Comparing Performance Metrics of Machine Learning Methods in Price Prediction of Houses

Aditya Raj
Student ID: x20143311
M.Sc in Data Analytics
National College of Ireland
Dublin, Ireland
x20143311@student.ncirl.ie

***Abstract* – The scope of this study is to compare extensively the performance metrics of various machine learning methods on housing datasets. This study is expected to carry out in accordance with KDD, Data Mining methodology. The datasets undergo rigorous data preprocessing to omit discrepancies, anomalies and patterns. We split the data into train and test to apply different regression models like Decision Tree Regressor, Random Forest Regressor, Extreme Gradient Boosting Regressor, Linear Regression, Extra Tree Regressor, Ridge Linear Regression, Elastic Net Regression and Lasso Regression. They are implemented on different housing datasets to draw inferences on the features involved in Price Prediction of houses in Melbourne, Germany and Beijing. We will evaluate the accuracy of models to conclude on choosing the model with best fit for each dataset.**

Keywords— Accuracy, Decision Tree Regressor, Elastic Net Regression, Extra Tree Regressor, Extreme Gradient Boosting, KDD, Lasso Regression, Linear Regression, Machine Learning, Price Prediction, Random Forest Regressor, Ridge Linear Regression

I. INTRODUCTION

World has witnessed a major advancement in technology with introduction to machine learning. The novelty of machine learning has drawn attention of people and organizations across mélange of domain. Due to its exceptional capabilities of application and efficacy, it is proclaimed to be one of the most desired skills in the market. Some of the major areas of implementation are Predictive Analytics, Forecasting, Virtual Assistants, Social Media Services, Fraud Detection, Online Recommendations etc.

Machine Learning is a part of an elite conglomerate that plays a latent but pivotal role in constructing the global economy. Another such affiliate is known as Property Investment. The enormity of transaction involved in Property Sell/Purchase is so large, that it needs to be scrutinized and evaluated watchfully. The role of machine learning has relegated the conventional idiosyncrasy embraced in an event of property investment. It addresses few possible concerns related to property purchase/sell like fraud (seller/buyer/agent) detection and price prediction. Accurate valuation of a house property is a decisive step to ensure whether a purchase is ideal. A set of historical records pertaining to purchase of house property can be efficiently used to train a model, to identify the factors that cater to price valuation.

This research will try to emulate a scenario-based case-study and provide insights related to machine learning methods and their performance on global housing datasets. We intend to develop an understanding of different machine learning methods and assess their performance metrics. We plan to develop a prototype of price prediction that can be generalized based on the details of property.

The datasets used in the study belong to multiple geographical locations scraped from Kaggle. The source of these datasets has been verified for veracity and authenticity.

A. Melbourne Housing Data (Year 2016 Data)

This is Melbourne housing clearance details, which includes Suburb, Address, Rooms, Type (House Variant), Method (Sale or Purchase Type), Bedroom, Land Size, Building Area, Year of Construction, Price, and more. This dataset contains 21 attributes and 34857 observations. Price is the target variable that can be estimated using the dataset's other independent variables.

B. Germany Housing Data (As of 14th July 2020)

This is German Real Estate dataset as of 14th July 2020. It has 26 columns and 10552 records in total with possible predictors like Living Space Area, Usable Area, Rooms, Bedrooms, Bathrooms, Floors, Year Built, Furnishing Quality, Renovation Year, Energy Consumption based fields, Garage and more. These attributes can manipulate the prediction of the prices.

C. Beijing Housing Data (Year 2011-2017)

This dataset includes the details of housing in Beijing from 2011 to 2017. It contains 26 columns and 318851 rows of data which stores the sale metadata and house particulars. Details like Trade Time (Transaction Date), DOM (Days on Market for Sale), Followers, Square Feet, Living Room, Drawing Room, Kitchen, Bathroom, Floor, Building Type, Renovation Condition and more, are the features that can be examined for potential predictors.

We have used regression methods like Decision Tree Regressor, XGBoost, Random Forest Regression, Extra Tree Regressor, Linear Regression, Ridge Linear Regression, Elastic Net Regression, Lasso Regression. All these methods are expected to generate set of performance metrics based on features (predictors) available in the dataset. These metrics can be dissected to reject or approve the model and examine its performance. The response variable in this study is the price attribute. The output of this research will help in fixating the most competent machine learning method to predict house prices and answer the research question below:

Which is the most efficient machine learning method based on performance metrics for optimum prediction of house prices?

The paper is further sub-categorized based on the methods of implementation and processes involved in Data Mining such as KDD (Knowledge Discovery in Databases). Furthermore, we have referred to several research papers as part of the literature review tangential to price prediction and outline their research findings, limitations and future scope. Moreover, we will elaborate on steps involved in Data Preprocessing, Distribution Plots, Correlations, Identifying Patterns and Variable Transformations. This will help in gauging the hyper parameters and feature selection for our model. Upon implementation of these machine learning methods, we will evaluate the performance metrics to conclude on the most optimized and precise model for our datasets.

