House Price Prediction Models

Abstract

This report outlines the development of a machine learning model for predicting house sale prices, serving the real estate market by providing accurate property valuations. The project involves comprehensive data preprocessing and feature engineering, utilizing a blend of Linear Regression, Gradient Boosting Machines, and Decision Trees. Model evaluation incorporates metrics like MAE and R2. Anticipated challenges include data quality and model complexity.The project's overarching goal is to furnish a valuable tool for stakeholders, facilitating more informed decision-making processes within the dynamic real estate industry

1 Introduction

In the ever-evolving real estate landscape, informed decision-making stands as a pivotal factor for stakeholders seeking accuracy in property valuation. This report unfolds the comprehensive development of a predictive model tailored to estimate house sale prices, presenting a crucial tool for buyers, sellers, and real estate agents. Acknowledging the intricate interplay of variables influencing property values, this project embarks on harnessing the power of machine learning to refine the predictive accuracy of house prices.The dataset, rich with information encompassing location, size, age, and the number of bedrooms and bathrooms, forms the cornerstone for this endeavor. Through a systematic methodology, the report navigates the multifaceted process of data preprocessing, delving into cleaning, handling missing values, encoding categorical variables, and standardizing numerical features. Feature engineering is explored as a means to augment the model's predictive prowess by uncovering latent patterns within the data. A curated selection of machine learning algorithms, including Linear Regression for interpretability, Gradient Boosting Machines (gbm) for robust performance, and Decision Trees for modeling nonlinear relationships, forms the backbone of our predictive strategy. An ensemble method is strategically employed to synergize the strengths of these models, emphasizing a holistic approach to accuracy.As the report unfolds, we scrutinize anticipated challenges, emphasizing data quality, feature selection, and model complexity management to ensure the avoidance of overfitting while maintaining high predictive accuracy. The assessment metrics, encompassing Mean Absolute Error (MAE) and R-squared (R2), serve as benchmarks for evaluating the effectiveness of our predictive model.Ultimately, this project aspires to deliver not just a predictive model ensemble, but a comprehensive tool that empowers real estate stakeholders with nuanced insights, fostering a new era of data-driven decision-making in the dynamic and competitive real estate market.

2 Problem Statement

The central challenge addressed by our project revolves around the development of a predictive model that can precisely estimate house sale prices. Recognizing the complexity of this task, our objective is to delve into the analysis of historical sales data, aiming to discern and leverage patterns inherent in the relationships between various house characteristics and their corresponding selling prices. The overarching goal is to provide stakeholders in the real estate market with a powerful tool for precise property valuations, facilitating informed decision-making. This involves not only capturing the nuances of individual features such as location, size, age, and amenities but also understanding the intricate interplay between these factors in influencing property values. As the real estate market continues to evolve, the need for an accurate and adaptable predictive model becomes paramount for assisting stakeholders in navigating the dynamic landscape of property transactions.

3 Literature Survey

Every common man's first desire and need is for real estate property. Investing in real estate appears to be very profitable as the property rates do not fall steeply. Investing in real estate appears to be a difficult task for investors when one has to select a new house and predict the price with minimum difficulty. There are several factors which affect the price of a house and all these factors need to be taken into consideration to predict the price effectively. Also building such models for prediction needs much research and data analysis as many researchers are already working on it to get better results.

V. S. Rana, J. Mondal, A. Sharma and I. Kashyap 2020 [8] have used various regression algorithms to predict the house prices, like XG Boost, Decision Tree Regression, SVR, and Gradient Boosting Machines (gbm). After applying all these algorithms on to the dataset a comparison for the accuracy is done at the end. From which the maximum accuracy of 99

T. D. Phan, 2018 [2] is House Price Prediction using machine learning algorithms: A case study of Melbourne city, Australia. This is a thorough case study for analyzing the dataset to give some useful insights on the housing industry of Melbourne city in Australia. They have used various regression models. Starting with the data reduction to applying PCA (Principal Component Analysis) steps to get the optimal solution from the dataset. Then they have applied SVM (Support Vector Machine) for the competitive approach. Thus several methods are implemented to get the best results out of it.

M. Jain, P. Chawla, H. Rajput, and N. Garg 2020 [3] A house price forecast system that employs certain techniques. They have employed a straightforward machine learning procedure in this, which involves pre-processing, data cleaning, visualization, and k-fold cross validation for the output results. At last, they have presented a graph that closely resembles both the real and forecast prices, demonstrating a respectable level of accuracy based on their working model.

N.N. Ghosalkar and S. N. Dhage 2018 [5], Real Estate Price value using Linear Regression are using simple Linear Regression technique to give the price value for the houses.Through this paper they have tried to have best fitting line (relationship) between the factors of the real estate taken into consideration and used various mathematical techniques like MSE (Mean Squared Error), RMSE(Root Mean Squared Error) etc.

After reviewing various articles and research papers about machine learning for housing price prediction the article now focuses on understanding current trends in house prices and homeownership. The proposed system uses a machine learning model to predict prices with high accuracy.

