Proceedings of the Fifth International Conference on Trends in Electronics and Informatics (ICOEI).
IEEE Xplore Part Number:CFP21J32-ART; ISBN:978-1-6654-1571-2

Prediction of House Price Using Machine Learning Algorithms

[1]G Kiran Kumar
Dept. of Information Technology
MLR Institute of Technology,
Hyderabad, Telangana, India
ganipalli.kiran@gmail.com

[2]D Malathi Rani
Dept. of Electronics and Communications Engineering
Marri Laxman Reddy Institute of Technology and Management
Hyderabad, Telangana, India
duggi.malathi@gmail.com

[3]Neeraja Koppula
Dept. of Information Technology
MLR Institute of Technology,
Hyderabad, Telangana, India
kneeraja123@gmail.com

[4]Syed Ashraf
Dept. of Information Technology
MLR Institute of Technology
Hyderabad, Telangana, India
syedashraf612@gmail.com

2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) | 978-1-6654-1571-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOEI51242.2021.9452820

***Abstract*— People are very careful when they want to buy a new house with market strategies and their budgets. The objective of this paper is to predict the house prices for non-house holders based on their aspirations and financial provisions. By analyzing different parameters like area of the house, square feet of the house, no of floors in the house etc. This research work has utilized the dataset from Kaggle.. An analysis is performed by applying advanced machine learning regression techniques such as Linear regression, KNN Regression, Random Forest Regression, Decision Tree Regression, Extra Trees Regression etc. to attain the most efficient and least error driven regression technique. From the analysis performed, an observation has been made that Catboost Regression Algorithm has outperformed other algorithms. The model predicts the final output with respect to correlated attributes in the dataset.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Machine learning is a subset of artificial intelligence. The machine is learning by itself and testing through the existing dataset using certain algorithms and gives more accurate results than a manual work. Machine learning algorithms uses existing data as input to predict new output values for new input values [1] . Machine Learning is majorly bisected into two sections, namely supervised learning and unsupervised learning. Supervised learning is where the program is trained on a predetermined set of data and to be able to predict when a new data is given. Unsupervised learning is where the program tries to find the relationship and the hidden pattern between the data.

There are numerous machine learning algorithms in recent times that can be implemented and applied on to predict the desired output, but each and every algorithm works differently for various kinds of datasets. We are going to measure the performance of the algorithms upon house pricing prediction datasets with respect to the Root Mean Square Error (RMSE) and implement the best performing algorithm in the final model[2].

The dataset used in the analysis phase will undergo preprocessing which will help to improve the data redundancy and missing data, this will improve training and the accuracy of output prediction[6][7].

II .LITERATURE SURVEY

The presently existing process for buying or selling a house is through a broker, who acts as a third party between the buyer and the seller and someone who takes commission from both sides for the deal that he gets for his customers. This has been the part of real estate for a long time now and effects the house prices as the third party looks to make profit out the deal, which isn't fair to both the buyer and the seller as the buyer has to spend extra money to the third party for the deal and the seller has to share some part of what they earn[4][5]. Mousavi et al proposed a combination of cross validation, swarm optimization and support vector regression is proposed in predicting the cost of new product development [8][9]. Zhigang Jiand et al proposed a data driven based

decomposition and integration method to predict EOL products re-manufacturing cost [10].

In another work, prediction of cost to pay to independent drivers is done using Ventral hernia repair cost variability concept. It is very helpful in understanding the needs of individual drivers, controlling the cost [11][12].

Neeraja Koppula et al explained about machine learning algorithm which is based on purely graph is discussed and is used in machine translation and disambiguation [13][14].

III. PROPOSED WORK

We have designed a system which overcomes the traditional method of going to a broker to buy a house or sell a house. Using the proposed system, automated price predictor any person can get their queries resolved and would not need to consult any third party to get their house approximate prediction as the third party may keep their part of the commission from the deal and could also ask for the service charges to both sides of the deal.

Hence, using this system, real estate standards can change and improve to new heights while being very helpful to common people who are looking to buy or sell a house without being corrupted by the real estate brokers and saves people from being conned.

The software can be deployed as a website or as a mobile application as per the convenience of the user and used by anyone and everyone via the respectful applications.