II. Related Work

The vision of implementing machine learning in price prediction of houses has proved to be a revolutionary maneuver in real estate industry. This subsection outlines inferences, limitations and drawbacks from distinct research that were conducted in order to predict house prices. They help in laying foundation for our research and setting expectations with respect to each ML method.

A. Varma, A. Sarma, S. Doshi and R. Nair [1] developed a prediction model for housing prices that used Linear Regression, Boosted Regression, Forest Regression. The efficiency of the model is further increased with implementing neural networks. The limitations explained in their research is subjected to computational power, which can be augmented to enhance the output of the model. [6] Moreover, they pressurize on the possibility of integrating UI/UX design for interactive visualizations using Augmented Reality. Also, they suggest to gather users feedback to provide better recommendations.

J. Manasa, R. Gupta and N. S. Narahari [2] applied Linear Regression, Ridge Regression, [4] Lasso Regression, SVR (Support Vector Regression), XGBoost Regression Models to predict house prices. The XGBoost and Support Vector Regression models performed better than other models with R-square value of 0.78 and 0.79 respectively. There is a possibility of improvement in these scores with treatment of outliers and implementation of more advanced machine learning techniques like Neural Networks, Random Forest and Particle Swarm Optimization.

S. Neelam and G. Kiran [3] conducted a similar study by evaluating different machine learning techniques for price prediction of property. They have streamlined the process by analyzing the dataset and extracting correlation between parameters. They applied 4 different algorithms like Decision Tree Regression, Lasso Regression, Logistic Regression and Support Vector Regression with accuracies as 84.64%, 60.32%, 72.81% and 67.81% respectively. They critically evaluate the possibility of using a classifier method for a regression problem which boded well in their research [7]. They also infer that more volume of data can be beneficial in presenting a more sustainable model and even help in minimization of errors.

Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong [8] performed a research by implementing 3 machine learning algorithms, Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB). They use correlation matrix to establish the relationship between price and features of the dataset. It is observed that Housing floor area has the highest influence on the price of the property followed by Age of the Property, Distance and Travel time from Central District and Floor Level. Based on the performance of the all the methods, RF and GB outperforms SVM in terms of performance and GB is slightly better compared to RF for minimization of error. But it doesn’t conclude that one particular ML method supersedes the other due to the ambiguity resulting from specific data-oriented results.

C. R. Madhuri, G. Anuradha and M. V. Pujitha [9] investigated different regression methods for Forecasting Variations on House Price. They used extensive number of regression models like Multiple Linear Regression (MLR), Ridge Regression (RR), Lasso Regression (LR), Elastic Net Regression (ER), Adaboost Regression (AR) and Gradient Boosting Regression (GBR) for Forecasting the prices of houses with low error rate. Their research outlines that Gradient Boosting Regression performs with a higher accuracy and relatively lower error magnitude.

D. Banerjee and S. Dutta [5] have stated about various inferences that have been previously drawn by researchers about features that affect the rise or decline in house prices. One of the referred researches, J. Kahn [11] has explicitly outlined the erratic behavior of price against productivity growth which ascends for a certain duration and observes a decline later. D. Banerjee and S. Dutta [5] through their research, train their models using classification methods like Support Vector Machine (SVM), Random Forest (RF) and Artificial Neural Networks (ANN). In summary, the results show that Random Forest (RF) classifier outclasses in performance.

V. Sucharita, S. Jyothi and P. V. Rao [12] applied methods of machine learning in order to effectuate visual recognition of prawn species. Using Support Vector Classification, Neural Networks and K-Nearest Neighbour, they rivalled their performance statistics and observed that SVM classification exhibited extravagant results over its contemporaries.

Mauro Castelli, Maria Dobreva, Roberto Henriques, Leonardo Vanneschi [13] developed a model to predict the number of days a property remains in the market using Lasso Regression, Ridge Regression, Elastic Net Regression and Neural Networks. Lasso Regression outperformed other regression algorithms by selection of discriminators for prediction. The only limitation raised in this research was language of few variables was in Cryllic for which the translation was impossible and the volume of data could have been more for better prediction outcome.

Lu, Sifei & Li, Zengxiang & Qin, Zheng & Yang, Xulei & Goh, Rick [14] proposed and developed a hybrid regression model using Ridge, Lasso and Gradient Boosting. They use the concept of Creative Feature Engineering where Lasso is extremely effective in feature selection and helps in reducing Mean Square Errors of Ridge and Gradient Boosting Regression models. The best regression results for test data using hybrid model is 0.11260 with model influence of 65% Lasso and 35% Gradient Boosting Regression.

The findings from the above research elucidates that there are multiple factors that influence the performance of machine learning methods like feature selection, correlations, volume of test and train data on which the model is built. But there is no specific machine learning algorithm that judiciously works for all types of datasets with invariable split of categorical and numerical attributes.

III. METHODOLOGY

The Data Mining method used in conducting this research adheres to the standards of KDD (Knowledge Discovery in Databases). KDD consists of 5 primary stages that include Data Selection, Data Preprocessing, Data Transformation, Data Mining and Interpretation/Evaluation.

