2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)
University of Malaya, Kuala Lumpur, Malaysia. Sep 24-26, 2021

2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) | 978-1-7281-9951-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/GUCON50781.2021.9573829

Real Estate Price Prediction using Data Mining Techniques

Sayar Kumar Dey
Department of Computer Science
Birla Institute of Technology and Science, Pilani,
Dubai International Academic City,
Dubai, United Arab Emirates
h20200013@dubai.bits-pilani.ac.in

Siddhaling Urolagin
Department of Computer Science
APP Centre for AI Research (APPCAIR),
Birla Institute of Technology and Science, Pilani,
Dubai International Academic City,
Dubai, United Arab Emirates
siddhaling@dubai.bits-pilani.ac.in

***Abstract*—The objective of this paper is to create a model for data mining using knowledge of real estate to predict property prices. It uses data set that have many dimensions to train and build a model. This is done in mainly two phases: data pre-processing and model building. In the first phase data is cleansed and normalized. Abnormal values are removed to make the data standardized. The second phase is mainly concerned with building various types of models to train and see which model performs the best and thus selecting the best model to use in the prediction of price for real estate properties. The models that have been used are elastic net, kernel ridge, lasso, random forest, SVM, XGBoost, LGBM and gradient boosting. Out of all these elastic net gave the highest accuracy score of 94%. Cross validation of models is done in phase two as well as methods to improve accuracy is discussed.**

***Index Terms*—Data Mining, Prediction, Real Estate,Regression**

I. Introduction

Real Estate is an evergreen market. It generates a lot of data that can be mined to gain meaningful insights. Data mining has a lot of applications in this industry. It can make it so that there is no need to depend on middle man or agents. It makes decision making easy. It can help us make more economic choices. It will help in fraud detection too.

Data Mining is a vast field that helps in deriving meaning out of data. It uses many mathematical tools such as statistics and probability to create models that help in predicting an outcome for a given set of inputs. In early days, real estate decisions were based on experience and emotions. Any experienced real estate dealer would influence the market prices for properties based on his manual calculation. The process of doing so is not only tedious but also prone to error. There is always a chance of bias wherever emotions are involved. Thus we can notice abnormal pricing in many areas of real estate. The reason data mining is so useful in determining prices in real estate is mainly because of the volume of data that is generated in these industries. In comparison to past where manual interpretation data was done now its much easier as data mining specialized in mining insights from huge amount of data which otherwise would have been impossible. This paper is organized as follows. Section II explains the related work done in data warehousing for telecommunication industries. Section III describes the methodos used in the proposed system. Section IV describes the overall setup of the proposed system. Section V is about the results that this system achieves. Conclusions and recommendations for future work follow in section VI.

II. Literature Survey

In this section a brief study of some of the papers that have done related study in this field is presented. [17] uses data warehousing and regression techniques to find the best model that can be used to predict the price of property listings. [6] uses data mining software to evaluate real estate prices. [11] uses predictive modeling for credit card fraud detection. It also uses predictive modeling and tries to predict the correct result. [8] tries to identify hear diseases using decision tree technique. [15] tries to predict stock prices using hybrid neural network. [1] tries to analyze customer behavior using clustering and classification techniques. [9] tries to identify similar attributes between datasets using advance data mining techniques. [3] tries to mine data from educational applications to draw insights in how data has been usen in such applications. [16] tries to predict churn using hybrid data mining systems. [14] tries to use data mining for medical diagnosis [7] tries to detect defects using neural networks. [2] works on using data mining to detect covid -19 in animals. It is a a very interesting application of data mining. [4] tries to explain the various methods used in data mining. [13] applies data mining to education so that ineternet learning can be improved. [10] tries to optimize data learning algorithms based on ant colony logic. [12] tries to analyze data records of vaccinated and infected patients of covid-19. [5] is a study of data mining on healt care systems.

III. Methodology

The entire process is carried out in two phases In first phase, data pre-processing is done. Then the data is gathered. It is followed by data exploration and analysis. Data cleaning is then carried out. Outliers detection is done on the result. Null values are then evaluated. If it is required then new features are added. The data is then prepared for learning purposes. If the data is skewed then it is evaluated accordingly. After all these

