeric	Basement full bathrooms
BsmtHalfBath	numeric	Basement half bathrooms
FullBath	numeric	Full bathrooms above grade
HalfBath	numeric	Half baths above grade
Bedroom	numeric	Number of bedrooms above basement level
Kitchen	numeric	Number of kitchens
KitchenQual	text	Kitchen quality
TotRmsAbvGrd	numeric	Total rooms above grade (does not include bathrooms)
Functional	text	Home functionality rating
Fireplaces	numeric	Number of fireplaces
FireplaceQu	text	Fireplace quality
GarageType	text	Garage location
GarageYrBlt	numeric	Year garage was built
GarageFinish	text	Interior finish of the garage
GarageCars	numeric	Size of garage in car capacity
GarageArea	numeric	Size of garage in square feet
GarageQual	text	Garage quality
GarageCond	text	Garage condition
PavedDrive	text	Paved driveway
WoodDeckSF	numeric	Wood deck area in square feet
OpenPorchSF	numeric	Open porch area in square feet
EnclosedPorch	numeric	Enclosed porch area in square feet
3SsnPorch	numeric	Three season porch area in square feet
ScreenPorch	numeric	Screen porch area in square feet
PoolArea	numeric	Pool area in square feet
PoolQC	text	Pool quality
Fence	text	Fence quality
MiscFeature	text	Miscellaneous feature not covered in other categories
MiscVal	numeric	$Value of miscellaneous feature
MoSold	numeric	Month Sold
YrSold	numeric	Year Sold
SaleType	text	Type of sale
SaleCondition	text	Condition of sale
SalePrice	numeric	The property's sale price in dollars.

Figure 3: Data Description

3.3 Missing values

There may be missing values in the dataset. Missing values in data may interpret inaccurate predictions. Considering missing values, we need to find the relation between missing values and dependent target feature. The objective of finding this relationship is to determine how much an individual feature's missing value is effecting the dependent feature.

These are some of the visualizations (bar graphs) that shows relationship between individual missing value features and SalesPrice in below figures 4 – 8.

