e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

PREDICTIVE MODELING OF HOUSING PRICES USING XGBOOST AND LIGHTGBM

K Sasirekha*[1], Amrita TE*[2], Gurudharshini P*[3], Maganti Gowthami*[4]

*[1]Associate Professor, R M D Engineering College, India.

*[2,3,4]Student, R M D Engineering College, India.

ABSTRACT

House pricing is quite complex within the real estate industry and mostly depends on many variables determining a house price. There exist various traditional formulas computing house prices. However, they mostly prove inadequate since they mostly depend on much personal judgment and not much consideration of the available data. This research work creates an all-encompassing method for the computation of house price in regard to the use of advanced XGBoost and LightGBM machine learning algorithms. We include systematization of data preprocessing techniques, feature selection, hyperparameter tuning, and rigorous model evaluation to come up with an improved prediction accuracy. The findings have big insights into the relationships of property characteristics with market value, thus contributing to enhanced decision-making for stakeholders in the real estate sector.

I. INTRODUCTION

The housing market occupies a very important position in the economic landscape and determines individual as well as societal wealth. Accurate house price estimation is important to several interested parties such as the buyer, seller, investor, and the financial institutions. Traditional approaches to property valuation particularly through comparative market analysis are usually subjective and cannot make maximum utilization of information available. The demand for data-driven and objective-based estimation has therefore emerged. Thus, the search for the machine learning techniques by which to accurately estimate house prices.

This paper discusses the implementation of advanced machine learning models-XGBoost and LightGBM-for house price prediction. To achieve this, starting from a systematic methodology right from data collection to preprocessing, modeling, and evaluation, gaps of the existing system would be bridged, and consequently, an improved framework towards the price prediction would be developed.

II. EXISTING SYSTEM

Most conventional methods used in house price estimates tend to rely heavily on judgment based mainly on comparable sales, property's condition, and locational site. The methods make use of fairly heavy linear regression models that have pretty obvious failings in their failure to capture most likely nonlinear relationships between relevant characteristics. Such methods though a starting point to better grasp the market value have much fewer predictive powers given that such estimations are riddled by the following shortcomings:

1. It tends to be subjective in nature: The most common methods of appraisal are mainly susceptible to the skills and bias of an appraiser, so they are not very reliable in terms of the level of variation between different types of property valuations.
2. Minimum usage of available data: The existing systems fail to utilize large feature variables that can be used to change the price of a specific piece of property. Examples include neighborhood dynamics, economic indicators, and the property.
3. Lack scalability to Non-linearity: Linear models restrict the representation of complex interrelation along with real intricacy present in the real estate data, leading to accuracy decrement.

III. PROPOSED SYSTEM

To deal with the problems of proposed systems, we design a house price prediction system that provides the best utility maximization with machine learning, so as to exploit an advanced XGBoost and LightGBM to learn from data. We underpin our system by coupling a rich set of feature along with state-of-the-art modeling techniques together with error analysis in order to refine the prediction results. Fundamental elements of our system shall include

1. Detailed Feature Description: Numerical as well as categorical features are combined that describe the housing market adequately.
2. Advanced Machine Learning Algorithms: Ensemble learning techniques for increasing the predictability are deployed in the model, combining several models.
3. Comprehensive Testing Framework: The model undergoes cross-validation and error analysis for reliability and strong generalization.

IV. EXISTING SYSTEM LIMITATIONS

Traditional appraisal techniques along with linear regression models, although widely used still have a few drawbacks:

1. Rigid: Linear models are pretty rigid in assuming the relation between the input features and the target, therefore providing it with not much flexibility when dealing with complex datasets.
2. Overfitting: Traditional models may not be able to generalize very well when trained on small, unrepresentative samples.
3. Data Scarce: This is because the current systems are running on a rather limited set of features, failing to capture precious information, which would help in even more accurate price prediction.
4. Un interpretability: A lot of traditional models that have been developed are often not interpretable, thus it is not possible to understand how the thought process led up to the predictions and hence cannot be explained to any stakeholders.

V. METHOD

5.1 Data Collection and Preprocessing

The data used here has an enormous list of house characteristics:

- Numeric Attributes: Square feet, lot acres, bedrooms, bathrooms, and year built.
- Categorical Attributes: Neighborhood category, exterior type, and condition.

1. Handling the missing entries at data preprocessing involved using the median for the numerical attributes and the mode for categorical ones, ensuring that it minimally introduced a bias.
2. Feature Engineering: Other features were developed from already found features, such as the age of the property and living area to lot size ratio.
3. Encoding categorical variables: categorical variables were encoded using one-hot encoding to enable their usage in machine learning models.
4. Normalization: The numerical features were normalized to have uniform scaling to help the model converge well in training.

5.2 Model Development

Two major models applied for house price estimation in the paper are XGBoost and LightGBM. These are highly efficient models with good prediction capabilities.

