Machine Learning Engine for Real Estate Price Estimation

Youssef Roman[1], Abdul-Rahman Mawlood-Yunis[1], [*,†]

[2]*Wilfred Laurier University, 75 University Ave W, Waterloo, ON N2L 3C5*

Abstract

Accurate price estimation is crucial for informed decision-making in the real estate industry. This study explores machine learning (ML) methods for predicting housing prices in Ontario's Halton Region from 2022 to 2023, using a dataset of over 7,000 detached home transactions. Data preprocessing involved feature engineering, including economic indicators like prime rates. Exploratory data analysis revealed transaction patterns and market shifts linked to interest rate changes. ML techniques, including linear regression, Random Forest, and XGBoost, were employed, with models achieving R-squared values between 0.93 and 0.997. Decision Tree and Random Forest models were the most effective in capturing price variability. Additionally, a Flask-based price estimation tool was developed, and trained on several regions of the Greater Toronto Area (GTA) allowing users to predict home prices based on specific property features. The study demonstrates the value of ML in enhancing real estate market efficiency by providing reliable price predictions, benefiting stakeholders such as homebuyers, sellers, and investors.

Keywords

Machine Learning, Multiple Listing Services (MLS), Feature Engineering, XGBoost.

1. Introduction

Given the close relationship between property values and economic conditions, appraising houses is an essential task for stakeholders such as developers, investors, homeowners, and appraisers. Informed decision-making in real estate transactions is made possible by accurate forecasts, which are also essential for investment planning and market stability.

Previous research has employed various machine learning (ML) techniques, such as Linear Regression [1, 2, 3], Random Forest [4, 2, 5], and Recurrent Neural Networks (RNNs) [6], to forecast housing prices, demonstrating the importance of feature engineering and the integration of diverse algorithms for improved accuracy. Approaches like linear regression and ensemble methods, such as XGBoost, have shown promise by incorporating factors like LSTAT score and crime rate per capita [5], while deep learning methods logistic regression, convolutional neural networks, and long short-term memory (LSTM) networks have been employed to predict prices by considering real estate characteristics and time-series data [7]. Time-dependent factors were further analyzed using Auto-Regressive and Moving Average (ARMA) models [7], highlighting the significance of temporal patterns in price prediction.

However, a lot of the current research ignores the significance of outside market factors, like shifts in prime rates, which are essential to comprehending variations in housing prices. By adding outside economic variables to the forecasting model, this study seeks to close that gap and provide a more thorough method of price prediction. To improve forecast accuracy and relevance, our model incorporates additional economic indicators, such as the Prime Rate of the Bank of Canada, in addition to property-specific attributes.

AISD-2024: Second International Workshop on Artificial Intelligence: Empowering Sustainable Development, October 2, 2024, co-located with the Second International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2024), Virtual Event, Lucknow, India.

roma0130@mylaurier.ca (Y. Roman); amawloodyunis@wlu.ca (A. Mawlood-Yunis)
https://scholar.google.ca/citations?user=pokpmw0AAAAJ&hl=en (A. Mawlood-Yunis)
0000-0002-1455-8844 (A. Mawlood-Yunis); 0009-0005-7645-9891 (Y. Roman)

The remainder of this article is structured as follows: Section 2 discusses the raw dataset and the engineered features derived from it. Section 3 presents the exploratory data analysis and key insights gained from the data. In Section 4, various ML models are developed, and their performance is evaluated using appropriate metrics. Finally, Section 5 summarizes the research findings and provides recommendations for future work.

2. Data Preparation and Preprocessing

The dataset employed in this study is derived from the REALM MLS Software, specifically targeting detached homes sold within the Halton Region for the years 2022 and 2023. The Halton Region, encompassing the cities of Oakville, Burlington, Milton, and Halton Hills, serves as the focal geographical area for this analysis. The dataset contains over 7,000 records, providing a substantial basis for a comprehensive examination of the real estate market trends and dynamics in this region during the specified period. Table 1 presents a brief description of the raw dataset. Each row in the table represents a home feature and its corresponding description obtained from the REALM MLS records. The additional Engineered Features such as Canadian Prime Rates and numerical PPSQFT are mentioned in section 2.2.

Table 1
Feature Description

Feature	Description
Address	This column represents the street address or location of the detached home within the Halton Region.
Beds	It indicates the number of bedrooms in the detached home.
Washrooms	Specifies the number of bathrooms (including full and half) in the detached home.
Type	Describes the type or style of the detached home, such as bungalow, two-story, or split-level.
Sold Price	Reflects the final selling price of the detached home at the time of sale.
SqFt	Represents the total square footage of the detached home, indicating its size or living area.
MLS#	Stands for Multiple Listing Service number, a unique identifier for the property within the REALM MLS Software system.
Sold Date	Indicates the date when the detached home was sold.
Approx Age	Represents the approximate age or year of construction of the detached home, providing insight into its construction period.