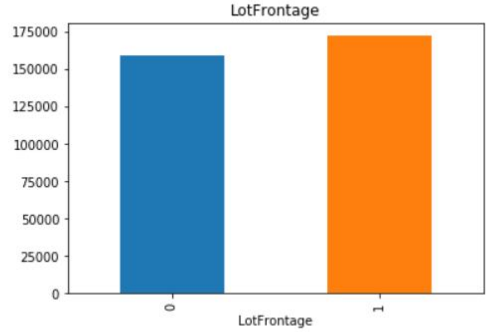


Figure 4: Relationship between LotFrontage's missing values and SalesPrice

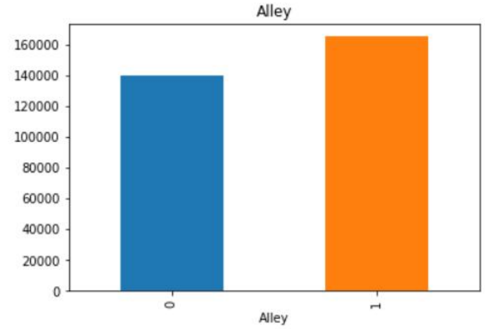


Figure 5: Relationship between Alley's missing values and SalesPrice

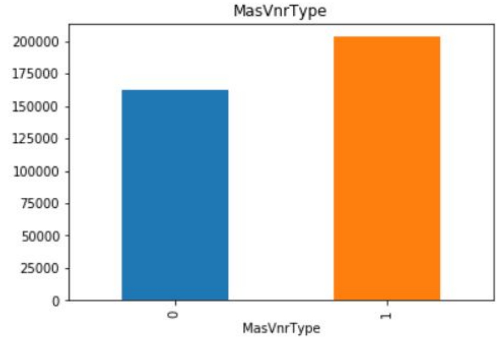


Figure 6: Relationship between MasVnrType's missing values and SalesPrice

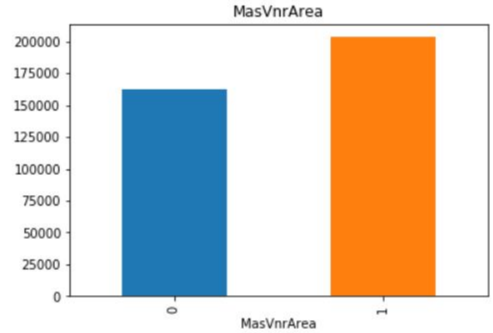


Figure 7: Relationship between MasVnrArea's missing values and SalesPrice

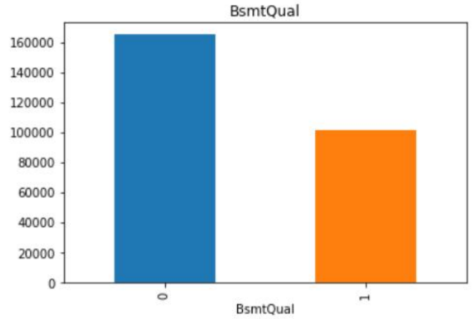


Figure 8: Relationship between BsmtQual's missing values and SalesPrice

In these graphs, '1' represents missing (Nan) values and '0' represents un-missing values. X-axis represents independent feature and Y-axis represents SalesPrice. An efficient way to solve the issue of numeric missing values is to replace them with Mean/Median/Mode value. Another ways are replacing with zero/constant and repeated value within each column.

3.4 Logarithmic transformation

Logarithmic transformation is a technique which deals skewed data and reduce the data variability. The below scatter plot graphs figures 9 – 13 presents logarithmic transformation of some continuous features.

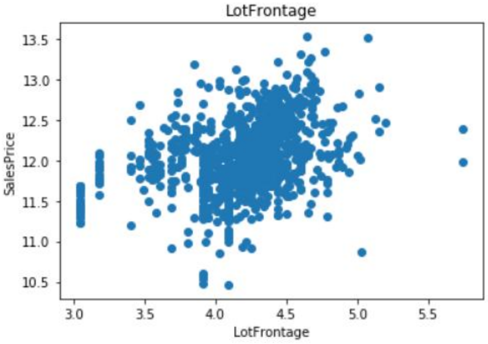


Figure 9: LotFrontage log transform

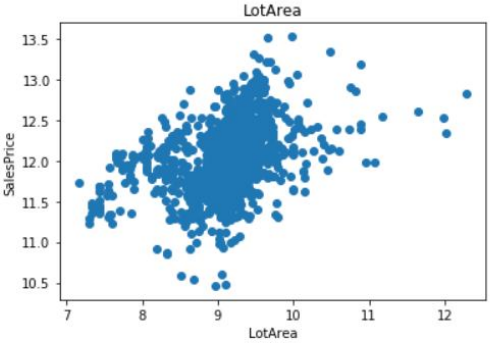


Figure 10: LotArea log transform

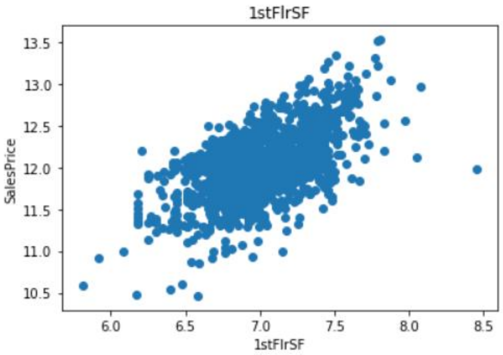


Figure 11:1stFlrSF log transform

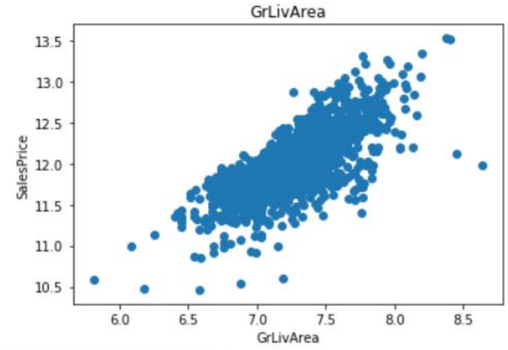


Figure 12: GrLivArea log transform

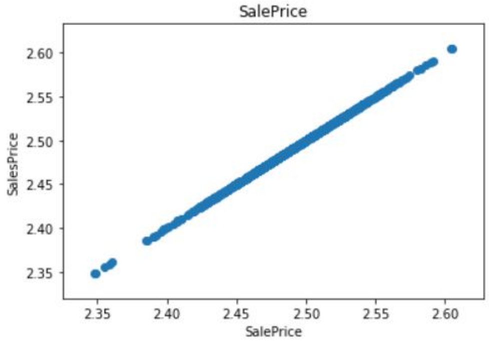


Figure 13: alePrice log transform

3.5 Outliers

Outliers are the extreme observations of data that deviates from rest of the data. The below boxplots figures 14 – 18 are some of the observed outliers of continuous features.

