Residential Property Price Prediction Using Machine Learning: MakanSETU

2022 International Seminar on Application for Technology of Information and Communication (iSemantic) | 978-1-6654-8839-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/iSemantic55962.2022.9920395

Mr. Yash Panchal
Department of Computer Engineering
Universal College of Engineering
University of Mumbai
Vasai, India
yashp.2920@gmail.com

Mr. Manan Mer
Department of Computer Engineering
Universal College of Engineering
University of Mumbai
Vasai, India
mananmer9769@gmail.com

Mr. Abhiroop Ghosh
Department of Computer Engineering
Universal College of Engineering
University of Mumbai
Vasai, India
abhiroop.ghosh23@gmail.com

***Abstract*—MakanSETU is an emerging and advanced solution in the Real Estate industry. Real Estate Industry is at boom in the 21st century and trading Real Estate has become a great opportunity for Real Estate owners as well as others. The projection of Real Estate industry in business acquisitions is expected to reach 11 trillion USD. However, there is no proper solution to deal with inaccurate prices of properties online. The system proposed in this paper uses Native and new age Machine learning algorithms to predict and validate value of residential properties. Supervised learning is used in the system along with multiple Regressors to obtain the best result. Some of the regression algorithms used are Simple Linear regression, Decision tree regression, Random Forest regression (100 n-trees, 200 n-trees, and 500 n-trees), and Extreme Gradient Boost regression algorithm. The development of this system has followed a series of Data Collection, data handling, data processing, EDA, Feature engineering and Feature selection. The system enables investors to get a fair value of a property. The system is considered successful and ready to implement in the real work.**

Keywords— Machine Learning, Supervised Learning, Real Estate, Real Estate Price Prediction, Regression Algorithm, Data Analysis, Feature Engineering, Real Estate Analysis

I. INTRODUCTION

Machine learning is a field of AI that uses algorithms and technologies to extract meaningful information from enormous volumes of data. Global pandemic constraints have had a direct impact on traditional real estate operations — and in an unexpected way, for the better. Thousands of firms, realtors, appraisers, mortgage lenders, and other businesses have been pushed to incorporate quickly expanding PropTech [1] to manage the situation, and with good reason. In the short term, real estate AI apps may handle planned data flows, learn user behavior, streamline, and carry out operations, as well as provide more accurate assessments and market forecasts. There have been several attempts to use Machine Learning [2] and Artificial Intelligence to assist real estate, but there have always been challenges with accuracy, resulting in employment losses in this area. This project uses artificial intelligence to increase accuracy while also offering a platform for real estate brokers. It not only makes it safer for individuals to examine their needs, but it also protects clients from bogus data. MakanSETU offers value evaluations, value prediction, and CRUD operations to the Real Estate agents and end users in a single platform which is based on Python based "Django MVT framework". This project also includes the integration of several modules, which improves accuracy and customer satisfaction.

II. LITERATURE REVIEW

Sayan Putatunda presented research, "PropTech for Proactive Pricing of Houses in Classified Advertisements in the Indian Real Estate Market" [1], in which a machine learning strategy is proposed for addressing the problem of house price prediction in classified ads specific for Indian real estate sector. It checks the performance of various machine learning algorithms such as Random Forest, Gradient boosting, and Artificial neural networks using a real-world dataset. In terms of prediction accuracy, he has discovered that the Random Forest technique is the best. The only back fall of this work is very less Data to train and work upon.

Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, and Wayan Firdaus Mahmudy present research, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization" [2], in which they utilize regression analysis and particle swarm optimization (PSO) technique to estimate housing prices in Malang based on "Nilai Jual Objek Pajak (NJOP)" prices. PSO is used to choose the influenced variables, and regression analysis is used to get the best coefficient for prediction. To predict home prices, several studies utilizing linear regression and particle swarm optimization methods have been conducted. The optimum parameter settings were 1800 particles, 700 iterations, with inertia weights of 0.4 and 0.8, resulting in a lowest prediction error of IDR 14.186. The other model's error prediction values remain high. The suggestion is to test Non-linear Machine learning algorithms for better accuracy.

