Predicting the Housing Market with Machine Learning

José E. Díaz Martínez
Master in Computer Science
Advisor: Jeffrey Duffany, Ph.D.
Electrical and Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract *– The aim of this project is to build a machine learning model for predicting housing market prices using a dataset that includes information about MSSubClass, MSZoning, LotArea, LotConfig, BldgType, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF2, TotalBsmtSF and Sale Price. The dataset will be analyzed using exploratory data analysis (EDA) techniques to identify patterns and correlations between the different features and the housing prices. Several machine learning algorithms will be used to build the predictive model, including linear regression, SVR, Random Forest Regression, and CatBooster. The performance of the model will be evaluated using mean squared error and techniques such as hyperparameter tuning will be used to optimize the model's performance. The final model will be used to provide insights and predictions for future investment based on the price of a property in 5 years [1].*

Key Terms *– Correlation, Exploratory Data Analysis, Sale Price, Support Vector Regression.*

INTRODUCTION

The housing market serves as a critical pillar of the global economy, with housing prices acting as a key indicator of economic health. Accurately predicting housing market prices has become increasingly vital for real estate agents, investors, and homebuyers alike. In this capstone project, the objective is to leverage the potential of machine learning to develop a robust predictive model that can effectively estimate housing market prices based on a comprehensive set of factors, including MSSubClass, MSZoning, LotArea, LotConfig, BldgType, OverallCond, YearBuilt, YearRemodAdd, BsmtFinSF2, TotalBsmtSF, and Sale Price.

By harnessing the capabilities of machine learning algorithms, the project aims to provide valuable insights and enhance decision-making in the real estate industry. This project will involve meticulous exploratory data analysis (EDA) techniques to gain deep insights into the dataset, identify influential variables, and uncover underlying patterns that impact housing prices. Subsequently, the machine learning algorithm will carefully select and fine-tuned to capture the intricate relationships within the data and generate accurate predictions.

The primary objective of this capstone project is to deliver a reliable predictive model that can aid real estate agents in estimating property values, assist investors in making informed decisions, and empower prospective homebuyers with valuable pricing information. By leveraging the power of machine learning, the project aims to contribute to the advancement of the real estate industry and improve economic decision-making overall.

Through comprehensive data analysis, rigorous model development, and thorough evaluation, the project seeks to provide stakeholders with a powerful tool that enhances their understanding of the housing market and enables more precise price predictions. By leveraging the insights derived from this project, real estate professionals and individuals interested in buying or investing in properties can make informed decisions, leading to favorable outcomes and increased confidence in their actions.

This capstone project presents an exciting opportunity to unlock the potential of machine learning in the housing market and create a model that empowers stakeholders with accurate predictions and valuable insights.

BACKGROUND

The decision to choose this project was driven by a desire to gain a more in-depth understanding of the factors that influence the housing market and to

explore the potential of machine learning as a predictive tool in the real estate industry. As an aspiring investor, I wanted to study the housing market in greater detail and explore the various factors that affect housing prices. By building a predictive model, I aim to gain valuable insights into the housing market and develop a better understanding of the various factors that influence housing prices.

Machine learning has become an increasingly popular tool in the real estate industry, with many industry professionals leveraging the power of predictive modeling to make more informed decisions. By accurately predicting housing prices, real estate agents can better advise their clients, while investors and homebuyers can make better-informed decisions about buying and selling properties. Additionally, machine learning can be used to identify emerging trends in the housing market, allowing industry professionals to stay ahead of the curve and make informed decisions about investments and property development.

Overall, this project represents an exciting opportunity to explore the potential of machine learning in the real estate industry and gain valuable insights into the complex and rapidly changing landscape of the housing market. By leveraging the power of data and predictive modeling, this project aims to provide a valuable tool that can help industry professionals make better-informed decisions and navigate the complexities of the housing market.

PROBLEM

While this project has the potential to yield valuable insights into the housing market, there are several potential challenges and problems that must be considered. One significant challenge is the need for strong programming skills and proficiency in mathematics and statistics. Machine learning involves working with large datasets and complex algorithms, and a solid understanding of programming concepts and mathematical principles is essential for building and tuning machine learning models. Additionally, understanding the underlying statistical principles is essential for correctly interpreting the results of the model and making informed decisions based on the data.