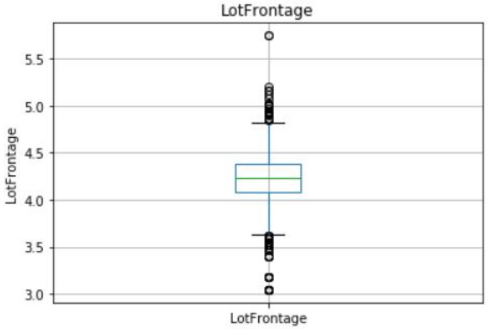


Figure 14: LotFrontage Outliers

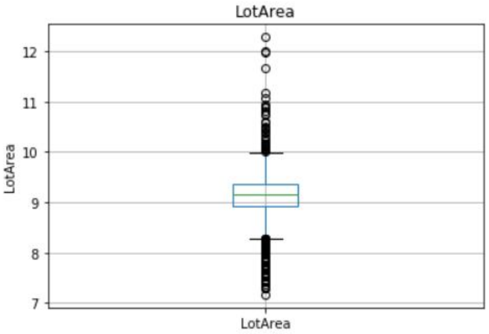


Figure 15: LotArea Outliers

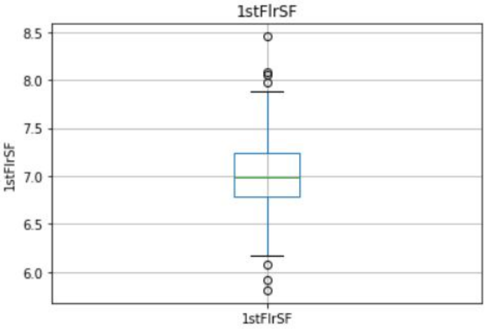


Figure 16:1stFlrSF Outliers

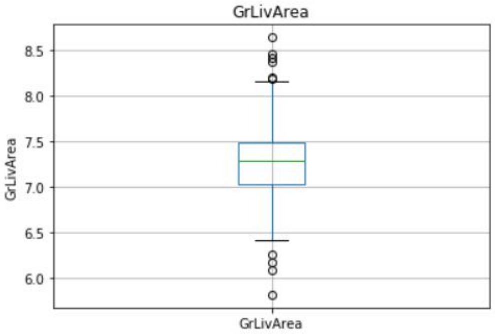


Figure 17: GrLivArea Outliers

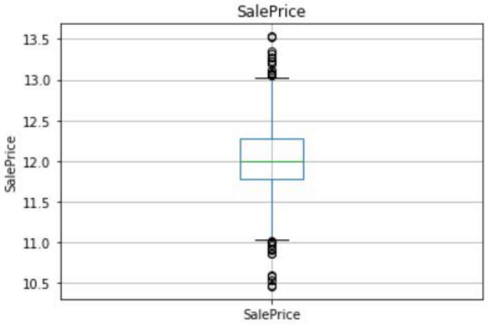


Figure 18: SalePrice Outliers

3.6 Correlation

Correlation is a quantify that shows how much a feature is dependent on another feature. The below heatmap figure 19 is showing how strongly one feature is related to other feature.