The models made in "Real estate value prediction using multivariate regression models" [3] work by R Manjula, Shubham Jain, Sharad Srivastava, and Pranav Rajiv Kher using different features such as the square feet of the home, the number of bedrooms, the atmosphere, and so on. As a result, each feature in their model is assigned a weight using "Feature Engineering", which defines how relevant that feature is to their model's prediction. "Zillow.com" and "Magicbricks.com" are two examples of companies that have a large dataset of home prices that they use to predict values using machine learning. They calculated the root mean squared error value for each of the models as suggested. It used the

following approaches to achieve this: Polynomial Regression, Simple Regression Model, Multivariate Regression Model. Use of mixture of these models might give high bias while a high complexity model gives high accuracy.

According to Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei's work "Housing Price Prediction through Improved Machine Learning Techniques" [4], the House Price Index (HPI) is used to measure price variations in residential housing in various countries. The HPI is a repeat sales index that analyses average price changes in repeat sales and refinancing on the same properties. Different algorithms like XGBoost, LightGBM and hybrid models for the coupling effect are used but the best accuracy was found in Stacked Generalization Regression despite of its complexity on a dataset of 300000 data instances. Future work includes examining tree-based regression models for better accuracy.

A Decision tree machine learning technique is used to develop a prediction model to estimate selling values for any real estate property in the work, "House price forecasting using Machine Learning" [5], given by Alisha Kuvalekar, Shivan Manchewar, Sidhika Mahadik, and Shila Jawale. Additional parameters such as air quality and crime rate were added in the dataset to help with price estimation. These characteristics are uncommon in other prediction system's datasets, which differentiates this system. To connect the learned model to the user interface, the Flask Framework is employed. When it comes to predicting real estate values, the method has an accuracy of 89 percent.

Few scholars holding PhD, Dr. G. Naga Satish, Dr. V. Raghavendran, Mr. M.D. Suganan Rao, and Dr. Ch. Srinivasulu, have found in their work, "House Price Prediction Using Machine Learning" [6], that Lasso Regression and Boosting algorithm have best estimations in Lodging value prediction. Lasso regression was chosen due to its better adaptivity on model selection. However, Lasso Regression does not out stand their expectations while trained on the dataset when compared to XGBoost Regressor. The work suggests a good User interface which can ease up the process for the end users for price prediction.

As, Maharshi Modi, Ayush Sharma, and Dr. P. Madhavan says in the research work, "Applied Research On House Price Prediction Using Diverse Machine Learning Techniques" [7], it is viable for the researchers to develop more accurate and efficient models, nowadays. A model which offers more enhanced and accurate result with a user friendly interface has been developed which has overcome the previous less accurate and overfitting models. Various model including Tree-based models, probabilistic models, Classification models, and Boosting models are used and coupled using stacking technique. The work suggests to use ensembling technique to improve the accuracy in future.

III. Proposed System

MakanSETU is a fully fledged platform for real estate agents as well as end users. MakanSETU proposes a system architecture which offers number of operations for the users. Price predictions and Value evaluation are some of the main focuses of the project. The detailed system architecture of MakanSETU is shown in Figure 1. The proposed system is a website where algorithms which are trained using millions of data instances is used, looking for property valuations, location, house modifications, and even some uncommon features of the residential property. This hybrid strategy, which employs numerous algorithms as well as regressions, yields far more accurate results. The web technology assists both purchasers and real estate brokers, ensuring that no human resources or jobs are lost. Investors or purchasers have certain specifications that they may enter into the portal, which are then compared to the database as well as estate agency inputs to select the most suitable property. To maintain