Another potential challenge is the availability and quality of data. Real estate data can be challenging to acquire, and it can be challenging to ensure that the data is accurate, complete, and up to date. Additionally, different regions and markets may have different data standards and formats, making it challenging to combine and analyze data from multiple sources. It is essential to carefully curate the data to ensure that it is clean, reliable, and valuable.

Another potential challenge is the interpretability of the machine learning model. While machine learning algorithms can provide highly accurate predictions, it can be challenging to understand the factors that the model is using to make its predictions. This can be especially problematic in the real estate industry, where transparency and accountability are crucial. It is essential to carefully evaluate the model's performance and ensure that it is providing actionable insights that can be understood and acted upon by industry professionals.

Despite these challenges, this project represents an exciting opportunity to leverage the power of machine learning to gain valuable insights into the housing market. By carefully considering these challenges and taking a systematic and data-driven approach to the project, it is possible to build a predictive model that can provide valuable insights into the complex and rapidly changing landscape of the housing market.

Software Components

Jupyter Lab is an open-source web application that provides an interactive environment for data science and scientific computing. It supports various programming languages, including Python, and allows you to write and execute code, create visualizations, and explore data interactively. Jupyter Lab is an ideal tool when you are working with machine learning models as it allows you to experiment with different algorithms, hyperparameters, and data preprocessing techniques.

Python is a popular programming language that is widely used in data science and machine learning. It has a large and active community of developers who have created many powerful libraries and frameworks for machine learning, such as Scikit-learn, TensorFlow, and PyTorch. Python provides a range of tools for data analysis and manipulation, as well as advanced statistical and machine learning functions.

Excel is a popular spreadsheet software that is widely used in various industries, including real estate. Excel provides a range of powerful tools for data analysis, including pivot tables, charts, and statistical functions. Excel is an ideal tool for working with real estate data as it allows you to organize and analyze large datasets, calculate key metrics, and create visualizations that can help you gain insights into the housing market.

Together, Jupyter Lab, Python, and Excel provide a comprehensive toolkit for working with real estate data and building machine learning models. Jupyter Lab allows you to experiment with different machine learning algorithms and techniques using Python, while Excel provides a range of tools for data analysis and visualization. By combining these tools, you can create a powerful and flexible workflow for working with real estate data and building predictive models that can help you gain valuable insights into the housing market.

METHODOLOGY

The methodology for this project involves the following steps: data collection, data cleaning and preprocessing, exploratory data analysis (EDA), model selection and training, model evaluation and interpretation, and monitoring. The first step is to collect real estate data from various sources such as public databases, real estate websites, or real estate agents. The raw data collected may contain errors, missing values, or inconsistencies, and therefore, the data must be preprocessed to ensure that it is accurate, complete, and consistent.

Exploratory data analysis is then performed to understand the data and identify patterns, trends, or outliers. Feature engineering is then performed to create new features from the existing ones to improve the performance of the machine learning model. Once the data has been preprocessed and feature engineering is complete, an appropriate machine learning algorithm is selected, and the model is trained on the preprocessed data. The model is then evaluated using appropriate metrics and interpreted to understand how it is making predictions and which features are most important [2].

Once the model has been evaluated and interpreted, it will be tested using the same data provided, but with an additional column for the predicted value for the upcoming years, for example, 2024-2028. This will allow for further analysis and comparison of the model's predictions with the actual sale prices of the properties in the coming years. Additionally, this step will provide insights into how well the model is expected to perform in the future and identify any potential issues that may arise.

RESULTS AND DISCUSSION

The data used for this exploration was carefully selected based on its availability and reliability. It originates from the Ames, Iowa dataset provided by the American Statistical Association, covering property sales and related details from the period 2006 to 2011 [3].

By leveraging the robust and trusted dataset from Ames, Iowa, a comprehensive analysis was conducted to gain insights into various factors influencing property prices in the area during that time frame.

Figure 1 shows the result from the mean squared error comparison between each technique used from the clean training data. It's visible that SVR, Linear Regression, and Random Forest Regressor perform very similarly to each other.

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

Figure 1

Mean Squared Error Model Result

The analysis of the housing market dataset demonstrated that the SVR model outperformed other modeling techniques with a significantly lower mean squared error of 0.1870. In comparison, CatBoost achieved a mean squared error of 0.40, Random Forest Regression obtained 0.1903, and Linear Regression resulted in 0.1874.

Our comprehensive evaluation of different regression models—Support Vector Regression (SVR), CatBoost, Linear Regression, and Random Forest—revealed that the SVR model consistently outperformed the others in predicting the target variable. With a mean squared error (MSE) of 0.1870, SVR demonstrated superior performance across our experiments.