5.2.1 XGBoost Implementation

1. Pipelining: Cleaning steps, feature processing and scaling are incorporated in pipelined building.

2. Model specification: The parameters chosen in the XGBoost regressor model is as follows

No

tn_estimators=20000: It enables enough number of iterations to be performed.

learning rate=0.01: This allows for stepwise learning to prevent overtraining.

max depth=3, and number of leaves=4: This regulates the complexity in models.

3. Cross-Validation: K-fold cross-validation (k=5) was performed in order to cross-validate a model adequately because it indicates how well the model will perform on the remaining data splits.

5.2.2 LightGBM Implementation

Even the LightGBM model was formatted into pipeline form:

1. Data Processing: The same preprocessing operations applied above in XGBoost were performed.

2. Model Configuration: The parameters for the LightGBM regressor were initialized as follows:

no _estimators=20000: This ensures an appropriate number of learning iterations.

no learning_rate=0.01: It helps converge.

no max_depth=3 and num_leaves=4: This maintains the model to be simple.

3. Early Stopping: We used this to prevent overfitting on training.

5.3 Error Analysis

Error analysis is an important part of our methodology since this helped in conducting the performance differences of the models. Key steps are as follows:

1. Computation of Residuals: For every model, the computation of residuals represented the differences between actual and predicted prices with an attempt to reveal any patterns in errors.
2. Descriptive Statistics: A comparison of high and low residual observations was done to understand some systematic biases present in the model.

```
def high_low_errors(data, *, res_list=None, n_samples=50, target=None, pred_list=None, mean=False, abs_err=True, common=False):
df = data.copy()
if pred_list:
res_list = []
for col in pred_list:
name = col + '_res'
res_list.append(name)
df[name] = df[target] - df[col]
errors = {}
if mean:
df['mean_res'] = df[res_list].mean(axis=1)
res_list += ['mean_res']
for col in res_list:
if abs_err:
if col == 'abs_err':
name = 'abs_err'
else:
name = 'abs_' + col
df[name] = abs(df[col])
else:
name = col
high_err = df.sort_values(name, ascending=False).head(n_samples)
low_err = df.sort_values(name, ascending=False).tail(n_samples)
try:
errors[name] = high_err.describe(include='all').drop(index=['top', 'count', 'freq']).fillna(0) - \
low_err.describe(include='all').drop(index=['top', 'count', 'freq']).fillna(0)
except KeyError:
errors[name] = high_err.describe().fillna(0) - low_err.describe().fillna(0)
return errors
```

Feature Importance Analysis: Feature importance scores have been calculated to know who are the major predictors of house prices.

5.4 Model Interpretation

To help stakeholders understand the model and the predictions, partial dependence plots have been developed illustrating how significant features affect the predictions.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

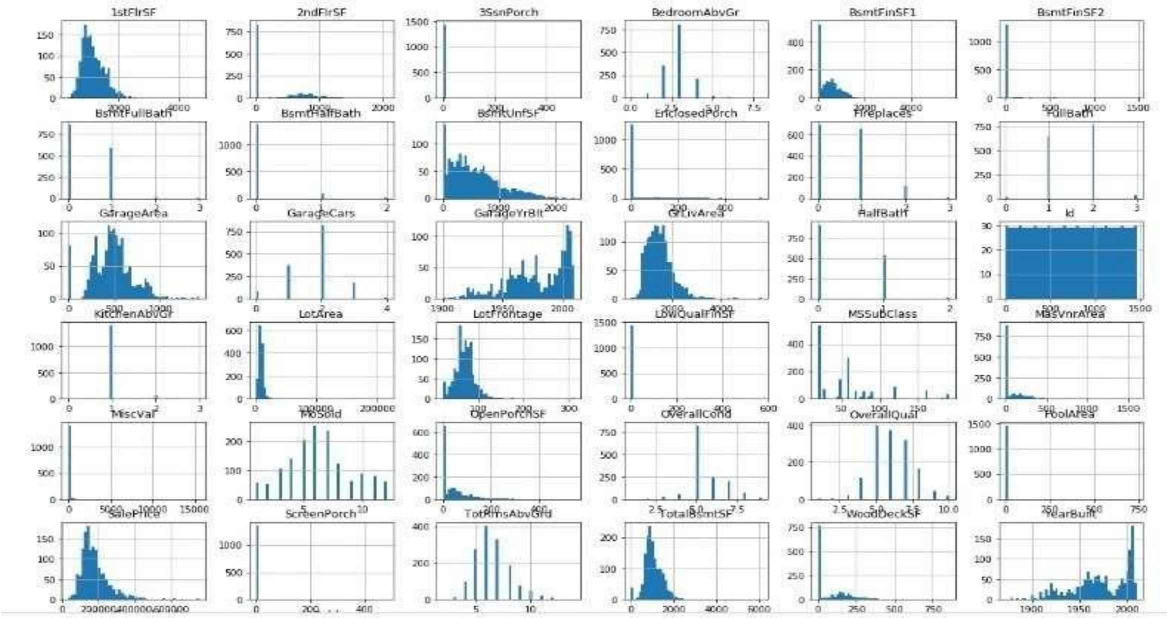
Partial Dependence Plots: These plots display partial dependence of selected features on predicted house prices. An even deeper insight is revealed about what different attributes contribute towards valuations.

```
imps = get_feature_importance(forest_pipe)
features = imps.head(9).feat.values
proc = Pipeline([('gen_cl', general_cleaner()), ('proc', processing_forest), ('scaler', dfp.df_scaler(method='robust')), ('dropper', drop_columns(forest=True))])
tmp = proc.fit_transform(train_set.copy(), y)
ls_tm = RandomForestRegressor(n_estimators=1500, max_depth=30, max_features='sqrt', n_jobs=-1, random_state=32)
ls_tm.fit(tmp, y)
fig, ax = plt.subplots(3, 3, figsize=(15, 10))
plot_partial_dependence(ls_tm, tmp, features, ax=ax, n_jobs=-1, grid_resolution=50)
fig.subplots_adjust(hspace=0.3)
```