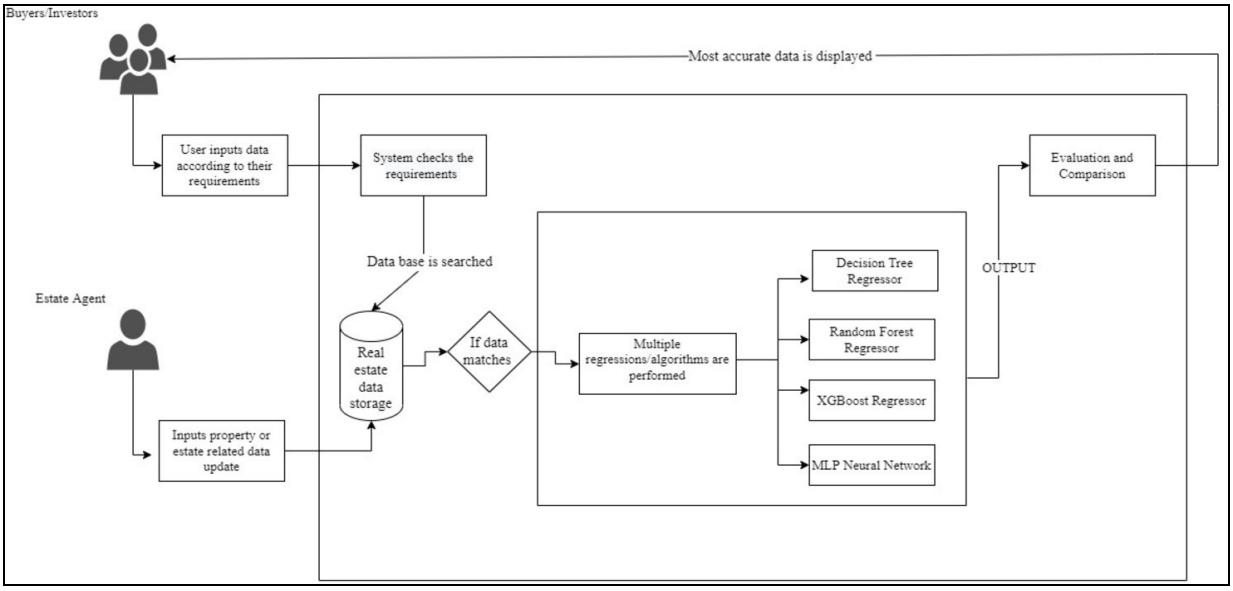

Fig. 1: System Architecture of MakanSETU

the model efficient, feature engineering was combined with machine learning. Features are defined as inputs such as number of bedrooms, hall, kitchen, square foot, and so on, which are then searched in the data store. Multiple techniques, such as Decision Tree Regression, Random Forest Regression, XGBoost Regressor, and MLP Neural Network, are used after locating the relevant and cleansed data. When the output or results are received, the accuracy is analyzed and compared to discover the best possible outcome; once the evaluation is complete, the buyer is given the most accurate data. Various modules have been incorporated to substantially increase accuracy.

IV. Methodology

Several machine learning methods are implemented in the proposed framework, which are subsequently considered as the predicting regressors. Different machine learning regression algorithms are compared and analyzed, and the results are recorded.

A. Data Collection

Data is the most important part of any machine learning program. Data is the base on which a ML algorithm can perform learning the trends to create a solution. It can be in different forms numerical, categorical, binary, etc. Most of the data on internet are unstructured, but platforms like Kaggle and many other provides structured datasets for research and study purpose. The Dataset, “Housing Prices in Metropolitan Areas of India”, that we have used is sourced from Kaggle [8] also another dataset from GitHub, “House-price-prediction” [9] is used for training different ML algorithms for MakanSETU.

B. Exploratory Data Analysis and Data Processing

Data are classified as structured and unstructured data. Data is present on the web in variety of forms text, numerical, binary, etc. But to train specific ML model we need dataset in a format specific to the Algorithm used in the model. MakanSETU is trained on 2 different datasets combined in a single dataset. Data processing means formatting the dataset in the required format to get the desired output from the model. In this project the datasets were preprocessed to overcome missing data, and converting text categorical data into numerical data by one-hot-encoding on Kaggle. Merging two different datasets, [8] and [9], and formatting them into required format was a main task in this project. After formatting, the dataset undergoes a set of operation to get the insights from the data.

Figure 2 visualizes the trend between one of the most importance features in the prepared dataset, number of bedrooms with the output variable Price.

Similarly, Figure 3 shows the relation between the data values of feature Area and Price. Visualizing is the best way to understand the data and what the data wants us to know.

C. Feature Selection

It's the process of selecting appropriate characteristics for your machine learning model based on the sort of problem you're attempting to answer. It is accomplished by considering features which are most impactful. It helps in the reduction of noise in our data as well as the amount of our input data. Obtaining Feature importance using the correlation between them is the most common method of Feature selection used in the industry.