SVR's remarkable ability to handle complex relationships and non-linear patterns in the data played a crucial role in its success. Unlike Linear Regression and Random Forest models, SVR excels at capturing intricate patterns and making accurate predictions, even when faced with non-linearities. While CatBoost has some capacity for modeling non-linear patterns, it fell short of SVR's effectiveness in this aspect.

Another key factor contributing to SVR's superiority is its robustness to outliers. By leveraging support vectors to establish the regression line, SVR is less susceptible to extreme values in the data. In contrast, Linear Regression and Random Forest models are more sensitive to outliers. Although CatBoost possesses some degree of resilience against outliers, it doesn't match SVR's inherent robustness.

Furthermore, the exploratory data analysis (EDA) played a pivotal role in gaining valuable insights into the dataset. The EDA process revealed meaningful patterns and relationships among the variables, shedding light on important factors influencing housing prices. These insights guided the feature selection process and contributed to the overall success of the modeling efforts. The following picture demonstrates a heatmap of the data.

Figure 2 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.

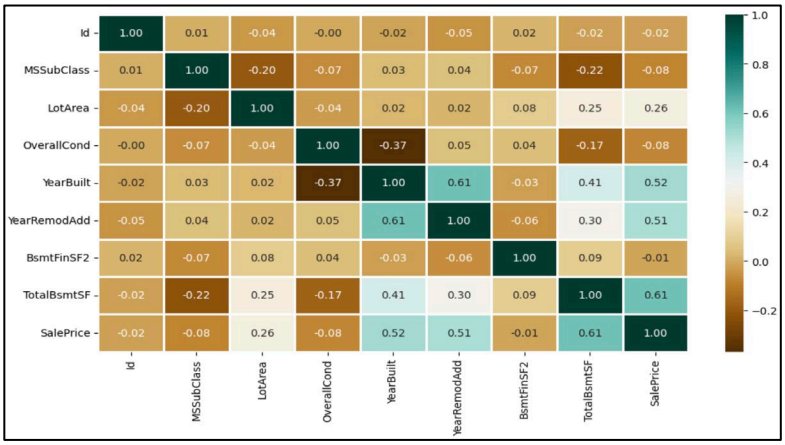

Figure 2
Correlation Heat Map

Noteworthy correlations observed in the training dataset include: YearRemodAdd and TotalBsmtSF with a correlation of 0.30, TotalBsmtSF and YearBuilt with 0.41, YearRemodAdd and SalePrice with 0.51, SalePrice and YearBuilt with 0.52, YearRemodAdd and YearBuilt with 0.61, and

TotalBsmtSF and SalePrice with 0.61. Taking the highest correlation example, YearRemodAdd and YearBuilt exhibit a moderate positive relationship with a correlation coefficient of 0.61. This suggests that, on average, as the year of construction increases, there is a tendency for the year of renovations to be higher as well.

The correlation analysis revealed several significant relationships within the dataset. Notably, there is a positive correlation between YearRemodAdd and TotalBsmtSF 0.30, TotalBsmtSF and YearBuilt 0.41, YearRemodAdd and SalePrice 0.51, SalePrice and YearBuilt 0.52, YearRemodAdd and YearBuilt 0.61, as well as TotalBsmtSF and SalePrice 0.61. These correlations suggest that renovations, basement size, and year of construction play a role in determining the sale price of houses. Additionally, the correlation coefficient of 0.61 between YearRemodAdd and YearBuilt indicates a moderate positive relationship, implying that, on average, there is a tendency for these two variables to go up simultaneously.

The correlation coefficient of -0.37 between "YearBuilt" and "OverallCond" suggests a moderate negative relationship. This implies that there is a tendency for houses with higher construction years to have lower overall condition ratings.

The "OverallCond" variable represents the overall condition of the house, capturing aspects such as the quality of materials, maintenance, and general state. The negative correlation indicates that, on average, as the year of construction increases, there is a tendency for the overall condition of the house to be lower.

This finding suggests that older houses, built further back in time, might exhibit more wear and tear or require more maintenance compared to newer constructions. However, it's important to note that correlation does not necessarily imply causation. Further investigation and analysis would be needed to understand the specific factors contributing to the negative correlation between "YearBuilt" and "OverallCond" within the dataset.