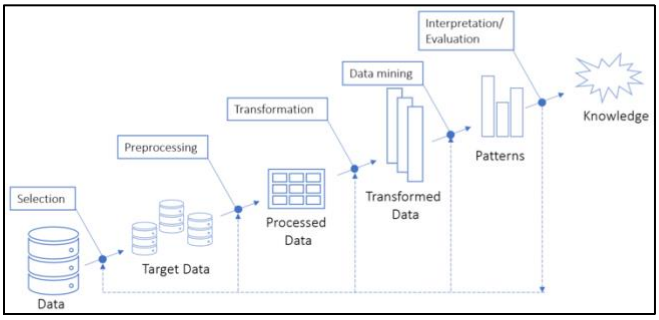

Fig. 1 KDD Process Breakdown

The first stage is selection of data from pool of resources that would facilitate our research. In our instance, we have identified Melbourne Housing, Germany Housing and Beijing Housing Datasets as our target datasets for implementation of machine learning methods. The second stage, Data Preprocessing caters to the requirement of omission of missing values and data discrepancies. The third stage, Data Transformation is not mandatory as it depends on the quality of data, whether it fulfils the pre-requisite of machine learning implementation. The fourth stage, Data Mining is usage of machine learning methods, application of statistics and mathematical functions to extract useful information from the processed data.
The final stage is evaluation of metrics and interpret the results to check the suitability of the machine learning model. These results help us to approve or discard the model.

A. *Melbourne Housing Dataset*

1) *Data Selection and Preprocessing*

We read the Melbourne Housing CSV file and store in the dataframe. We check for missing values in the dataset and visualize using a bar plot. We use Pandas Profiling to gain insights about our dataset. The output report of profiling is stored in HTML.

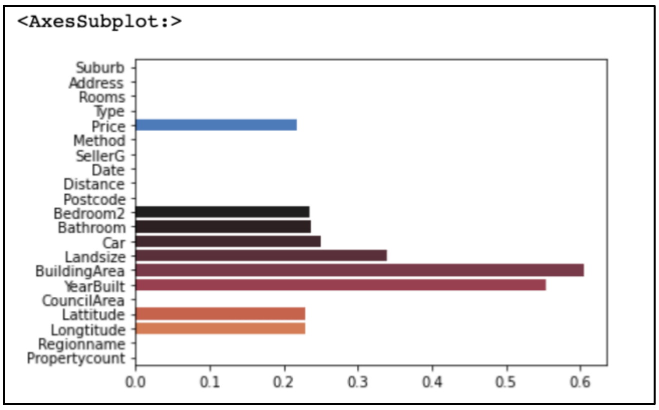

Fig. 2 Count of Missing Values for Variables in Melbourne Dataset

From the above plot, we can observe that BuilingArea and YearBuilt Variable has high percentage of Missing Values (>50%) and Price which is our Target Variable has around 20 percent missing values. We will visualize the distribution of Price across predictor variables like Type, Method, CouncilArea and Regionname.

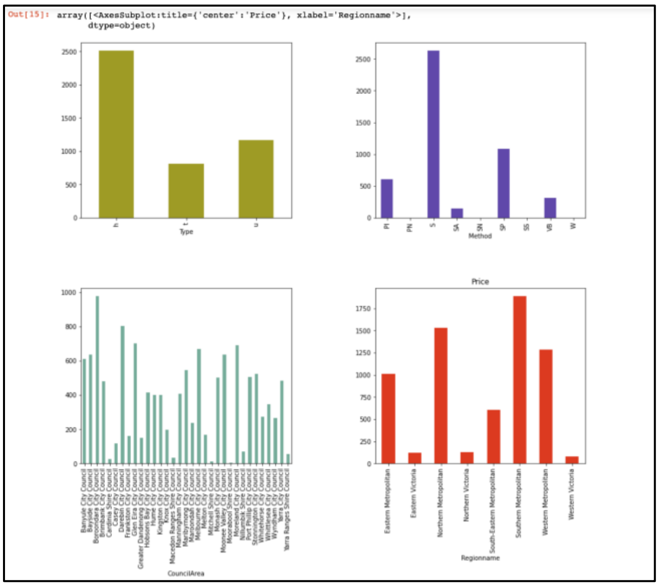

Fig. 3 Distribution of Price across Type, Method, CouncilArea and Regionname

From the above plot, we can see even distribution of price across the categorical values for these variables.

2) *Data Transformation*

As part of data preprocessing, we will try to evaluate the correlation matrix of numerical variables from the dataset.
We extract the column names of numerical variables from the dataset for building a correlation matrix.

From the below correlation matrix, it is observed that Room and Bedroom have a strong correlation of 0.96, we will drop Bedroom2 column from our dataset. We also drop the rows with missing values of Price.