978-1-7281-9951-1/21/$31.00 ©2021 IEEE

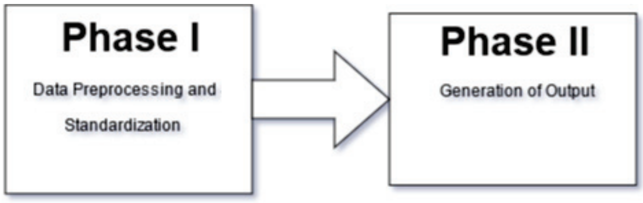


Fig. 1. Architecture of Price Prediction Model

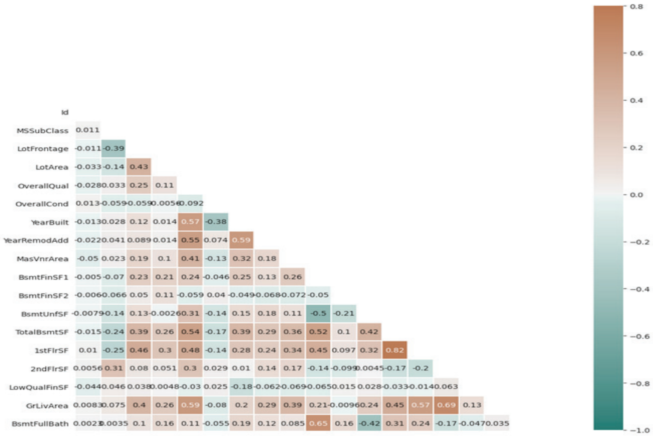


Fig. 2. Correlation Graph

steps phase two starts. In phase two the models are initialized. A Comparitive study is made on different machine learning models. Results are computed using train/test split.Then evaluation is done using cross validation method. After this the top models are further improved. Finally other approaches are discussed for further improvement like Stacking.

IV. Experimental Setup

The entire process is done using Python and data preprocessing is done in this stage 5. Data exploration and analysis also takes place here through correlation graphs 2. Data cleaning is done using null value removal and dimension reduction 3 4. Then outliers detection is done 8.Then new features are added if necessary. Then the data is checked for skewness 9

V. Experimental Results

After the data is loaded onto the system the model to be used is initialized 5. To find the best model many models are compared amongst each other and this comparison is done using the error in prediction given by each technique 6 7. Computation is done on training and testing data and models are evaluated using cross validation.

As per the experimental results Elastic Net preforms good in train/test with score (94.0779%) for training and score (89.5601%) for testing score. It has also low Root mean square error for training and testing. It reached high mean score with cross validation also (91.1784%) with Standard deviation (1.1132%). It also has a low mean value of Root mean square error values (11.6398%) for cross validation with Standard deviation (1.1601%) 10 11.

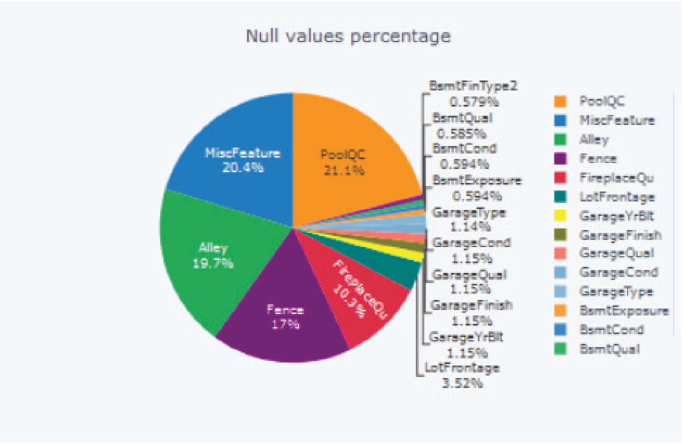


Fig. 3. Null Values

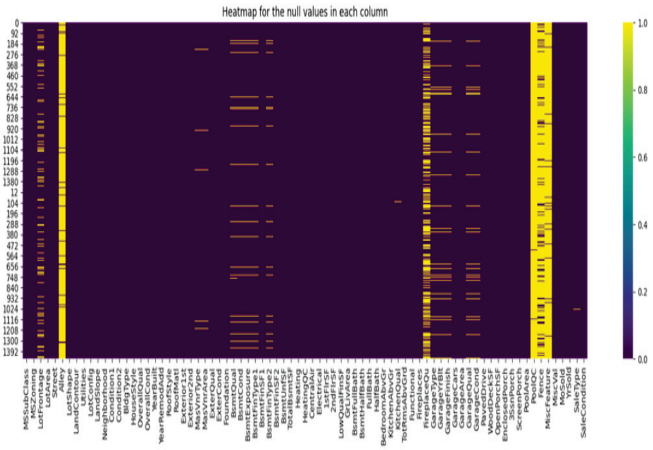


Fig. 4. Heat Map for Null Values per column

VI. Conclusion and Future Work

Data Mining has a lot of potential in real estate industry. Depending on the application many different models can be used as per requirement. One of the interesting methods by which accuracy can be increases is stacking. As per the setup used in this paper , it was found out that elastic net performed consistently in comparison to all models. So, in order to improve the accuracy even more Stacking was tried. Stacking is the best choice because unlike bagging, in stacking, the models that are used are different but the dataset is same.