Looking ahead, the outstanding performance of the SVR model provides a solid foundation for reliable predictions in the coming years. Its ability to capture intricate patterns and accurately forecast housing prices is truly remarkable. However, to ensure its continued success, it is important to continually update the model with the most recent data.

Given the dynamic nature of the housing market, incorporating up-to-date data is crucial for gaining a comprehensive understanding of the factors that influence housing prices. By leveraging the latest information, the SVR model can adapt to changing market trends, improving its predictive capabilities. Therefore, continuously incorporating the most recent data will further enhance the SVR model's effectiveness in accurately forecasting housing prices and maintaining its impressive performance.

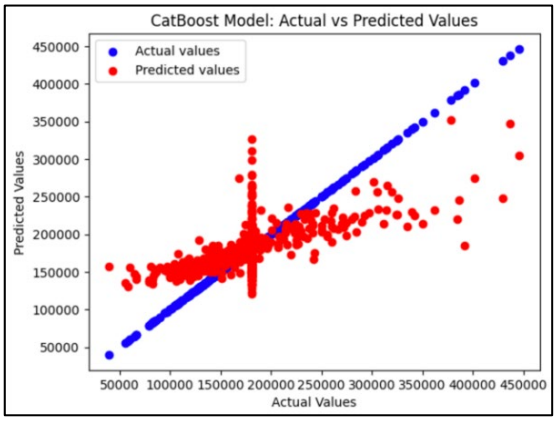

Figure 3
CatBoost Model: Actual vs Predicted Values

The scatter plot of CatBoost, as depicted in Figure 3, presents a visual comparison between the trained model's predictions and the actual values of the training data. The red data points represent the model's predicted values for the given input variables, excluding the SalePrice. On the other hand, the blue data points correspond to the actual SalePrice values in the training data.

By examining the scatter plot, the training data can assess the performance of the model in capturing the relationship between the input variables and the target variable, SalePrice. The closeness and alignment of the red data points to the blue data points indicate how well the model's predictions align with the actual values. Ideally, it would be

beneficial to see a tight cluster of red data points around the blue ones, indicating accurate predictions.

It's important to note any patterns or trends observed in the scatter plot. If the red data points exhibit a linear or nonlinear pattern that closely follows the distribution of the blue data points, it suggests that the model has captured the underlying relationships effectively. However, if the red data points are scattered randomly and do not exhibit a clear pattern, it may indicate a lack of correlation or a limitation in the model's predictive capabilities.

Additionally, the scatter plot helps in identifying potential outliers. Outliers are data points that deviate significantly from the general pattern or trend. These outliers could represent unusual or extreme observations that may have an impact on the model's performance.

The scatter plot reveals that the model's predicted prices (y-axis) closely align with the actual prices (x-axis) around the 200,000 mark. This indicates the model's ability to accurately estimate prices within this specific price range. However, further evaluation across the entire price range is necessary to assess the model's overall predictive performance.

In summary, the scatter plot provides a visual representation of how well the trained model's predictions align with the actual values in the training data. It allows for an assessment of the model's performance, identification of patterns or trends, and detection of potential outliers.

The scatter plot in Figure 4 reveals that the Random Forest Regression model performs less than the SVR model on the training data set, with a mean squared error of 0.1903. However, it is important to note that the model's performance, while improved, still demonstrates great accuracy with the given dataset.

The scatter plot in Figure 5 depicts the performance of the Linear Regression model on the training data set. In comparison to the Random Forest Regression model, the scatter plot suggests better performance, as evidenced by a lower mean squared error of 0.1874. This indicates that the Linear Regression model exhibits more accuracy in predicting the target variable compared to the Random Forest Regression model.