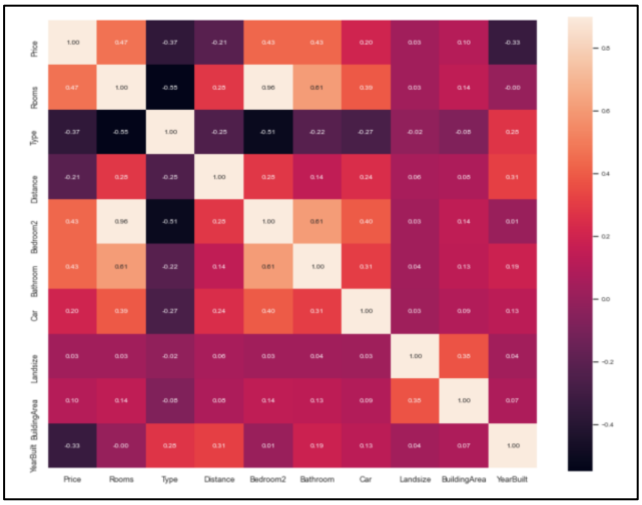

Fig. 4 Correlation Matrix of Numerical Variables

3) *Data Mining*

The preliminary step of Data Mining is Splitting the data into Train and Test. I have also used PCA algorithm to reduce the number of features to expedite the model training. We then move on to implementation of regression algorithms.

a) Decision Tree Regressor

The is a machine learning method that follows the tree like structure to predict values of target variable based on observations. Decision Trees are of two types- Classification and Regression. Decision Tree Classification predicts the target variable with discrete values whereas Decision Tree Regression predicts the target variable with continuous values.

b) Xtreme Gradient Boosting

XGBoost is a high-speed and high-performance gradient boosted decision tree implementation.

The loss function of our base model (e.g., a decision tree) is used as a proxy for minimizing the overall model's error in regular gradient boosting. As a rough approximation, XGBoost uses the 2nd order derivative.

The interpretation of results will be comprehended in the next section.

B. Germany Housing Dataset

1) *Data Selection and Preprocessing*

We selected this dataset due to its difference with first dataset geographically and few more additional variables. We start with dropping an additional index column from this dataset. We check for missing values in the Germany dataframe and visualize the results using a plot.

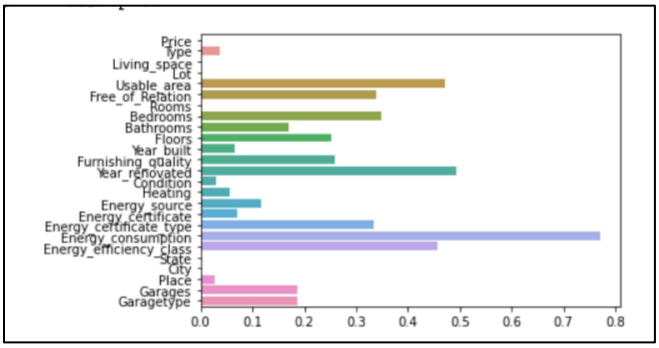

Fig. 5 Count of Missing Values in the Germany Dataset

From the above plot, we can infer that Energy Consumption variable has 80% missing values in the dataset. This column will not assist the model in prediction of price of the property.

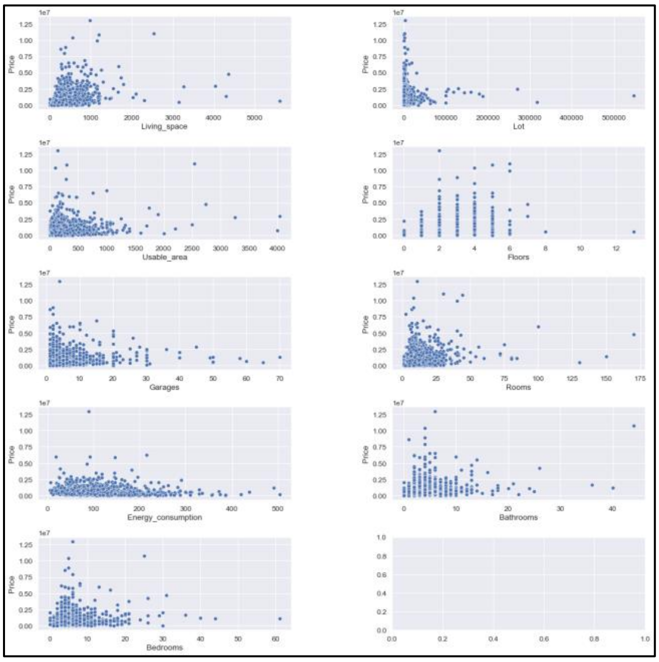

Fig. 6 Price Distribution Plots in the Germany Dataset

We can observe the random distribution through scatter plots of Price with different numerical as well as categorical variables within the Germany dataset.

