Predicting the Housing Market with Machine Learning

Author: José E. Díaz Martínez
Advisor: Dr. Jeffrey Duffany
Electrical and Computer Engineering and Computer Science Department

Abstract

This capstone project focuses on building a machine learning model to predict housing market prices using a dataset that includes key housing features. Through exploratory data analysis (EDA), we identified patterns and correlations between features and housing prices. Multiple machine learning algorithms, including linear regression, SVR, Random Forest Regression, and CatBooster, were utilized to construct the predictive model. Performance evaluation using mean squared error highlighted SVR as the most accurate model. Hyperparameter tuning optimized the model's performance. Leveraging the reliable Ames, Iowa dataset from 2006 to 2011, the analysis provided insights into factors influencing property prices. The SVR model outperformed others with a mean squared error of 0.1870. Linear regression, Random Forest Regression, and SVR demonstrated similar performance, showcasing their effectiveness in predicting housing prices. The project offers valuable predictions and insights for future investments based on property prices in the next five years.

Introduction

The housing market serves as a critical pillar of the global economy, with accurate predictions of housing market prices becoming increasingly vital for various stakeholders. In this capstone project, the objective is to develop a robust machine learning model for estimating housing market prices based on a comprehensive set of factors. By leveraging machine learning algorithms, we aim to provide valuable insights and enhance decision-making in the real estate industry. Meticulous exploratory data analysis (EDA) techniques will be employed to uncover patterns and influential variables impacting housing prices. The machine learning algorithm will be carefully selected and fine-tuned to capture intricate relationships and generate accurate predictions. The primary goal is to deliver a reliable predictive model that aids real estate agents, assists investors, and empowers prospective homebuyers with valuable pricing information. This project contributes to advancing the real estate industry and improving economic decision-making through comprehensive data analysis, rigorous model development, and precise price predictions.

Background

The housing market is influenced by a myriad of factors, and accurately predicting housing prices remains a significant challenge. Machine learning techniques offer promising solutions in this domain. Support Vector Regression (SVR) stands out for its ability to capture complex relationships and non-linear patterns, surpassing Linear Regression, Random Forest, and CatBoost models. SVR's robustness to outliers further enhances its performance. Exploratory data analysis (EDA) plays a crucial role in gaining insights into the dataset, revealing patterns and relationships that influence housing prices. This project's motivation stems from a desire to understand the factors shaping the housing market and explore machine learning's potential as a predictive tool. Accurate price predictions empower real estate agents, investors, and homebuyers, fostering informed decision-making. Leveraging machine learning in the real estate industry is increasingly prevalent, enabling stakeholders to navigate the complex housing market landscape more effectively.

Problem

Accurately predicting housing prices in a complex market requires a reliable machine learning model. Ensuring accurate interpretation of results and informed decision-making further demands understanding underlying statistical principles. This project aims to overcome these challenges and develop a robust predictive model. By leveraging machine learning techniques, valuable insights and predictions for future investments based on housing prices can be provided. The goal is to empower stakeholders with accurate information, aiding in decision-making and enhancing understanding of the housing market dynamics.

Methodology

The methodology for this project comprises several key steps to ensure accurate predictions and reliable insights. The initial step involves collecting real estate data from various sources, such as public databases, real estate websites, or real estate agents. The collected data may contain errors, missing values, or inconsistencies, necessitating data cleaning and preprocessing to ensure accuracy, completeness, and consistency.

Exploratory data analysis (EDA) is then conducted to gain a deeper understanding of the dataset and identify patterns, trends, and outliers. Feature engineering techniques are applied to create new features from existing ones, enhancing the performance of the machine learning model. Once the data is preprocessed and feature engineering is complete, an appropriate machine learning algorithm is selected.

The selected model is trained using the preprocessed data, and its performance is evaluated using suitable metrics. The interpretation of the model's results helps understand how it makes predictions and identifies the most important features driving its predictions. Furthermore, the model is tested using the provided data, including an additional column for predicted values for upcoming years (e.g., 2024-2028). This step enables analysis and comparison of the model's predictions with actual sale prices, providing insights into its future performance and identifying potential issues.

Monitoring of the model's performance is essential, ensuring its continued effectiveness over time. Ongoing analysis and evaluation help identify any necessary adjustments or refinements to improve the model's accuracy and reliability.
Overall, this methodology encompasses data collection, cleaning and preprocessing, EDA, model selection and training, model evaluation and interpretation, as well as monitoring, to develop a robust predictive model that enhances understanding of the housing market and aids decision-making.