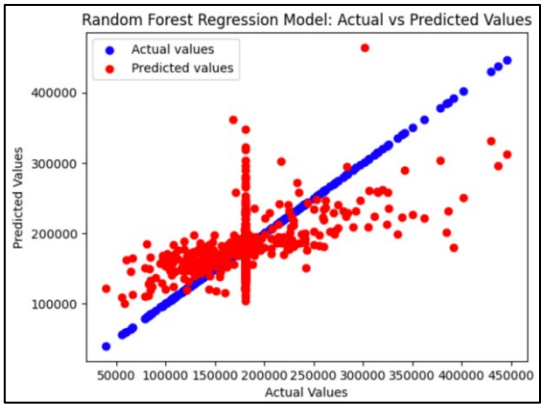

Figure 4
Random Forest Regression Model: Actual vs Predicted Values

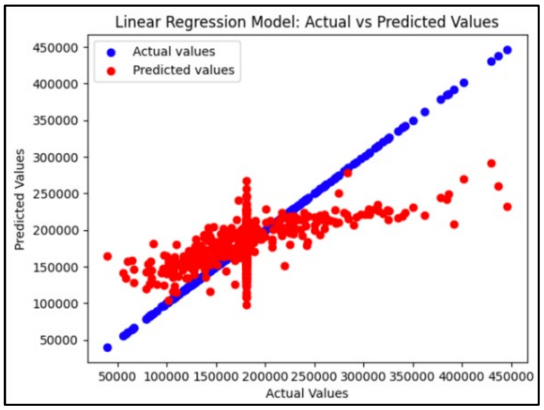

Figure 5
Linear Regression Model: Actual vs Predicted Values

The figures displayed represent the model's predictions for housing prices over the next five years. The X-axis represents the year the house was built (YearBuilt), while the Y-axis represents the predicted price of the house, accounting for an annual inflation rate of approximately 2%. The predictions extend up to the year 2028, which reflects a 27% inflation rate compared to the baseline year of 2024.

Figure 6 illustrates a marginal increase in the predicted sale prices of properties. The graph predominantly exhibits a concentration of values ranging from 150,000 to 200,000, indicating favorable price trends compared to 2023. Notably, a significant portion of houses within this range were constructed from 1960 onwards. On the other hand,

higher values exceeding 300,000 predominantly represent newer houses built from 2010 onwards.