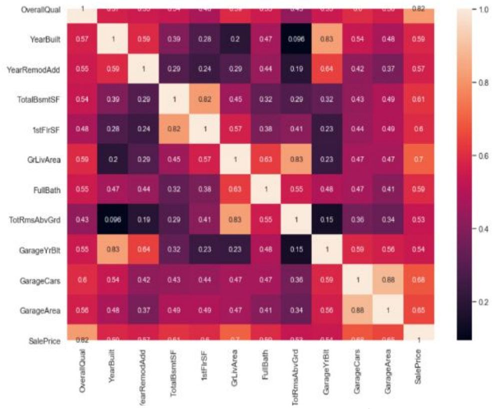


Figure 19: Correlation between features

4. ALGORITHMS

Before executing model, we consider the data into 'x' and 'y'. x consists of input of independent features. y consists of output of dependent feature. Here, y is SalePrice which is targeted feature. We split this data into training data and testing data i.e. x_train, y_train, x_test, y_test. To train the model, training data will be utilised. After we make the model to learn with train data, we give testing data (x_test) as input to the model. It estimates the outcome as y_test. The performance of a model can be measured in accuracy score. The accuracy score is difference between predict value score and real value score. 100% accuracy of a model leads to overfitting. Very least accuracy leads to underfitting.

4.1 Linear Regression Technique

Linear regression is one of the supervised machine learning methodology in which it performs on labelled data and predicts dependable feature from historical undependable features. It determines the linear relation of depend variable with independent variables.

4.2 Random Forest Regression Technique

Random forest regression is a supervised machine learning algorithm in which trees are executed parallel and generates mean prediction of each trees. Firstly it takes the sample input data and build number of decision trees. These decision trees run in parallel manner and estimate predictions. Considering all the predictions of trees, we calculate average for final result. Figure 20 shows random forest regression.

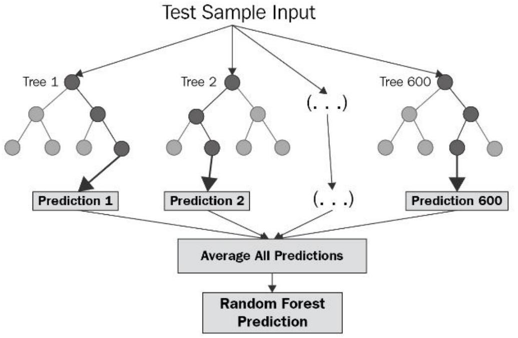


Figure 20: Random Forest regression workflow

4.3 Gradient Boosting Regression

Gradient boosting regression technique enables to construct a strong model upon weak regression model. Initially, it predicts using a model. Then it builds another model by solving the issues of first model. This process continues either whole dataset completes or until most of the model utilised. In our experimental analysis, it is proven that gradient boosting regressor predicted with high accuracy among other models.

4.4 Ada Boost Regression

Ada Boost Regression is a machine learning regression technique in which the predictions are estimated by taking mean of weak regressors. It is quick to detect the small changes of outliers and noisy data. Ada boost is alias name of Adaptive boosting.

4.5 Decision Tree Regression

Decision tree is one of the machine learning techniques in which dataset splits into small subsets and evaluate. Both continuous and categorical outcome features can be worked by decision tree. Decision tree regression considers some features as an input and trains the model to estimate the future outcome feature. Decision trees has an ability to observe the non-linear correlation of independent attributes with depend target attribute. Decision tree looks like following below figure 21.

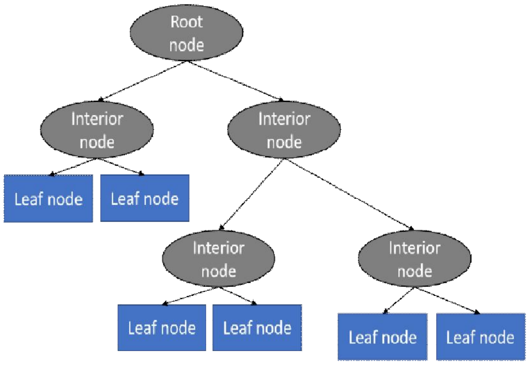


Figure 21: Decision Tree

4.6 TPOT Regression (AutoML)

Automated machine learning (AutoML) is a way of automatizing the computer to choose the model based on given dataset. So, it can give assistance to data scientists to pull the appropriate model. TPOT is an AutoML tool and the acronym is Tree-Based Optimization Tool. It provides open access to utilise packages. The procedure is same as other machine learning models. But the difference is, it automates the process of feature selection, feature pre-processing, feature construction, selection of model and parameter optimisation. In our experimental procedure, TPOT regressor chosen XGBRegressor as a model automatically based on given input dataset. Figure 22 shows TPOT regressor flow of work.