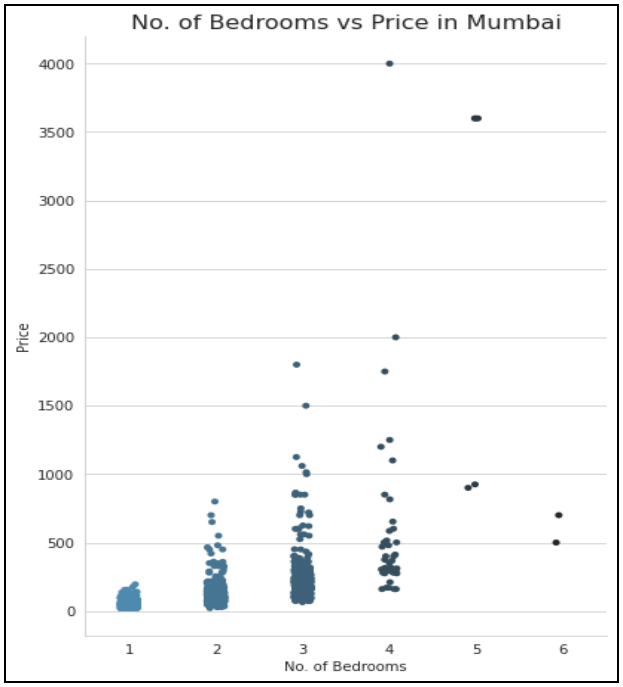

Fig. 2: Number of Bedroom vs Property Price(Rs.100k)

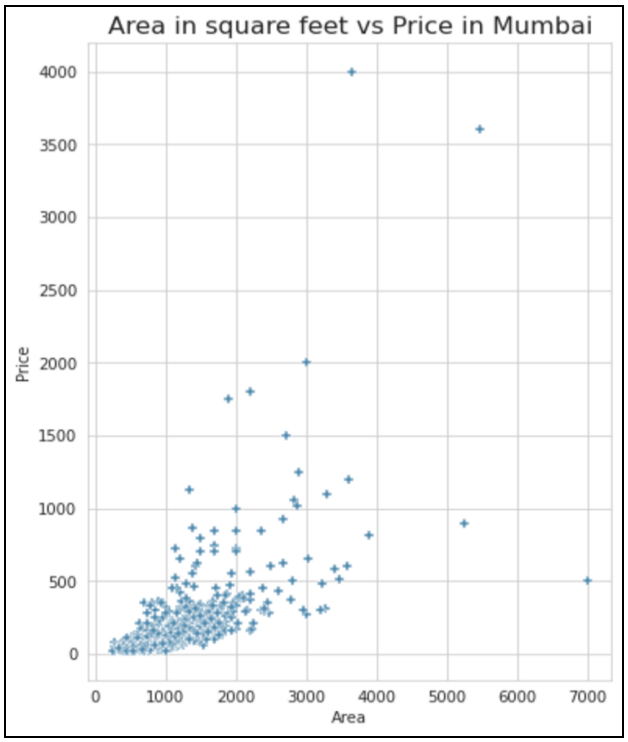

Fig. 3: Area(sq. ft.) vs Property Price(Rs.100k)