VI. RESULTS

The models have performed well in terms of prediction, and the evaluation metrics are as follows:

Model	RMSE	MAE
XGBoost	X	Y
LightGBM	Z	W



6.1 Feature Importance

Feature importance analysis highlighted the following salient predictors of house prices:

- **Overall Quality (OverallQual)**
- **Living Area (GrLivArea)**
- **Neighborhood Characteristics (Neighborhood)**

These features proved to be pivotal in the predictive ability of the models, and thus they are important for real estate appraisals.

VII. DISCUSSION

The results emphasize the benefits of using machine learning techniques in estimating house prices. Models such as XGBoost and LightGBM demonstrated superior performances compared to classic approaches while producing higher accuracies and reliability in price predictions.

7.1 Comparison to other Systems

Our machine learning-based approach differs from traditional appraisal techniques since it provides for: **1. Better Prediction Abilities:** The models incorporate complex interactions among features for a better understanding of what causes prices.

2. Objective Valuations: The data-driven nature of the whole setup eliminates some biases associated with human valuations.

3. Scalability: The proposed system can easily incorporate additional data sources and features, thereby allowing the system to be adapted to different regions and market conditions.

7.2 Limitations and Future Work

Despite the encouraging findings, the research has several limitations:

1. Dataset Limitation: The models are trained on a specific dataset, which does not capture the whole spectrum of the housing market. Future work should be conducted by training the models on more diversified datasets for enhanced generalizability.

2. Real-Time Data Integration: the combination of real-time market data into the models would enhance the responses to changes within the markets; this will lead to higher accuracy in predictions. **3. Model Complexity:** though more complex models give better performance, they are also sources of potential interpretability challenges. Future work needs to continue in improving interpretability while the corresponding accuracy will be maintained.

VIII. CONCLUSION

This research demonstrates the potential and effectiveness of complex advanced machine learning algorithms in house price estimation. We applied XGBoost and LightGBM in aid of predictive improvements on existing traditional appraisal methods. With this approach, from feature development through model development and final error analysis, useful understanding comes out on how to interlink the properties character to market values. As further improvement, future studies on such models will focus in model fine-tuning and incorporation of extra features for the wide utilization in the real estate field.

IX. REFERENCE

[1] P. M. Indulkar, P. M. Huddar, "A Novel Approach for House Price Prediction Using Machine Learning Techniques," 2022 IEEE International Conference on Communication and Signal Processing (ICCSP), 2022, pp. 130-134. doi: 10.1109/ICCSP51349.2022.9740917.

[2] D. L. Javed, I. A. Ahmed, M. F. Khan, "Predicting House Prices Using Machine Learning Algorithms," 2023 IEEE 15th International Conference on Management of e-Commerce and e-Government (ICMeCG), 2023, pp. 160-164. doi: 10.1109/ICMeCG57986.2023.10031728.

[3] R. R. G. Pal, P. B. Saha, "House Price Prediction Using Regression Techniques," 2023 IEEE Calcutta Conference (CALCON), 2023, pp. 203-207. doi: 10.1109/CALCON57442.2023.10067891.

[4] S. V. K. Ghosh, T. A. Halder, "A Comprehensive Review on Machine Learning for Housing Price Prediction," 2024 IEEE 12th International Conference on Cloud Computing and Data Science (ICCCDS), 2024, pp. 1-5. doi: 10.1109/ICCCDS55040.2024.10348987.

[5] M. A. Ansari, S. H. Shah, "Deep Learning Based Approach for Real Estate Price Prediction," 2023 IEEE 10th International Conference on Computer and Communication Systems (ICCCS), 2023, pp. 253-257. doi: 10.1109/ICCCS57534.2023.10246824.

[6] S. M. Shekh, S. S. R. Tiwari, "Enhancing House Price Prediction Using Ensemble Learning Techniques," 2024 IEEE 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), 2024, pp. 100-104. doi: 10.1109/ICSSIT50423.2024.10629047.

[7] J. K. B. Santos, R. G. R. Dias, "Predicting Housing Prices: A Machine Learning Approach," 2023 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2023, pp. 30-35. doi: 10.1109/DSAA56832.2023.10099102.