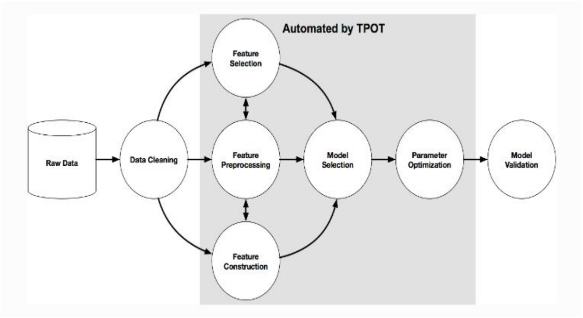
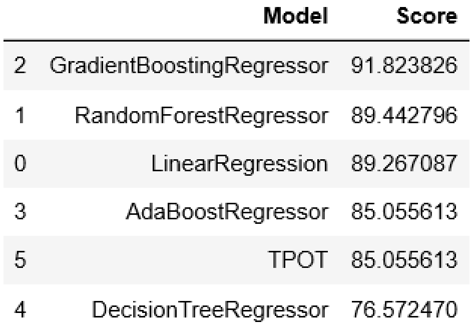


Figure 22:TPOT Regressor

5. MODEL EVALUATION

The accuracy score of all models are represented in the following below figure 23. Gradient boosting regressor model gave high accuracy among remaining models which is 91.8%.

Random forest regressor and regressor got 89% accuracy. And the least with 76% accuracy score is Decision tree regression model. TPOT and Adaboost regressor estimated with the accuracy of 85%.



	Model	Score
2	GradientBoostingRegressor	91.823826
1	RandomForestRegressor	89.442796
0	LinearRegression	89.267087
3	AdaBoostRegressor	85.055613
5	TPOT	85.055613
4	DecisionTreeRegressor	76.572470

Figure 23: Result Analysis

The accuracy score is also plotted in the below bar graph figure 24

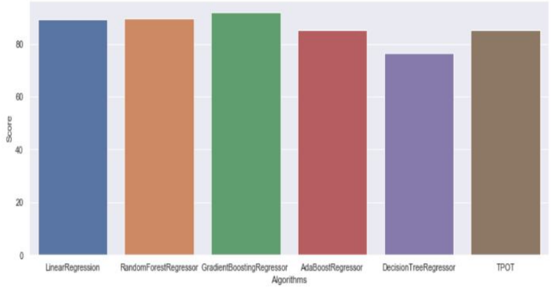


Figure 24: Results in bar graph

6. CONCLUSION

Prediction of house price is done using Regressor for house price dataset, comparing with different regression techniques like linear regression technique, decision tree regression algorithm, random forest regression technique, gradient boost regression, ada boost regression,. In further, TPOT regression (AutoML) is also performed. TPOT is an automatic tool that picks the best fit algorithm. The highest accuracy of 91% is from gradient boost regression. The customers and sellers will be benefitted with house price prediction.

REFERENCES

1. Sifei Lu, Zengxiang Li, Zheng Qin, Xulei Yang, Rick Siow Mong Goh**, A Hybrid Regression Technique for House Prices Prediction**, Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A*STAR), Singapore.
2. Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, **House Price Prediction Using Machine Learning And Neural Networks.**
3. Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, **Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization**.
4. G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu, **House Price Prediction Using Machine Learning.**
5. Nehal N Ghosalkar, Sudhir N Dhage, **Real Estate Value Prediction Using Linear Regression.**
6. D. Banerjee and S. Dutta, **Predicting the housing price direction using machine learning techniques**, 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, 2017, pp. 2998-3000. https://doi.org/10.1109/ICPCSI.2017.8392275
7. Bruno Klausde Aquino Afonso1, Luckeciano Carvalho Melo2, Willian Dihanster Gomesde Oliveira1, Samuel Brunoda Silva Sousa1, Lilian Berton1,**Housing Prices Prediction with a Deep Learning and Random Forest Ensemble**, Institute of Science and Technology – Federal University of S˜ao Paulo (UNIFESP)
8. T. D. Phan, **Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia**, 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, Australia, 2018, pp. 35-42. https://doi.org/10.1109/iCMLDE.2018.00017
9. Sayan Putatunda, **PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market**