4 Architecture

Linear regression, Decision Trees (DT), and Gradient Boosting Machines (GBM) are three fundamental algorithms widely used in the field of machine learning and statistics for predictive modeling. Each of these algorithms has its unique architecture and method of learning from data. Below is a detailed exploration of the architectures of these algorithms, suitable for inclusion in a report section.

4.1 Linear Regression

Linear regression is one of the simplest and most widely used statistical techniques for predictive modeling. It models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The basic form of linear regression is a linear equation that predicts a response variable as a function of one or more predictor variables.

Architecture:

- Equation: The linear equation can be represented as $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$, where Y is the dependent variable, X_i are the independent variables, β_i are the coefficients to be determined during the training process, and ϵ is the error term. - Parameter Estimation: The coefficients (β) are estimated using the least squares criterion, which aims to minimize the sum of the squared differences between observed and predicted values. - Model Evaluation: The goodness of fit of the model is typically evaluated using metrics such as R-squared and adjusted R-squared, which indicate the proportion of the variance in the dependent variable that is predictable from the independent variables.

4.2 Decision Trees (DT)

Decision Trees are a non-parametric supervised learning method used for classification and regression tasks. They model decisions and their possible consequences as a tree-like structure, where each internal node represents a ”test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (in classification) or a continuous value (in regression).

Architecture:

- Tree Construction: The tree is built by recursively splitting the training set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. - Splitting Criteria: The attribute for each split is selected based on a metric such as Gini impurity, entropy in classification problems, or variance reduction in regression problems. - Pruning: To prevent overfitting, the tree might be pruned by removing sections of the tree that provide little power in classifying instances.

4.3 Gradient Boosting Machines (GBM)

Gradient Boosting Machines are a powerful ensemble technique for regression and classification problems. They build a model in a stage-wise fashion like other boosting methods but generalize them by allowing optimization of an arbitrary differentiable loss function.

Architecture:

- Ensemble of Weak Learners: GBM constructs a forward stage-wise additive model; it builds the model in a step-by-step fashion, where each step involves adding a tree that best reduces the loss, given the current ensemble of trees. - Gradient Descent: The method uses gradient descent to minimize the loss when adding trees. At each stage, decision trees are fitted on the negative gradient of the loss function used in a classification or regression problem. - Regularization: GBM includes several regularization techniques, such as tree constraints, shrinkage (learning rate), and random sampling, to improve model performance and avoid overfitting.

Each of these algorithms has distinct features and is suited to different types of data and problems. Linear regression works well for linearly separable data, decision trees are useful for their interpretability and handling of non-linear relationships, while GBMs offer high performance at the cost of being more complex and less interpretable.

5 Result and Observation

A regression model’s effectiveness and performance can be evaluated based on a number of variables. Regression effectiveness of models cannot be assessed using metrics like F1 score, Precision, Recall, Accuracy scores, etc., unlike classification/clustering problems. This is due to the fact that, in contrast to classification difficulties, no data element’s precise future value can be predicted. The following metrics are what we’ll be using to assess the different regression algorithms that we tested using the provided dataset:

5.1 R-Squared

The R-Squared Score, also known as the R2 Score, indicates how closely the data points match the fitted regression/prediction line, or how accurately the line has been fitted to the provided dataset. R-squared is used in price prediction models to measure the proportion of the variance in the dependent variable (e.g., price) that is explained by the independent variables (e.g., predictors), indicating the goodness of fit of the model. It is the proportion of the overall variation that the model can account for. To put it another way, it’s a measurement of the distance between data points and the expected value. It makes sense that the accuracy of the regression model increases with its R2-Score value. A full correlation between the data points and the corresponding fitted regression value, which is almost impossible to obtain, is represented by a 100 on the R2-Score scale, which ranges from 0 to 100

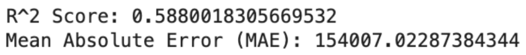
(MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{1}$$

R-Squared:

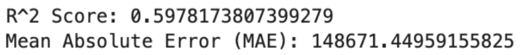
$$R - Squared = 1 - \left(\frac{SSR}{SST}\right) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

The precision attained for the Linear Regression



```
R^2 Score: 0.5880018305669532
Mean Absolute Error (MAE): 154007.02287384344
```

The precision attained for the Gradient Boosting Machine



```
R^2 Score: 0.5978173807399279
Mean Absolute Error (MAE): 148671.44959155825
```

5.2 Mean Absolute Error (MAE)

As the name implies, absolute error is just the amount of prediction error the model produces. Mean Absolute Error (MAE) is used in price prediction models to quantify the average magnitude of errors between predicted and actual prices, providing a straightforward measure of the model's accuracy in terms of absolute deviations from the true values.The difference between the actual value in the validation set and the value predicted by the model is known as the prediction error. It is possible for the error value to be positive or negative, indicating that the anticipated value may differ from the actual value. Nevertheless, whether an error is positive or negative, it still counts regardless of its indication. As a result, we calculate the absolute error and utilize it in further calculations. The result achieved is known as the mean absolute error (MAE), which is the average of all reported absolute errors.