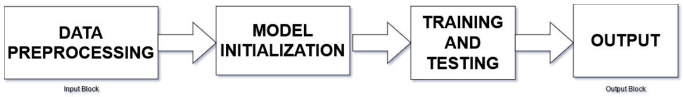
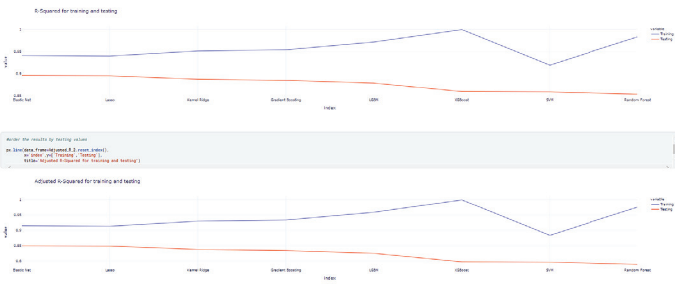


Fig. 5. Data Preprocessing and Model Initialization



```
#order the results by testing values

px.line(data_frame=Adjusted_R_2.reset_index(),
        x='index',y=['Training','Testing'],
        title='Adjusted R-Squared for training and testing')
```

Fig. 6. Output in R Square and Adjusted R square

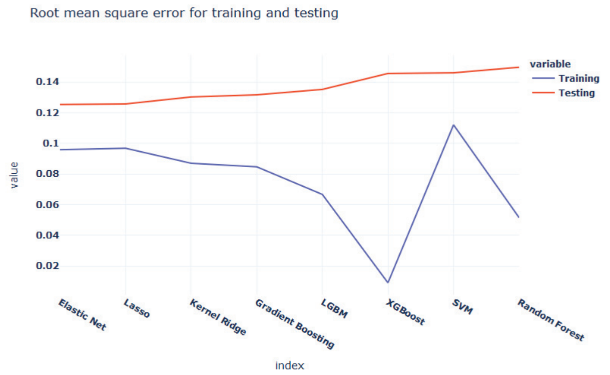


Fig. 7. Output in Root Mean Square

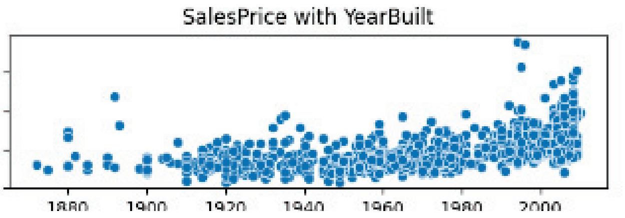


Fig. 8. Outlier Detection using correlation

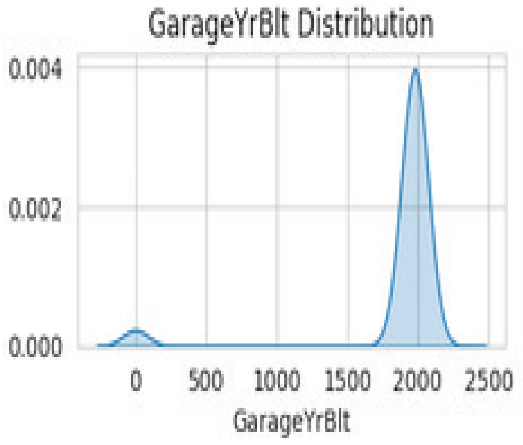


Fig. 9. Skewness Detection

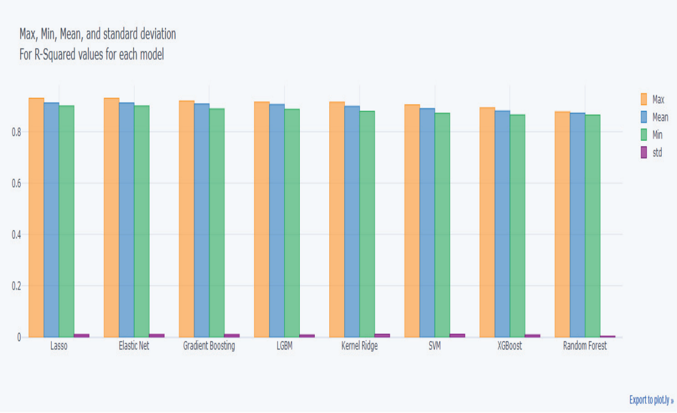


Fig. 10. Cross Validation Results

TABLE I
EXPERIMENTAL RESULTS FOR RMS ON TEST DATA

Model	Hold Out	Cross Validation
Elastic Net	0.94	0.91
Kernel Ridge	0.86	0.89
Lasso	0.93	0.90
Random Forest	0.51	0.90
SVM	0.89	0.89
XGBoost	0.009	0.87
LGBM	0.66	0.85
Gradient Boosting	0.85	0.86

When compared to boosting, stacking uses a single model that combines the outputs from the participating models. Elastic net, gradient boosting regression and lasso were used as participating models in stacking and the F-1 score was found out to be 96% which is already an improvement. However, it can be improved further depending on model selected.