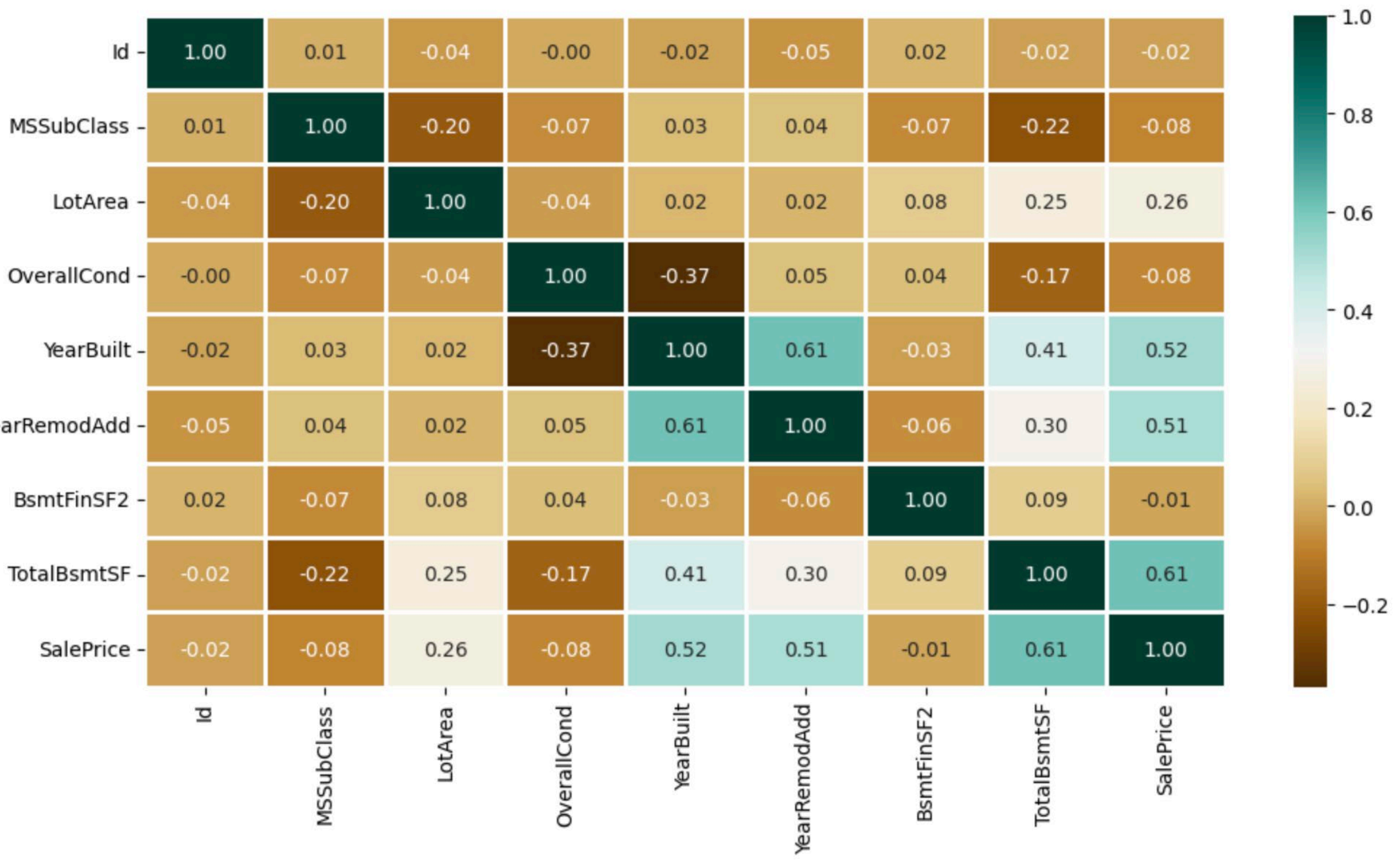


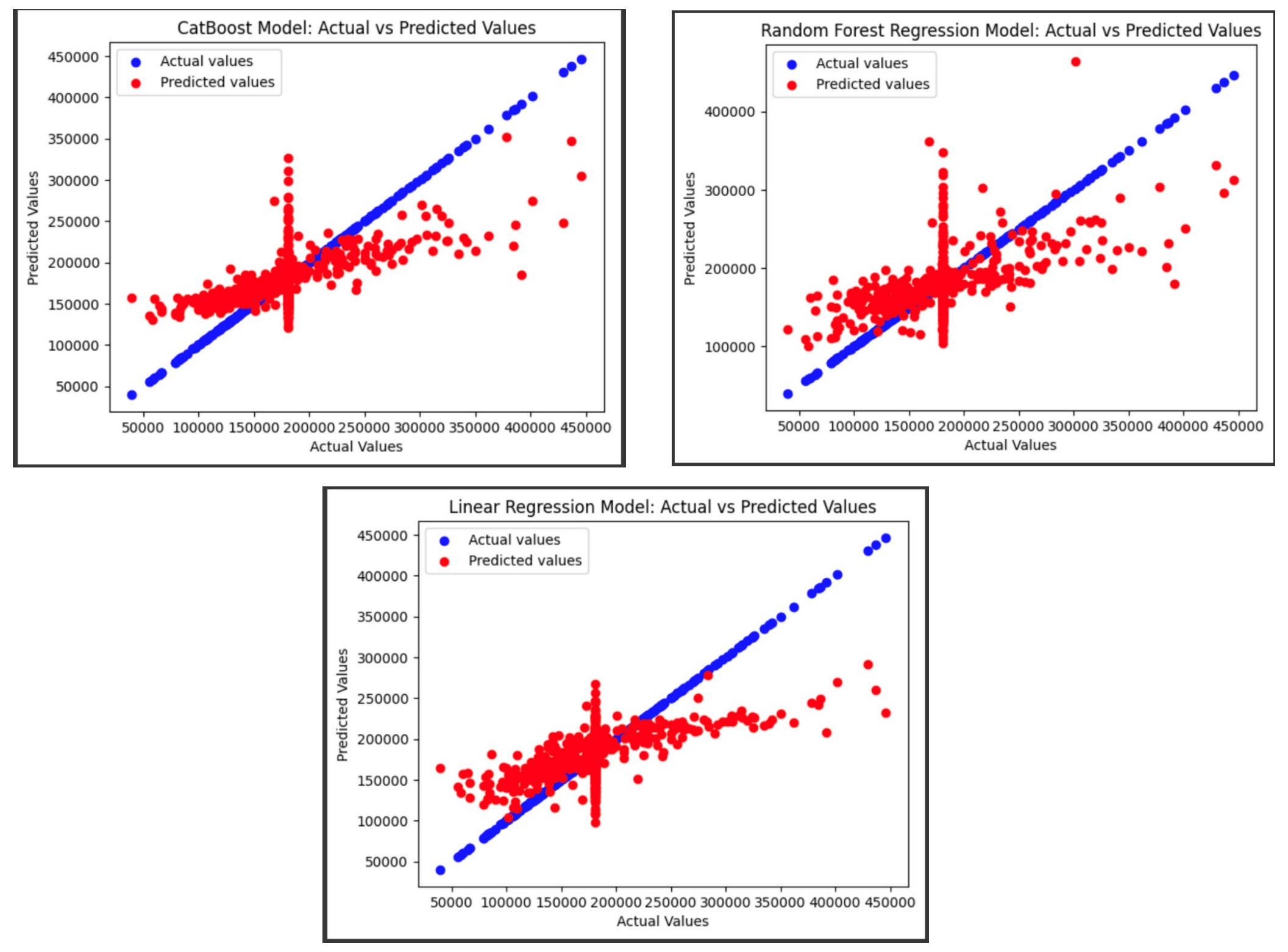
Figure 1: Presents a heat map used to visualize the intensity levels within the training data for the model. This heat map generates correlations between different variables in the dataset, using the original variables prior to data cleaning. Darker shades of green indicate higher correlations, while brown shades represent lower correlations. Correlation coefficients closer to 1 indicate a stronger positive relationship, whereas negative coefficients suggest an inverse relationship. A coefficient of 0 implies no correlation between the variables.

Results and Discussion

The predictions made provide valuable insights into the impact of general property price inflation relative to the year of construction. It is important to acknowledge that these predictions do not consider other factors that influence property prices, such as interest rates. Nevertheless, the results obtained using the SVR model indicate a strong correlation between the year of construction and the predicted sale prices. These findings shed light on the broader economic context and the effects of inflation on property values, underscoring the significance of historical construction periods in determining current market prices.