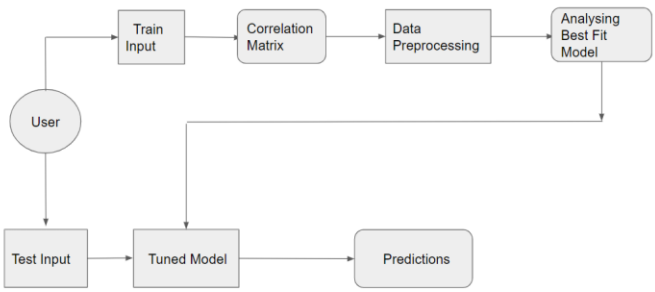

Fig. 1 - The Figure represents the system design of the model

The diagram in figure represents the flow of functioning of data in the model starting from the user on the left side and ends at the prediction point after being passed through various stages in the design cycle.

- The user enters the input dataset in csv format to the model
- Dataset is then passed to the correlation matrix[3] where the number of attributes are reduced as per the correlation to the labeled value for individual instances.
- The new reformed dataset is preprocessed to remove outliers and missing values which could hinder the flow of program
- The preprocessed data is used to analyze various regression algorithms using pycaret[3] to achieve the best performing algorithm
- The resultant regression technique is tuned to build the final model, which will be used to predict the user cases
- The user then gives a test input to the tuned model to make predictions.

IV. EXPERIMENTS

The experiments are performed on a real time database based in the Ames, Iowa. The dataset contains 1460 rows and 81 columns where 80 columns represent different attributes that act as a parameter for the experiment.

A. Performance Measure

We have evaluated the proposed model's performance measure using Root Mean Square Error (RMSE). RMSE is the standard deviation of the errors which occurs when a prediction is made on a dataset.[7]

$$RHMS = \sqrt{\frac{\sum_{i=1}^{N} (Predict_i - Actual_i)^2}{N}}$$

(1) Formula used to calculate the RMSE in the model.

We have chosen RMSE as the performance measure because it designates a higher weight to errors that are larger which indicates that RMSE is more effective when large errors are found, and they drastically affect the model's performance. Since our dataset contains huge values there is a possibility of larger errors. In this metric also, the performance of the model is better when the value is as low as possible. Hence, RMSE is best suited to measure this model's performance.

B. Selection Parameter

The results from the experiment are dependent on various attributes. The proposed experiment is heavily dependent on attributes such as OverAllQuality, GroundLivingArea, GarageCars, GarageArea and TotalBasementSurface have been considered. These considerations are considered based on the correlation matrix implementation of each attribute with respect to the labeled requirement which is the sales price.

V. RESULTS

```
compare_models(sort='RMSE')
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	1.815e+04	6.913e+08	2.604e+04	0.8283	0	0.1128	6.211
gbr	Gradient Boosting Regressor	1.886e+04	7.31e+08	2.684e+04	0.8178	0	0.1181	0.292
et	Extra Trees Regressor	1.888e+04	7.633e+08	2.742e+04	0.8118	0	0.1165	0.363
lightgbm	Light Gradient Boosting Machine	1.943e+04	7.7e+08	2.756e+04	0.8074	0	0.121	0.278
rf	Random Forest Regressor	1.924e+04	7.839e+08	2.784e+04	0.8063	0	0.1194	0.582
br	Bayesian Ridge	2.006e+04	7.898e+08	2.788e+04	0.8073	0	0.1246	0.053
ridge	Ridge Regression	2.012e+04	7.906e+08	2.789e+04	0.807	0	0.1251	0.104
llar	Lasso Least Angle Regression	2.012e+04	7.909e+08	2.79e+04	0.8069	0	0.1251	0.018
lasso	Lasso Regression	2.013e+04	7.91e+08	2.79e+04	0.8069	0	0.1252	0.108
lr	Linear Regression	2.013e+04	7.91e+08	2.79e+04	0.8069	0	0.1252	1.041
lar	Least Angle Regression	2.013e+04	7.91e+08	2.79e+04	0.8069	0	0.1252	0.078
huber	Huber Regressor	1.96e+04	8.104e+08	2.821e+04	0.8028	0	0.1193	0.045
par	Passive Aggressive Regressor	1.956e+04	8.281e+08	2.852e+04	0.7988	0	0.1188	0.071
xgboost	Extreme Gradient Boosting	1.986e+04	8.313e+08	2.864e+04	0.7949	0	0.1228	1.012
knn	K Neighbors Regressor	2.026e+04	8.78e+08	2.931e+04	0.7856	0	0.1226	0.038
ada	AdaBoost Regressor	2.292e+04	9.508e+08	3.065e+04	0.7624	0	0.1456	0.13
en	Elastic Net	2.157e+04	9.788e+08	3.109e+04	0.7615	0	0.131	0.019
omp	Orthogonal Matching Pursuit	2.41e+04	1.11e+09	3.314e+04	0.7272	0	0.1456	0.023
dt	Decision Tree Regressor	2.672e+04	1.452e+09	3.798e+04	0.6363	0	0.1656	0.025