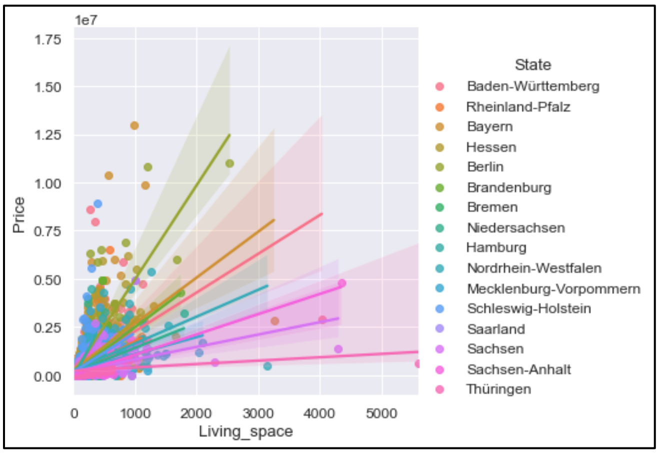

Fig. 7 Price vs Living Space across States in the Germany Dataset

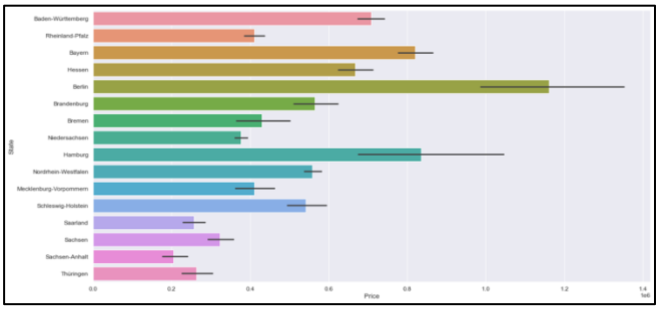

Fig. 8 Price Dsistribution across States of Germany

From the above first plot, we can infer that Bayern and Hessen have high prices of property with small values of Living Space. Whereas, from the second plot, we can infer that Berlin has the high prices of house followed by Bayern and Hamburg.

2) *Data Transformation*

We check for missing values in the dataset and try to impute values into two columns of Germany dataset. We assign mean values to Year Built and Garages. We drop all the other rows with missing values.

```
Missing Data :

 Price                     0
Type                      0
Living_space              0
Lot                       0
Usable_area               0
Free_of_Relation          0
Rooms                     0
Bedrooms                  0
Bathrooms                 0
Floors                    0
Year_built                0
Furnishing_quality        0
Year_renovated            0
Condition                 0
Heating                   0
Energy_source             0
Energy_certificate        0
Energy_certificate_type   0
Energy_consumption        0
Energy_efficiency_class   0
State                     0
City                      0
Place                     0
Garages                   0
Garagetype                0
dtype: int64
```

Fig. 9 Missing Values after dropping rows and imputation

All the variables are of integer datatype and we can directly create a correlation matrix to decipher the collinearity of variables.
From the below Correlation Matrix, it can be observed that Living Space, Rooms, Bathroom, Bedrooms have slightly high collinearity. In order to prepare data for model implementation, we created dummy values for Type, Free_of_Relation, Furnishing_quality, Condition, Heating, Energy_source, Energy_certificate, Energy_certificate_type, Energy_efficiency_class, State, City, Place, Garagetype. The resultant dataframe had 4924 columns upon creating dummy values.

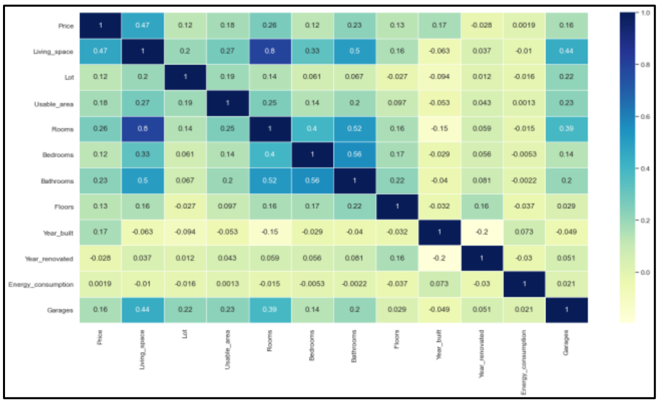

Fig. 10 Missing Values after dropping rows and imputation

3) *Data Mining*

Once the dummy variables are created, we create test and train split of the data of 80-20. We perform different machine learning methods on modified dataset with dummy variables.

a) *Linear Regression*

It analyses the linearity between the target variable (quantitative) and independent variable from the dataset to train the model. It predicts the possible outcome based on the linear relationship and some underlying assumptions.

b) *Decision Tree Regressor (Explained Previously)*

c) *Random Forest Regressor*

This method is inspired by concept of Decision Trees, it creates a "forest" of multiple decision trees on training dataset and returns predicted class or numerical output for corresponding classification and regression type problems. This model is said to have an edge over decision trees by resolving the complexity of overfitting exhibiting a better performance over the former.

d) *Ridge Linear Regression*

Ridge Linear Regression is the more advanced version of linear regression that is effective because there are a lot of variables and the data is multicollinear. Ridge Regression, unlike Lasso Regression, never reduces the coefficients to zero because it does not eradicate redundancy but rather through their impact on the trained model.

e) *Lasso Regression*

Lasso regression is a form of shrinkage-based linear regression. Data values are shrunk into a central point, such as the mean, in shrinkage. Easy, sparse models are encouraged by the lasso technique (i.e. models with fewer parameters).

f) *Elastic Tree Regression*

Elastic Net is a linear regression extension that incorporates regularization penalties into the loss function during testing.

C. *Beijing Housing Dataset*

1) *Data Selection and Preprocessing*

This dataset captures the housing details of Beijing from the year 2011-2017. It contains indifferent attributes like Days on Market, Followers, Building Type, Construction Time, Ladder Ratio, elevator etc. We plot different subplots to analyze the distribution of values count of different variables within the dataset.