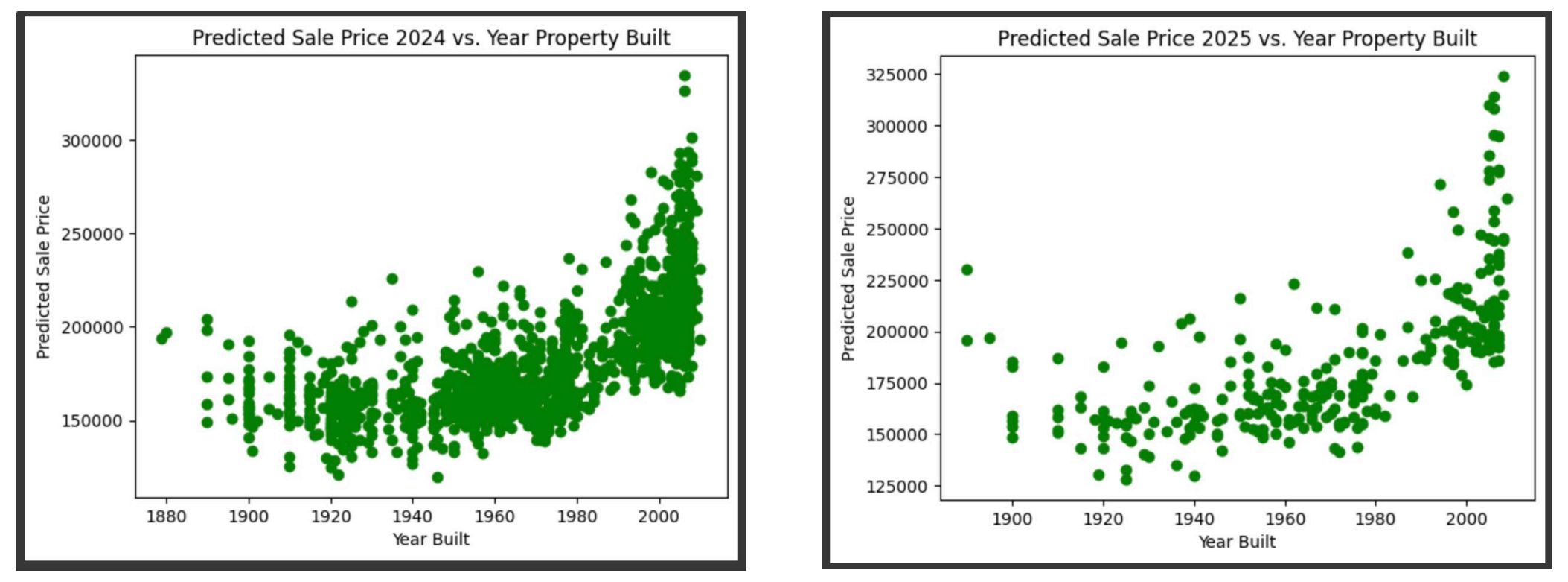
Method	Mean Squared Error Result
CatBoost Regression	0.40
Linear Regression	0.1874
SVR	0.1870
Random Forest Regressor	0.1903

Figure 2: Shows the result from the mean squared error comparison between each technique used from the clean training data. Its visible that SVR, Linear Regression and Random Forest Regressor perform very similar to each other.



Figures 3-5:

The scatter plot from figure 3- 5 visually compares the predictions made by different models: CatBoost Regressor, Random Forest Regressor, and Linear Regressor. The red dots represent the predictions from these models, while the blue dots represent the actual values. It evaluates the performance of each model and captures underlying relationships.



The figure 6-7 scatter plot predictions provide insights into the impact of property price inflation relative to the year of construction. The comparison of years in the scatter plots allows for an assessment of their predictive capabilities and sheds light on the influence of property price inflation in relation to the year of construction.

Conclusions

In this project, machine learning techniques were applied to predict housing market trends. Despite challenges in obtaining comprehensive data, valuable insights were gained from the available dataset. The evaluated models, including Linear Regression, Random Forest, and SVR, demonstrated similar performance in predicting housing prices. SVR excelled in capturing complex relationships, while Linear Regression offered simplicity and interpretability. Random Forest excelled in handling high-dimensional data. This project contributes to understanding the use of machine learning in housing market prediction, emphasizing the importance of data quality. Further research can advance predictive capabilities, benefiting stakeholders in the real estate industry.

Future Work

In future work, several avenues can be explored to enhance the project. Implementing a web scraping mechanism to extract data from real estate websites can provide a more comprehensive dataset. Continual data enrichment ensures the model remains up-to-date. Incorporating interest rates as a factor in the modeling process can improve accuracy. Exploring advanced machine learning techniques, such as deep learning or ensemble methods, may enhance predictive capabilities. These efforts aim to enhance the accuracy and robustness of the model and further advance understanding of the housing market dynamics.

Acknowledgements

I would like to express my heartfelt gratitude to Dr. Jeffrey Duffany for his invaluable guidance and support throughout this research project. His expertise and insightful feedback have been instrumental in shaping the direction of this study and encouraging me to pursue this research endeavor. I am truly grateful for his mentorship and the inspiration he has provided.

References

[1] U.S. Department of Housing and Urban Development. (n.d.). Real Estate Sales 2001-2018. Data.gov. Retrieved fromhttps://catalog.data.gov/dataset/real-estate-sales-2001- 2018

[2] Kaggle: House Prices: Advanced Regression Techniques. Retrieved from: https://www.kaggle.com/competitions/house-pricesadvanced-regression-techniques/overview

[3] De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. Journal of Statistics Education, 19(3). Retrieved from: https://jse.amstat.org/v19n3/decock/DataDocumentation.txt

[4] CatBoost Developers. (n.d.). CatBoost: CatBoost library repository. Retrieved from https://github.com/catboost/catboost