6 Conclusion

In our investigation, we conducted a comparative analysis between Linear Regression and Gradient Boosting Machines (GBM) for predictive analysis, focusing on key metrics like R-squared and Mean Absolute Error (MAE). GBM exhibited an R-squared of 0.5978 with a corresponding MAE of 148671, while Linear Regression demonstrated an R-squared of 0.5880 and a MAE of 154007. Through meticulous hyper-parameter tuning, our aim was to minimize prediction errors further. Additionally, we explored a novel approach of combining predictions from both models through simple averaging, which not only enhanced predictive accuracy but also had the potential to mitigate risks associated with overfitting, contingent upon data quality.

To ensure sustained accuracy amid project dynamics, we underscored the importance of maintaining high-quality input data and regularly updating the models. Our methodology and findings hold promise for diverse domains, including real estate, where these regression models can be applied to pertinent datasets, thereby empowering stakeholders with valuable insights for decision-making processes.

By shedding light on the efficacy of different predictive models, our study equips businesses and organizations with the tools necessary to leverage predictive analytics for informed decision-making and strategic planning, thereby driving productivity and efficiency across various industries.

References

[1] Towards Data Science: Predicting House Prices with Linear Regression.
https://towardsdatascience.com/predicting-house-prices-with-linear-regression-machine-learning-i

[2] ResearchGate: Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia.
https://www.researchgate.net/publication/330476986_Housing_Price_Prediction_Using_Machine_Learning_Algorithms_The_Case_of_Melbourne_City_Australia

[3] IJIRT: [Insert Title Here].
https://ijirt.org/master/publishedpaper/IJIRT152722_PAPER.pdf

[4] IRJMETS: [Insert Title Here].
https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2023/37154/final/fin_irjmets1682710472.pdf

[5] IEEE Xplore: [Insert Title Here].
https://ieeexplore.ieee.org/document/8697639

[6] SSRN: [Insert Title Here].
https://deliverypdf.ssrn.com/delivery.php?ID=3531120081240020790930060801240050241160090290870844

[7] ResearchGate: Machine Learning Approach for House Price Prediction.
https://www.researchgate.net/publication/371602053_Machine_Learning_Approach_for_House_Price_Prediction

[8] ResearchGate: House Price Prediction Using Optimal Regression Techniques.
https://www.researchgate.net/publication/349802688_House_Price_Prediction_Using_Optimal_Regression_Techniques

[9] IJRASET: House Price Prediction Using Machine Learning Algorithms.
https://www.ijraset.com/research-paper/house-price-prediction-using-machine-learning-algorithms

[10] DIVA Portal: [Insert Title Here].
https://www.diva-portal.org/smash/get/diva2:1354741/FULLTEXT01.pdf

[11] Tandfonline: [Insert Title Here].
https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558

[12] Semantic Scholar: [Insert Title Here].
https://www.semanticscholar.org/reader/60165a1034e15dbe03f24a89b680854c23978e94

[13] IJITEE: [Insert Title Here].
https://www.ijitee.org/portfolio-item/c97410111322/

[14] Wiley Online Library: [Insert Title Here].
https://onlinelibrary.wiley.com/doi/10.1002/eng2.12599

[15] University of Oslo: Interpretable House Price Prediction Using a Collection of Local ML Models.
https://www.mn.uio.no/math/english/people/aca/anderdh/hjort-2022-interpretable-house-price-predi
pdf

[16] Towards Data Science: Understanding Gradient Boosting Machines.
https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

[17] Princeton University: [Insert Title Here].
https://klusowski.princeton.edu/sites/g/files/toruqf5901/files/documents/cart_adapt.pdf

[18] ResearchGate: Predictive Data Analysis Using Linear Regression and Random Forest.
https://www.researchgate.net/publication/365066122_Predictive_Data_Analysis_Using_Linear_Regression_and_Random_Forest

[19] ResearchGate: House Price Prediction using Random Forest Machine Learning Technique.
https://www.researchgate.net/publication/358351395_House_Price_Prediction_using_Random_Forest_Machine_Learning_Technique

[20] IRJMETS: [Insert Title Here].
https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2022/21465/final/fin_irjmets1650969155.pdf

[21] Course Hero: [Insert Title Here].
https://www.coursehero.com/file/106855139/6435-Article-Text-11858-1-10-20210515pdf/

[22] IJEAST: [Insert Title Here].
https://www.ijeast.com/papers/146-148,%20Tesma0804.pdf

[23] IJEAST: [Insert Title Here].
https://www.ijeast.com/papers/247-254,Tesma511,IJEAST.pdf

[24] IJITEE: [Insert Title Here].
https://www.ijitee.org/wp-content/uploads/papers/v11i3/C97410111322.pdf

[25] SlideShare: House Price Prediction Using Machine Learning.
https://www.slideshare.net/irjetjournal/house-price-prediction-using-machine-learning