```
<catboost.core.CatBoostRegressor at 0x1c7c07ea390>
```

Fig. 2 - The above figure represents the comparison between applied algorithms

The above figure is used to depict the comparison of algorithms based on the performance measure of the model to predict the most suitable and least error prone output.

The model is trained on the Catboost Regression Algorithm with respect to the results in Fig. 2. Catboost is an algorithm that works extremely well with text, images and historical data which is the dataset for our project. Catboost gives a state-of-the-art result without the need to be trained extensively unlike many other machine learning algorithms used in the present day.

Catboost is a gradient boosting algorithm in machine learning which is an advanced form of gradient boost algorithm which performs effectively in predicting, forecasting values when provided with historical data. The model maintains its effectiveness even if not a lot of data is given for training unlike many Deep Learning models that require extensive data training.

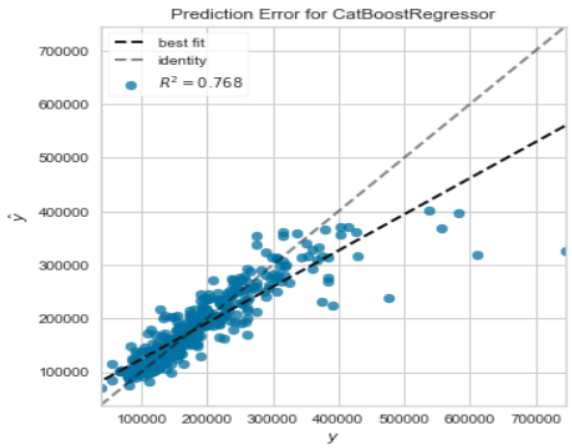

Fig. 3 - The diagram represents the error prediction fit of the tuned model

The figure represents a graph representing the deviation of the tuned model fit against the identity or ideal fit for the dataset which represents the error in prediction. The performance measure R Square is calculated to be 0.768 for the tuned model fit.

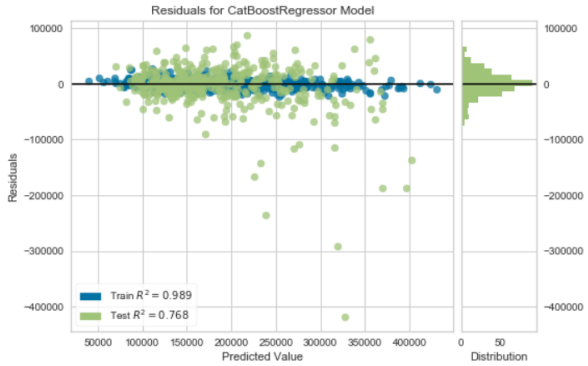

Fig. 4- The figure represents the residual graph denoting the predicted outputs from training dataset

The figure represents the residual graph which depicts the deviation of observed output value to the predicted output value by the model. The R Square values in the graph defines the accuracy of the training value and the predicted value respectively.

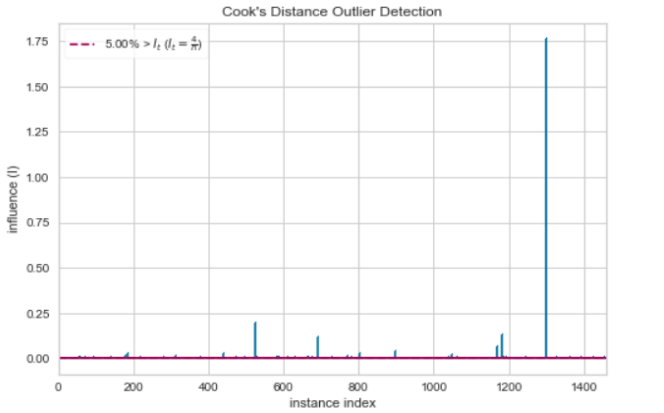

Fig. 5 - The figure represents the outliers present through the instances in the dataset

The figure represents a graph denoting the Cook's Distance Outlier Detection[5]. Cook's Distance is a commonly used estimate or measure of the influence of a data point when performing a least-squares regression analysis.

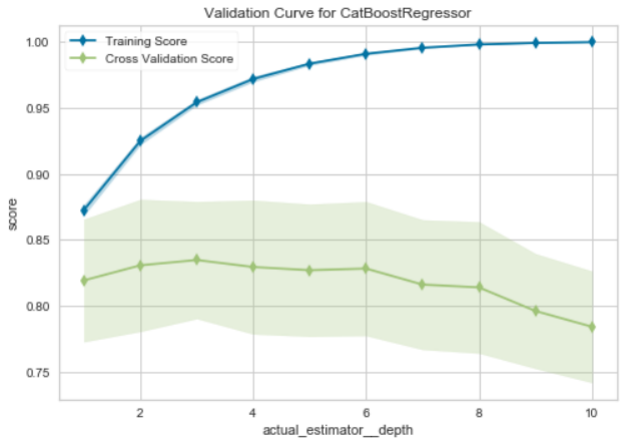

Fig. 6 - The figure denotes the validation for the training score

The figure represents the validation curve of the tuned model. Validation cure is the learning curve calculated from a hold-out validation dataset that gives an idea of how well the model is generalizing.
dataset.

VI. CONCLUSION

The dataset was taken from Kaggle having house prices for a state in the US. The analysis is performed on various regression algorithms using PyCaret. As observed from the analysis, catboost regression technique has stood out amongst all other algorithms, based on parameters some of which are OverAllQuality, GroundLivingArea, GarageCars, GarageArea, YearBuilt.. The model allows a person to easily find a price for their house or for their-to-be house without needing for them to consult a real estate broker or any third party for the same price calculation and also saving a lot of money that any third party would take for providing their service to the customer.