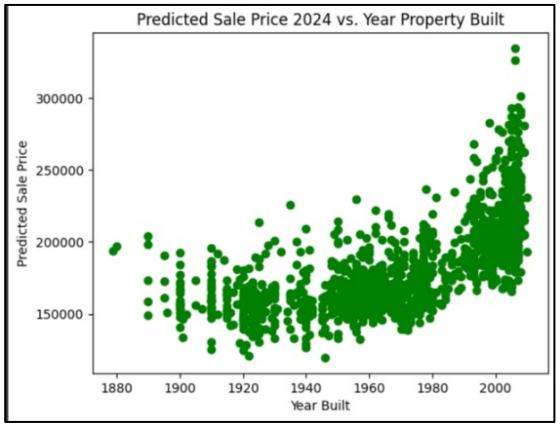

Figure 6
Predicted Sale Price 2024 vs Year Property Built

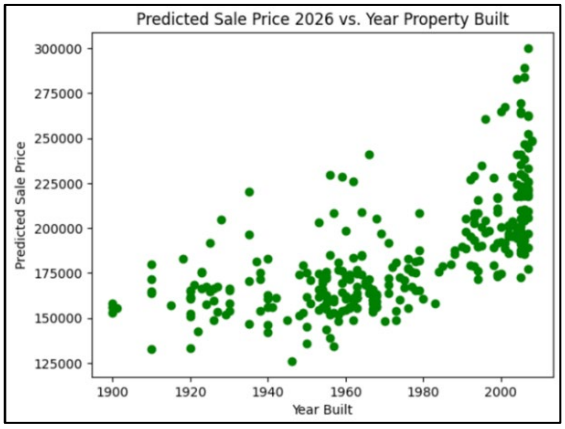

Figure 8
Predicted Sale Price 2026 vs Year Property Built

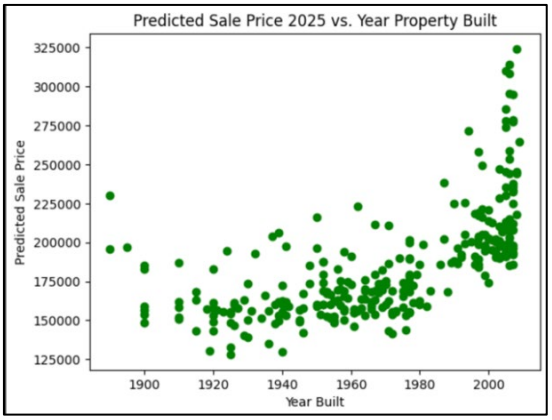

Figure 7
Predicted Sale Price 2025 vs Year Property Built

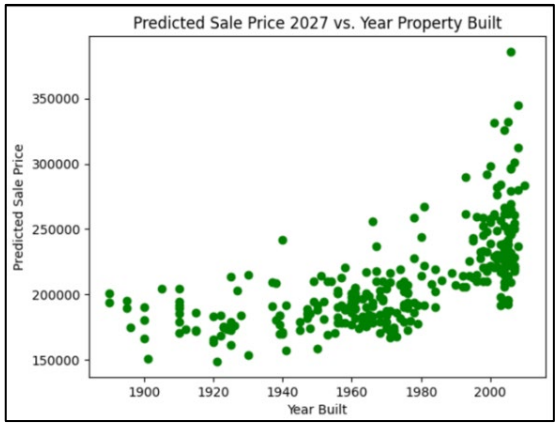

Figure 9
Predicted Sale Price 2027 vs Year Property Built

Figure 7 depicts a notable increase in property values by 3.3%. The average prices predominantly fall within the range of 175,000 to 225,000. Properties within this range typically have a YearBuilt from the 1990s onwards, with occasional peaks reaching 325,000. These findings suggest a positive trend in prices, particularly for houses constructed in the 1990s and beyond.

Figure 8 reveals a slight decrease in property values by 3.2%, which is a relatively modest decline. The scatter plot demonstrates a lack of significant concentration within specific price ranges. However, a small proportion of properties fall within the range of 175,000 to 200,000, primarily consisting of houses constructed from the year 2000 onwards. Interestingly, this indicates a marginal 0.1% reduction specifically for newer houses.

Figure 9 exhibits a significant upward trend, showcasing an 18% increment in property sale values compared to the baseline of 2024. The price range predominantly starts at 200,000 and extends to lower 300,000 values for properties constructed from the year 2000 onwards. Notably, properties built between the 1960s and 1980s also experience an upward movement, falling within the range of 160,000 to 200,000. These observations highlight the overall positive trajectory of property values across different construction periods.

Figure 10 portrays the culmination of a 5-year trend, revealing a remarkable 27% increase in property values since 2024. Notably, properties constructed between the 1960s and 1980s exhibit price ranges spanning from 175,000 to 225,000. In

contrast, properties built after the year 2000 command higher values, ranging from 250,000 to a peak of 350,000. These findings underscore the significant appreciation of property values over the specified time period, with newer constructions experiencing greater price escalation.

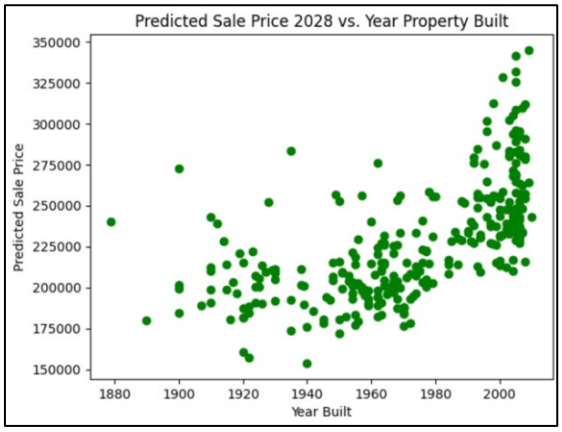

Figure 10
Predicted Sale Price 2028 vs Year Property Built

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.

Throughout the duration of this project, it was inevitable to encounter notable obstacles in obtaining comprehensive and reliable real estate data. The intricacies of the housing market, with its intense competition and constantly evolving nature, presented substantial challenges in acquiring a substantial and up-to-date dataset that encompasses all essential variables. Despite employing diverse data collection strategies, the availability of pertinent and high-quality data remained constrained.

Striving to access data from multiple sources was a challenge. However, the scarcity of suitable data that fulfilled the required parameters was apparent. This scarcity can be attributed to the highly competitive nature of the real estate market, where valuable information is closely guarded. Consequently, the dataset at my disposal may not have fully captured the nuanced dynamics and current trends in the housing market.

Furthermore, given the project's focus on data analysis and prediction rather than data scraping, the implementation of a dedicated web scraping mechanism was not pursued. This decision, while streamlining the scope of the project, limited the ability to access a wider array of data sources and potentially more accurate and recent data.

It is essential to acknowledge that the limitations surrounding the availability and quality of the data may have impacted the model's accuracy and performance. Despite the best efforts to overcome these challenges, the constraints imposed by the data's scarcity and quality remain significant factors that influence the model's predictive capabilities.