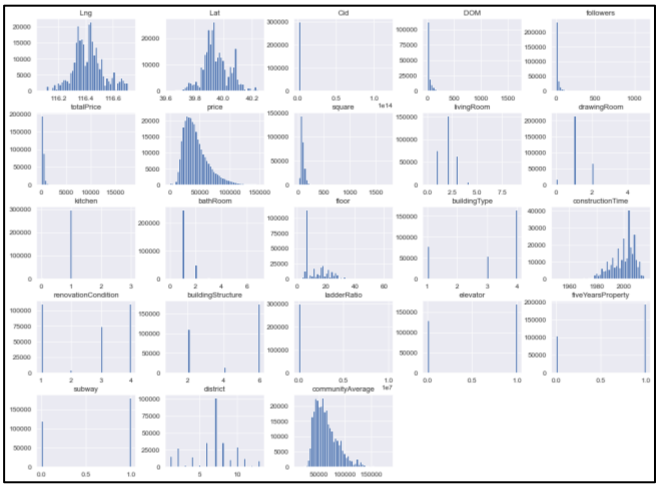

Fig. 11 Distribution Plots of variables within the Beijing Dataset

From the above plot, it seems the values and count seem to be evenly distributed. Few attributes like Price and Community Average seem to be right skewed and Construction Time is left skewed.

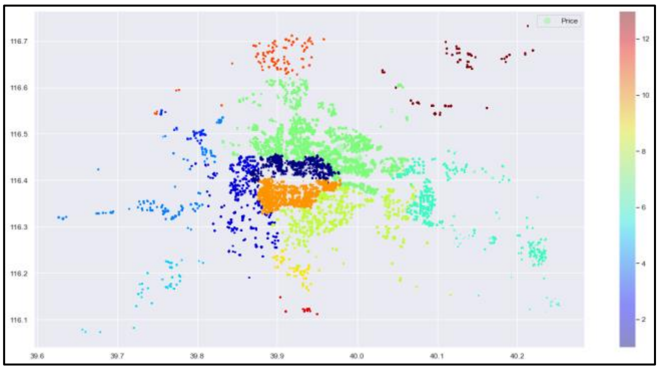

Fig. 12 Scatterplot of Price across District based on Geolocation (Lat,Long)

2) Data Transformation

We drop meta data columns from our dataframe like URL, ID, CID, DOM which are not influential in price prediction of houses. We analyze correlation matrix of our data for all the numerical fields available in our dataset.

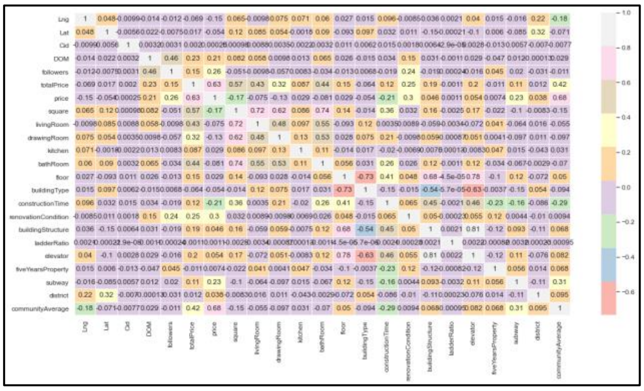

Fig. 13 Correlation Matrix of Numerical Variables in Beijing Dataset

We also extract year from Tradetime using the split function. We change the datatype of few columns like Building Type, Building Structure, Renovation Condition and District to String. Furthermore, we create dummy variables for all the categorical columns and new modified dataframe has 65 columns.

3) Data Mining

Upon creating dummy variables, we create a test and train split. We then apply set of machine learning methods on the train data to train our model and predict values on testing data.

a) Random Forest Regressor (Explained Previously)

b) Decision Tree Regressor (Explained Previously)

c) Xtreme Gradient Boosting (Explained Previously)

d) Extra Tree Regresor

Extra Tree Regressor is referred to as regressor of additional trees. This class introduces a meta estimator that uses averaging to increase statistical precision and control over-fitting by fitting a variety of randomized decision trees (a.k.a. extra-trees) on different sub-samples of the dataset. Consider min samples split as the minimum number.

IV. Evaluation Methods

A. Melbourne Housing Dataset

Upon the implementation of Decision Tree and Random Forest, the accuracy was evaluated and observed that Decision Tree outperformed Random Forest.