REFERENCES

[1] Abdi, F., Abolmakarem, S.: Customer behavior mining framework (cbmf) using clustering and classification techniques. Journal of Indus-

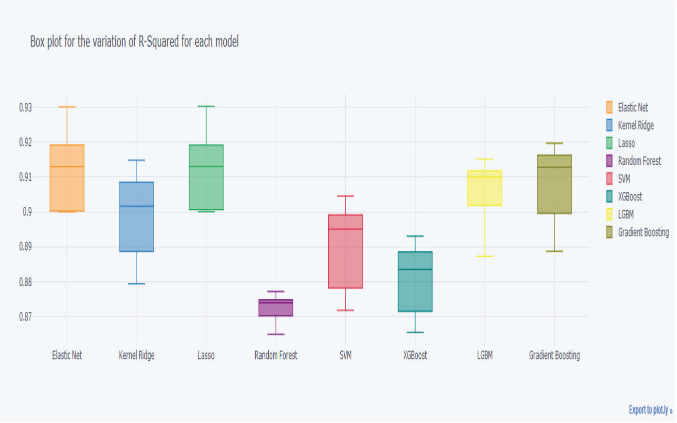


Fig. 11. Output in Boc Plot Graph

trial Engineering International 15(1), 1–18 (2019)

[2] Albahri, A., Hamid, R.A., Alwan, J.K., Al-Qays, Z., Zaidan, A., Zaidan, B., Albahri, A., AlAmoodi, A., Khlaf, J.M., Almahdi, E., et al.: Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (covid-19): a systematic review. Journal of medical systems 44, 1–11 (2020)

[3] Bakhshinategh, B., Zaiane, O.R., ElAtia, S., Ipperciel, D.: Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies 23(1), 537–553 (2018)

[4] Chung, H.M., Gray, P.: Data mining. Journal of management information systems 16(1), 11–16 (1999)

[5] Eickhoff, C., Kim, Y., White, R.W.: Overview of the health search and data mining (hsdm 2020) workshop. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 901–902 (2020)

[6] Hromada, E.: Real estate valuation using data mining software. Procedia engineering 164, 284–291 (2016)

[7] Jayanthi, R., Florence, L.: Software defect prediction techniques using metrics based on neural network classifier. Cluster Computing 22(1), 77–88 (2019)

[8] Mathan, K., Kumar, P.M., Panchatcharam, P., Manogaran, G., Varadharajan, R.: A novel gini index decision tree data mining method with neural network classifiers for prediction of heart disease. Design automation for embedded systems 22(3), 225–242 (2018)

[9] Narayana, G.S., Vasumathi, D.: An attributes similarity-based k-medoids clustering technique in data mining. Arabian Journal for Science and Engineering 43(8), 3979–3992 (2018)

[10] Nayak, J., Vakula, K., Dinesh, P., Naik, B., Mishra, M.: Ant colony optimization in data mining: Critical perspective from 2015 to 2020. In: Innovation in Electrical Power Engineering, Communication, and Computing Technology, pp. 361–374. Springer (2020)

[11] Patil, S., Nemade, V., Soni, P.K.: Predictive modelling for credit card fraud detection using data analytics. Procedia computer science 132, 385–395 (2018)

[12] Radanliev, P., De Roure, D., Walton, R.: Data mining and analysis of scientific research data records on covid-19 mortality, immunity, and vaccine development-in the first wave of the covid-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14(5), 1121–1132 (2020)

[13] Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10(3), e1355 (2020)

[14] Sangaiah, I., Kumar, A.V.A.: Improving medical diagnosis performance using hybrid feature selection via relieff and entropy based genetic search (rf-ega) approach: application to breast cancer prediction. Cluster Computing 22(3), 6899–6906 (2019)

[15] Senapati, M.R., Das, S., Mishra, S.: A novel model for stock price prediction using hybrid neural network. Journal of The Institution of Engineers (India): Series B 99(6), 555–563 (2018)

[16] Vijaya, J., Sivasankar, E.: Improved churn prediction based on supervised and unsupervised hybrid data mining system. In: Information and Communication Technology for Sustainable Development, pp. 485–499. Springer (2018)

[17] Wedyawati, W., Lu, M.: Mining real estate listings using oracle data warehousing and predictive regression. In: Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, 2004. IRI 2004. pp. 296–301. IEEE (2004)