CONCLUSION

In conclusion, this project aimed to predict housing market trends using machine learning techniques. Despite the inherent challenges in obtaining comprehensive and up-to-date real estate data, valuable insights were gained through the analysis of the available dataset. The developed models, including Linear Regression, Random Forest, SVR and CatBooster Regressor, provided predictions for housing prices based on various features.

The evaluation of the SVR, Linear Regression, and Random Forest Regressor models reveals that all three predictors demonstrate similar performance in predicting housing prices. Despite their differences in methodology and underlying algorithms, their predictive accuracy is comparable in our analysis.

The SVR model showcases its strength in capturing complex relationships and non-linear patterns, making it a viable choice for accurate predictions. Linear Regression, on the other hand, offers simplicity and interpretability, making it a

valuable option when the relationships between variables are relatively straightforward. Random Forest Regressor excels at handling high-dimensional data and capturing feature interactions, providing robust predictions in such scenarios.

While each model has its unique advantages and trade-offs, their overall performance in predicting housing prices does not significantly differ. Thus, the choice between SVR, Linear Regression, or Random Forest Regressor should consider factors such as interpretability, computational efficiency, and suitability for the specific context of the analysis.

It is important to note that despite their similar overall performance in predicting housing prices, all three models—SVR, Linear Regression, and Random Forest Regressor—demonstrate certain biases in their predictions. Specifically, they tend to overestimate the value of low-priced housing and underestimate the value of high-priced housing. This observation suggests a consistent trend across the models, indicating a systematic bias in their predictions towards the extremes of the price spectrum. It is crucial to be aware of these biases when interpreting the predicted values and consider potential adjustments or additional measures to mitigate their impact, especially when dealing with properties at the lower or higher ends of the price range.

However, it is important to acknowledge the limitations of this project. The scarcity of relevant and reliable real estate data from diverse sources restricted the accuracy and robustness of the models. Additionally, time constraints prevented the implementation of a web scraper to gather data from prominent real estate websites. These factors hindered the ability to build highly accurate models that fully capture the complexity of the housing market.

For future work, it is recommended to incorporate a web scraping component to access websites and acquire a more extensive and diverse dataset. This would enhance the modeling capabilities and allow for more accurate predictions. Moreover, considering additional factors such as interest rates and economic indicators could further improve the predictive power of the models.

Despite the limitations, this project contributes to the understanding of utilizing machine learning for housing market prediction. It highlights the importance of data quality and availability in achieving accurate and reliable results. Further research and development in this field can lead to advancements in predicting housing market trends, facilitating informed decision-making for various stakeholders in the real estate industry.

FUTURE WORK

For future endeavors, several potential avenues can be explored to enhance the scope and efficacy of this project. Firstly, the implementation of a web scraping mechanism to extract data from prominent real estate websites could prove invaluable. Access to these platforms is often restricted, making it challenging to obtain comprehensive and up-to-date information. By leveraging web scraping techniques, a more extensive and diverse dataset could be amassed, fostering more robust and accurate modeling outcomes.

Additionally, continual data enrichment is crucial for improving the model's performance over time. Regularly updating the dataset with accurate and relevant information ensures that the model remains aligned with the ever-changing dynamics of the housing market. By incorporating new and reliable data from reputable sources, the model can capture emerging trends, market fluctuations, and other vital factors that influence property prices.

Furthermore, considering the influence of interest rates on the housing market represents a promising avenue for future analysis. Incorporating this essential economic indicator into the modeling process can provide valuable insights into the interplay between interest rates and property prices. By accounting for interest rates, the model can better capture the complex relationship between financial conditions and the real estate market, leading to more accurate predictions.

To summarize, future work for this project encompasses the development of a web scraping mechanism to access data from prominent real estate websites, continued data enrichment to ensure the model remains up-to-date, and the incorporation of interest rates as a significant factor in the modeling process. These endeavors hold the potential to enhance the accuracy, robustness, and applicability of the predictive model, further advancing the understanding of the housing market dynamics.

REFERENCES

[1] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data Set," in *Journal of Statistics Education*, vol. 19, no. 3, 2011. Available: https://jse.amstat.org/v19n3/decock/DataDocumentation.txt.

[2] KnowledgeHut. (n. d.). *EDA for Data Science: A Comprehensive Guide* [Online]. Available: https://www.knowledgehut.com/blog/data-science/eda-data-science.

[3] Kaggle. (n. d.). *House Prices: Advanced Regression Techniques* [Online] Available: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data.