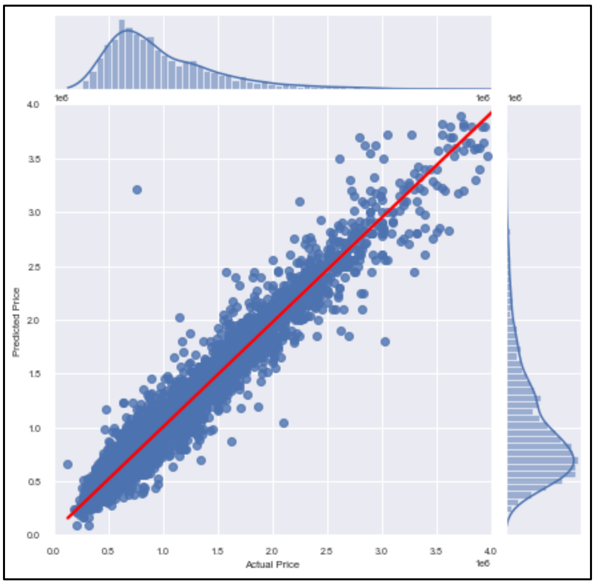

Fig. 14 Decision Tree Graph of Predicted vs Actual Price

From the above plot, we can see the linearity of Predicted vs Actual Price using the Decision Tree Regression Algorithm.

Melbourne Housing Dataset		
Decision Tree Regressor	Mean Absolute Error	Accuracy
	604.51	99.7937
Random Forest Regressor	Accuracy	
	Training	Testing
	82.84	76.94

Fig. 15 Performance Metrics of ML Methods on Melbourne Dataset

We also use built-in function which is used for Feature Selection of Top 10 features from the dataset than be used to build the model. From the plot below, it is observed in order that Distance, Landsize, BuildingArea, etc are influential when building a model based on XGBoost algorithm.

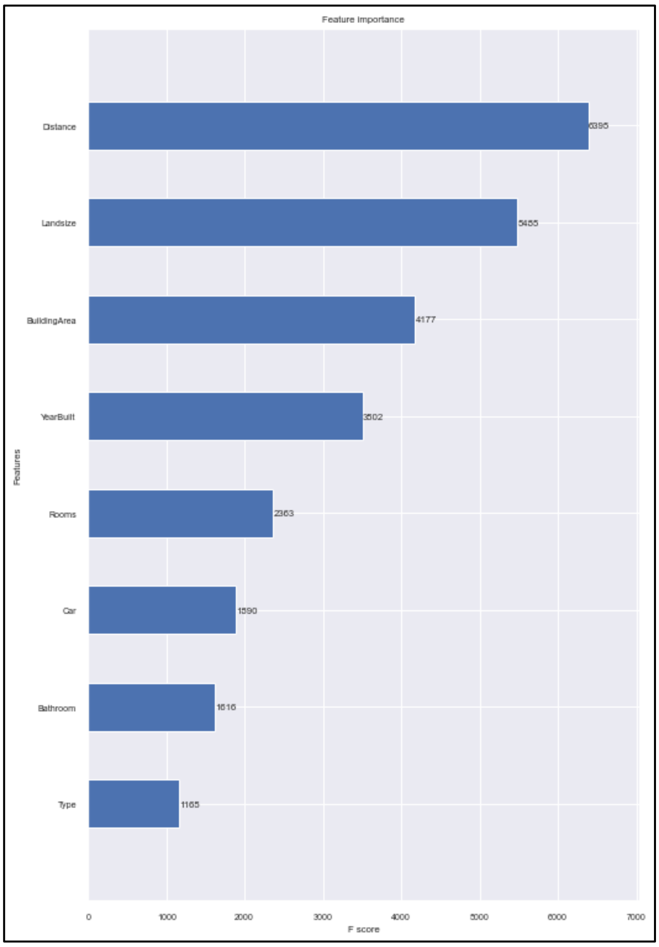

Fig. 16 Plot for Features Importance

From the performance metrics, we can say the Decision Tree was the ideal choice of machine learning algorithm for predicting price on Melbourne Dataset with an Accuracy of 99.79% and Mean Absolute Error of 604.51.

B. *Germany Housing Dataset*

We have implemented Linear Regression, Decision Tree Regressor, Random Forest Regressor, Ridge Linear Regression, Lasso Regression, Elastic Net Regression on Germany Dataset and analyzed their performance metrics.

Germany Housing Dataset				
Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
Linear Regression	229209.83	1.60959E+11	401197.02	42.16
Decision Tree Regressor	236632.32	2.0307E+11	450633.12	27.03
Random Forest Regressor	176718.46	1.24424E+11	352737.94	55.29
Ridge Linear Regression	176718.46	1.24424E+11	352737.94	43.82
Lasso Regression	230520.84	1.66229E+11	407712.25	40.27
Elastic Net Regression	186601.39	1.2559E+11	354387.32	54.87

Fig. 17 Performance Metrics of ML Methods on Germany Dataset

Random Forest Regressor performs the best out of 6 ML methods with an accuracy of 55%. The plot below shows the actual vs predicted values for Random Forest Regressor model.

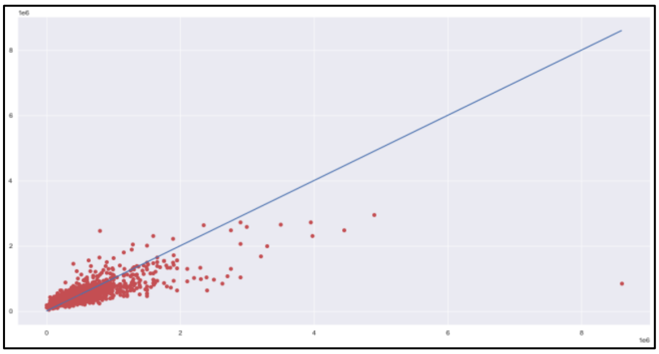

Fig. 18 Plot of Actual vs Predicted values of Random Forest

Therefore, we can say that Random Forest with Search Grid Validation was the ideal and optimum machine learning method to be used on Germany Housing Dataset with an Accuracy score of 55.29, MAE of 176718.46, MSE of 124424058967.42 and RMSE of 352737.94.