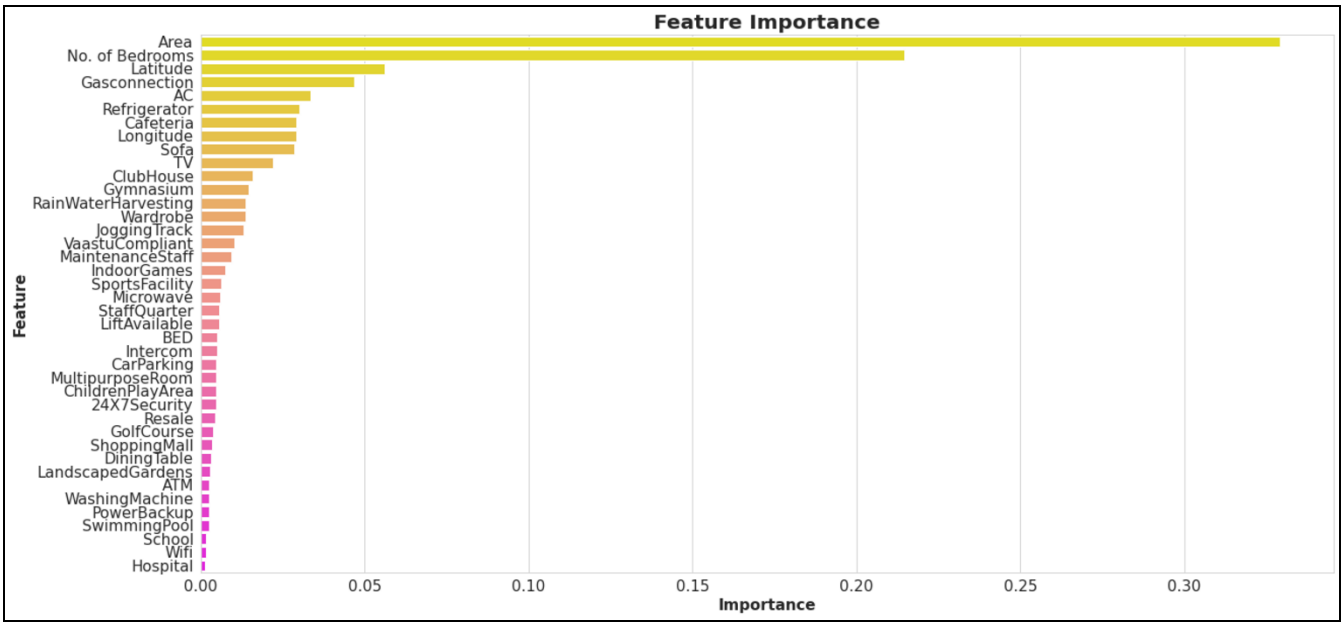

Fig. 4: Importance of Features using correlation with Property Price

In this project, Correlation of various features is calculated, to understand how different features are related to each other. Based on the permutation importance the Feature importance is evaluated for specific models like shown in the Figure 4.

D. *Model training*

For training different models of ML and DL, Scikit Learn module is used.

a) *Non-Linear Regression algorithms:* Decision Tree Regressor and Random Forest Regressors are used as the models to predict the property price in MakanSETU. A random forest is made up of a large number of individual decision trees that operate together as an ensemble, as the name suggests. The random forest creates a class prediction for each tree, and the class with the highest votes becomes our model's forecast.

b) *Gradient Boosting algorithm:* Gradient boosting is a machine that boosts the gradient of a trend. In the direction of straight-forward use of boosting computations there are a significant number boosting calculations in, CatBoost, Gradient Boosting, XGBoost, LightGBM, AdaBoost, Gentle Boost etc. Each boosting algorithm requires its own math. In addition, a minor variation can be observed at the same time as they are being applied. MakanSETU uses XGBoost algorithm for prediction as it has shown the best results. Although the XGBoost model generally achieves greater accuracy than a single decision tree, it does so at the expense of decision trees' inbuilt generalization ability.

c) *Simple Linear Regression algorithm:* The relationship between two continuous quantitative variables may be summarized and studied using a simple linear regression statistical approach.

- The predictor, explanatory, or independent variable is designated by the letter x.
- The response, result, or dependent variable is the other variable, denoted by y.

d) *Support Vector Regression algorithm:* The technique of supervised learning to predict discrete values, Support Vector Regression is employed. Both Support Vector Machines(SVMs) and Support Vector Regressors(SVR) are built on the same foundation. The primary assumption of SVR is to find the best-fitting line. In SVR, the hyperplane with the highest number of points is the best fit line. In MakanSETU, the SVR is trained using a Gaussian kernel rather than a Linear Kernel.

e) *Artificial Neural Network algorithm:* An artificial neural network technique with several layers is the multi-layer perceptron (MLP). Although obviously linear issues in the dataset used in MakanSETU may be addressed with a single perceptron, it is not well suited to non-linear instances. MLPs are neural network models that can estimate any continuous function as universal approximators.

V. RESULT

All the Models that were trained gave some error as there cannot be an ideal model for a solution. For regression Models Root Mean Square Error, Mean Square Error and R^2 coefficient is considered in order to compare them. In this project Regression algorithms are implemented and some parameters are being considered for model selection. In Table 1, Training/model score and R^2 score of different Machine learning Regressors are being compared. The R^2 coefficient/score represents the amount of variance in the output that our model is capable of predicting based on its features.

$$R^2(y_{true}, y_{pred}) = 1 - [\Sigma (y_{true}-y_{pred})^2 / \Sigma (y_{true}-\bar{y})^2]$$

where, $\bar{y} = (1/n_{samples}) \Sigma y_{true}$

And, Training score, also called as model score, is the R^2 value obtained by the default model.score() function of scikit learn library which uses multioutput =

'uniform_average' from scikit learn version 0.23 to keep consistent with default value of model.r2_score(). It clearly shows that XGBoost model has the highest R^2 Score and Training score.