ACKNOWLEDGMENT

I would like to acknowledge Kaggle for the resources and requirements I have consumed through the source.

.

REFERENCES

[1]. Chiramel, Sruthi, et al. "Efficient Approaches for House Pricing Prediction by Using Hybrid Machine Learning Algorithms." *Asian Conference on Intelligent Information and Database Systems*. Springer, Singapore, 2020.
[2]. Jain, Mansi, et al. "Prediction of House Pricing Using Machine Learning with Python." *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020.
[3]. Steiger, James H. "Tests for comparing elements of a correlation matrix." *Psychological bulletin* 87.2 (1980): 245.
[4]. Dhakal, Chuda Prasad. "Dealing with outliers and influential points while fitting regression." *Journal of Institute of Science and Technology* 22.1 (2017): 61-65.
[5]. Dharavath, Ramesh, et al. "t-SNE Manifold Learning Based Visualization: A Human Activity Recognition Approach." *Advances in Data Science and Management*. Springer, Singapore, 2020. 33-43.
[6]. Pumsirirat, Apapan, and Liu Yan. "Credit card fraud detection using deep learning based on auto-encoders and restricted boltzmann machines." *International Journal of advanced computer science and applications* 9.1 (2018): 18-25.
[7]. Reddy, G. Thippa, et al. "Analysis of dimensionality reduction techniques on big data." *IEEE Access* 8 (2020): 54776-54788.
[8]. Alfiyatin, Adyan Nur, et al. "Modeling house price prediction using regression analysis and particle swarm optimization." *International Journal of Advanced Computer Science and Applications* 8 (2017).
[9]. Mousavi, S. Meysam, Behnam Vahdani, and Majid Abdollahzade. "An intelligent model for cost prediction in new product development projects." *Journal of Intelligent & Fuzzy Systems* 29.5 (2015): 2047-2057.
[10]. Jiang, Zhigang, et al. "A data-driven based decomposition–integration method for remanufacturing cost prediction of end-of-life products." *Robotics and Computer-Integrated Manufacturing* 61 (2020): 101838.
[11]. Nisiewicz, Michael J., et al. "Validation and extension of the ventral hernia repair cost prediction model." *Journal of Surgical Research* 244 (2019): 153-159.
[12]Kumar, G. Kiran, and D. Malathi Rani. "Paragraph summarization based on word frequency using NLP techniques." AIP Conference Proceedings. Vol. 2317. No. 1. AIP Publishing LLC, 2021.

[13] Neeraja Koppula, and Dr. Padmaja Rani, "Hybrid Approaches for Word Sense Disambiguation : A Survey" in International Journal of Applied Engineering Research, 2015, Volume-10, Issue-23, PP. 43891-43895. ISSN:0973-4562.

[14] Koppula, Neeraja, B. Padmaja Rani, and Koppula Srinivas Rao. "Graph based word sense disambiguation." Proceedings of the first international conference on computational intelligence and informatics. Springer, Singapore, 2017.