C. *Beijing Housing Dataset*

We performed implementation of Random Forest Regressor, Decision Tree Regressor, XGBoost and Extra Tree Regressor on Beijing Dataset.

Beijing Housing Dataset				
Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Accuracy
Random Forest Regressor	35.19	5834.7	76.38	89.62
Decision Tree Regressor	48.37	7529.13	86.77	86.6
XGBoost	35.5	5303.56	72.82	90.56
Extra Tree Regressor	36.91	4784.52	69.17	91.48

Fig. 19 Performance Metrics of ML methods on Beijing Dataset

The performance metrics imply that Extra Tree Regressor presented the best results over other models for Beijing Housing with an Accuracy of 91.48, MAE of 36.91, MSE 4784.52 and RMSE of 69.17.

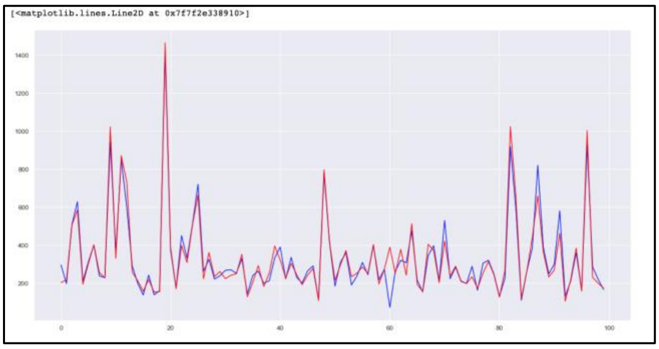

Fig. 20 Extra Tree Regressor Model – Actual vs Predicted Values

V. Conclusion and Future Work

The research conducted in this paper evaluated distinctive set of regression machine learning methods on housing datasets across geographical locations (Melbourne, Germany and Beijing). This paper is expected to pose answers to the research question about selection of machine learning with optimum performance against datasets of mercurial house parameters. Decision Tree, Random Forest and Extra Trees Regressor are all tree-based regression methods that suited our datasets.

This paper has scope for emphasis on future work with possible inclusion of more validation methods like cross-validation, k-fold cross validation, Leave-one-out Cross Validation, Nested Cross Validation, Time-series Cross Validation etc. The performance metrics of Beijing Dataset has room for additional optimization by possible use of hybrid regression methods.

VI. References

[1] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[2] J. Manasa, R. Gupta and N. S. Narahari, "Machine Learning based Predicting House Prices using Regression Techniques," *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 624-630, doi: 10.1109/ICIMIA48430.2020.9074952.

[3] S. Neelam and G. Kiran, "Valuation of house prices using predictive techniques", *Internal Journal of Advances in Electronics and Computer Sciences*, no. 6, 2018.

[4] S. Abhishek, "Ridge regression vs Lasso" in How these two popular ML Regression techniques work, Analytics India magazine, 2018.

[5] 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)*, 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.

[6] R. T. Azuma et al., "A survey of augmented reality," Presence, vol. 6, no. 4, pp. 355–385, 1997.

[7] Torgo, Luis, and Joao Gama. "Regression using classification algorithms." Intelligent Data Analysis 1.4 (1997): 275-292.

[8] Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong, "Predicting property prices with machine learning algorithm", Journal of Property Research, Volume 38, 2021, Issue 1.

[9] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," *2019 International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[10] 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, India, 2017, pp. 2998-3000, doi: 10.1109/ICPCSI.2017.8392275.

[11] J. Kahn, "What drives housing prices?", *Federal Reserve Bank of New York Staff Reports, 2008*.

[12] V. Sucharita, S. Jyothi and P. V. Rao, "Comparison of machine learning algorithms for classification of Penaeid prawn species," *2016 3rd International Conference on Computing for Sustainable Global Development* (INDIACom), New Delhi, India, 2016, pp. 1610-1613.

[13] Mauro Castelli, Maria Dobreva, Roberto Henriques, Leonardo Vanneschi, "Predicting Days on Market to Optimize Real Estate Sales Strategy", *Complexity, vol. 2020, Article ID 4603190, 22 pages, 2020*. https://doi.org/10.1155/2020/4603190

[14] Lu, Sifei & Li, Zengxiang & Qin, Zheng & Yang, Xulei & Goh, Rick. (2017). "A hybrid regression technique for house prices prediction" 319-323. 10.1109/IEEM.2017.8289904.

[15] https://www.kaggle.com/shubhankartiwari/melbourne-houses

[16] https://www.kaggle.com/ruiqurm/lianjia

[17] https://www.kaggle.com/scriptsultan/german-house-prices