TABLE I. TRAINING SCORE AND R^2 SCORE

Model	Training score	R^2 Score
Decision Tree Regressor	0.999	0.675
Random Forest Regressor	0.952	0.723
XGBoost Regressor	0.999	0.727
Simple Linear Regressor	0.605	0.306
Support Vector Regressor	0.563	0.557
MLP Regressor	0.362	0.351

In the given table 2, Mean Squared Error (MSE) and Root Mean Square Error (RMSE) of all the ML algorithms implemented in MakanSETU project is shown. Mean Squared Error is simply defined as the average of squared differences between the predicted output and the true output. Squared error is commonly used because it is unsure about whether the prediction was too high or too low, it just reports that the prediction was incorrect. And, Root Mean Squared Error is the square root of the MSE.

$$\text{MSE}\ (y_{\text{true}}, y_{\text{pred}}) = (1 / n_{\text{samples}})\ \Sigma\ (y_{\text{true}} - y_{\text{pred}})^2$$

Here Decision Tree, Random Forest and XGBoost Regressor show the least MSE and RMSE, respectively.

TABLE II. MEAN SQUARED ERROR & ROOT MEAN SQUARED ERROR

Model	MSE	RMSE
Decision Tree Regressor	27377	165
Random Forest Regressor	27802	166
XGBoost Regressor	28111	167
Simple Linear Regressor	47269	217
Support Vector Regressor	64382	253
MLP Regressor	92078	303

Lastly, Table 3 shows the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) of all the implemented algorithms in this project. Absolute error refers to the size of the difference between the prediction of an observation and its actual value. The magnitude of errors for the entire group of features into consideration is determined by MAE by averaging the absolute errors for a set of predictions and observations.

The mean of absolute relative errors is the definition of MAPE, a commonly used performance metric for regression:

$$\text{MAPE} = (1/n) * \Sigma\ (|\ actual - predicted\ |\ /\ |\ actual\ |) * 100$$

where, n is the total amount of predicted values generated by the regression model. Unlike other Models XGBoost regression shows the least Mean Absolute Error and Mean Absolute Percentage Error among all other regression models.

TABLE III. MEAN ABSOLUTE ERROR & MEAN ABSOLUTE PERCENMTAGE ERROR

Model	MAE	MAPE
Decision Tree Regressor	34.156	20.50
Random Forest Regressor	31.882	19.60
XGBoost Regressor	30.36	16.90
Simple Linear Regressor	99.314	98.60
Support Vector Regressor	76.369	41.00
MLP Regressor	154	95.50

After considering the above results XGBoost Regression algorithm is chosen to be the best predicting model with a decent accuracy and least error outputs. The chosen model is deployed using "Django MVT framework" as a Web App to cater the end users with a user-friendly Graphical User Interface to use the Model for predictions and validation. Figure 5, Figure 6, Figure 7 shows the GUI of the MakanSETU project.

The suggested system has all of the features of existing systems, but instead of solely dealing with non-spatial data set, it also works with geographical data which was not enabled during the research face as it required heavy finance to use Google Maps API.

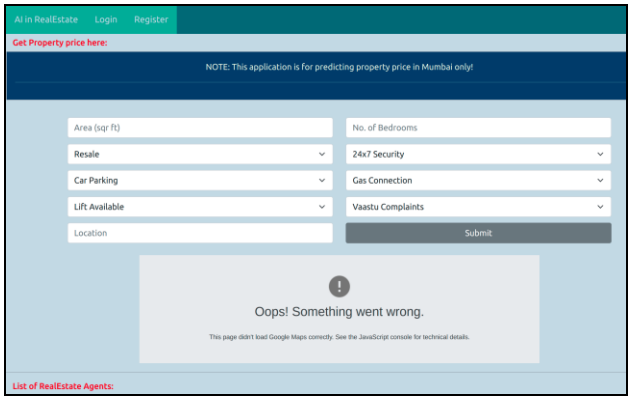

Fig. 5: User Home Page

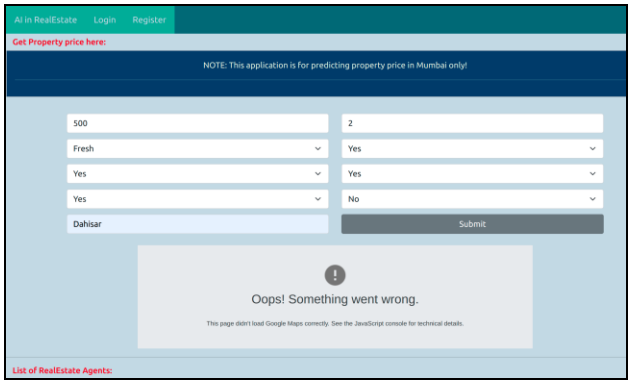

Fig. 6: User Home Page with Inputs

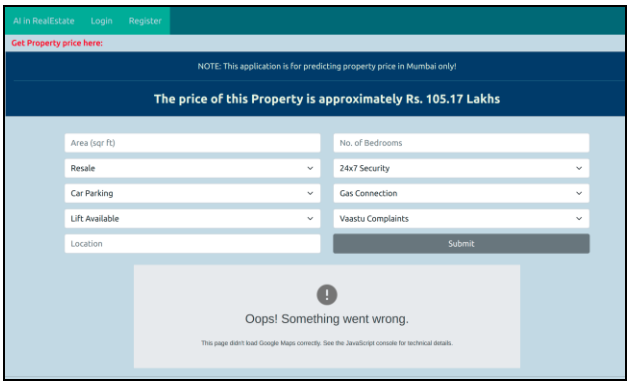

Fig. 7: User Home Page with output

User can Enter a location and MakanSETU will display the property price using the co-ordinates of the given location. The system will allow the user to quickly and conveniently search for a property to purchase or sell. A registered user can sell or rent his or her property. This mechanism protects clients and brokers against misleading marketing and data.

VI. Conclusion

After doing research on MakanSETU, it was discovered that Skyline AI [8] is the only form of AI employed in this industry, resulting in a significant reduction in the employment of brokers/agents, which is one of AI's most major drawbacks. The purpose/goal of this study is to enhance price prediction accuracy using traditional or advanced machine learning approaches on real-time data. It provides a platform for brokers/agents to display their most recent property prices, allowing buyers to choose the most appropriate and valuable source. AI is unavoidable in the real estate business due to its fast digitalization. Using predictive models, AI computers can identify market possibilities for agents to gain additional customers. Many various types of algorithms have been tested and implemented in MakanSETU, however a Non-Linear Tree-based Regression technique called XGBoost Regression has demonstrated to be the most accurate. MakanSETU has some Future work for other researchers in the Data Science and Artificial Intelligence domain. Finding better optimized form of the XGBoost model might get users more accurate results. Fraud detection in Real Estate is also an important topic, on which future researcher may work.

References

[1] Sayan Putatunda, “PropTech for Proactive Pricing of Houses in classified Advertisement in the Indian Real Estate Market”, Research Gate, 2019.

[2] A. Nur Alfiyatin, H. Taufiq, W.F. Mahmudy, Ruth Ema Febrita, “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization”, IJACSA, 2017.

[3] R Manjula, S Jain, S Srivastava and P.R. Kher, “Real estate value prediction using multivariate regression models”, IOP, 2017.

[4] Q. Truong, M. Nguyen, Hy Dang, Bo Mei, “Housing Price Prediction via Improved Machine Learning Techniques”, IIKI, 2019.

[5] A. Kuvalekar, S. Manchewar, S. Mahadik, S. Jawale, “House price forecasting using Machine Learning”, SSRN, 2020.

[6] Dr. G.N. Satish, Dr. V. Raghavendran, Mr. M.D. Suganan Rao, Dr. Ch. Srinivasulu, “House Price Prediction Using Machine Learning”, IJITEE, 2019.

[7] M. Modi, A. Sharma, Dr. P. Madhavan, “Applied Research On House Price Prediction Using Diverse Machine Learning Techniques”, IJSTR, 2020.

[8] R Bhatia, "Housing Prices in Metropolitan Areas of India", Kaggle, 2020, https://www.kaggle.com/ruchi798/housing-prices-eda-andprediction/data

[9] Shreyas, “House-price-prediction”, GitHub, 2019, https://github.com/Shreyas3108/house-price-prediction

[10] “Skyline AI”, 2017, https://www.skyline.ai/about/