2.1. Data Cleaning

Extraction of the dataset into an Excel file marked the start of the data cleaning procedure. The data was then imported as a data frame using the Panda Library in the Google Colab environment, which made tasks like cleaning, processing, and visualization easier. Because data integrity was of the utmost importance, duplicate rows based on address had to be systematically removed from the dataset. This stage made sure that every property in the analysis was represented separately.

Furthermore, the 'Map Location' column, which was found to be mostly Null, was removed. The 'Sold Date' column underwent conversion to datetime format for ease of analysis and manipulation. This conversion facilitated chronological analysis and time-based visualizations.

Additionally, during the cleaning process, the 'Bedrooms' column was meticulously parsed to extract total bedroom counts, accommodating diverse entries such as '3+1' or '4+1', which were then converted into numeric values representing the total number of bedrooms. Similarly, the 'SQFT' (square footage) and 'Approx Age' columns, initially presented as ranges, were processed to create new numeric features: 'Average SQFT' and 'Age Numeric', respectively. These enhancements resulted in the creation of single numerical values, enhancing compatibility with machine learning algorithms.

These steps ensured that the dataset was prepared thoroughly, addressing issues of duplication, missing data, and format inconsistencies, thereby laying a solid foundation for subsequent exploratory data analysis (EDA) and modeling phases.

2.2. Feature Engineering

A brand-new function called "ppsqft" was released; it computes the price per square foot and offers insightful data on the dynamics of property pricing. Prime rates were also added to the dataset to make it more comprehensive for analysis. These prime rates, which provide vital context about the current state of the economy and the interest rate environment during the sale period, are expressed as a percentage of the Prime rate of the Central Bank of Canada at the time of sale. The dataset was greatly enhanced by this augmentation, which added more variables to take into account when analyzing the complex dynamics of the Halton Region real estate market over the given period of time.

Furthermore, the 'Walk Score' of the properties was added using the Redfin Walk Score API. The Walk Score ranges from 0 to 100, indicating the walkability of a location. The scores are categorized as follows:

- Car-Dependent: 0-50
- Somewhat Walkable: 50-70
- Very Walkable: 70-89
- Walker's Paradise: 90-100

Table 2
Enriched And Engineered Features

Feature	Description
Age Numeric	Represents the approximate age or year of construction of the detached home, provided as a numeric value for analysis.
Bedrooms	The total number of bedrooms in the detached home, including basement bedrooms, represented a numeric value.
ppsqft	Price per square foot, calculated as the selling price of the detached home divided by its total square footage, providing insight into pricing dynamics based on size.
Canada's Prime Rates	Percentage of the Central Bank of Canada's Prime rate at the time of sale, offering context on economic conditions during the sale period.
Walk Score	A score from 0 to 100 indicates the walkability of the property's location, influencing its attractiveness and value.
Walk Score Description	Describes the walkability score category, such as Car-Dependent, Somewhat Walkable, Very Walkable, or Walker's Paradise.

The engineered features and the features that were added to the property features to enrich the data set are described in Table 2 above. The walkability feature adds more information about how convenient and accessible the properties are, which makes the dataset more complete. A deeper comprehension of the relationship between walkability and property values and buyer appeal is made possible by the incorporation of Walk Score data. The dataset underwent several improvements to maximize its usefulness for machine learning models. Firstly, label encoding was employed for two critical features: 'House Type' and 'City'. This transformation facilitated the integration of categorical data into the machine learning algorithms. For the 'City' feature, a parsing algorithm was applied to the 'Address' column to extract city names, including Oakville, Milton, Burlington, and Halton Hills, populating a new column aptly named 'City'. Additionally, to standardize the representation of property sizes, an 'Sqft Numeric' column was generated by converting the range of square footage values into their average. Similarly, the 'Age' feature underwent transformation from age ranges to average age values, where 'New' properties were assigned a value of 0. One notable augmentation involved the inclusion of 'Prime Rates' as a new column, enriching the dataset with economic indicators known to influence housing prices. This comprehensive feature engineering process aimed to refine the dataset, empowering machine learning models with enhanced predictive capabilities by incorporating nuanced real estate factors and economic conditions.

3. Exploratory Data Analysis

The bar graphs in Fig.1 illustrate the number of homes sold in 2022 and 2023. Remarkably, the general pattern for the quantity of homes sold is the same in both years, with a significant percentage of sales concentrated in the first quarter, especially in February, March, and April. But it's important to note one important finding about how the dynamics of the market changed over the course of the two years. As a result of the 2022-2023 increases in Prime Rates and interest rates that exceed 6.45% in 2023, there is no discernible decrease in the volume of transactions. Specifically, there was only about a 3% decrease in the number of transactions recorded in 2023 compared to 2022. This observation suggests that there is no potential correlation between changes in interest rates and the real estate market activity, indicating the influence of financial factors on buyer behavior and market dynamics. Furthermore, Fig.2 below portrays the percentage change of quantity of homes sold in the 2022 months versus the 2023 months, indicating a spike hike in the months of June and October. A 40% increase in homes sold in June and a 30% increase in October is observed, likely due to Prime Rates not increasing significantly in these periods, with only a 0.25% increase in interest rates.

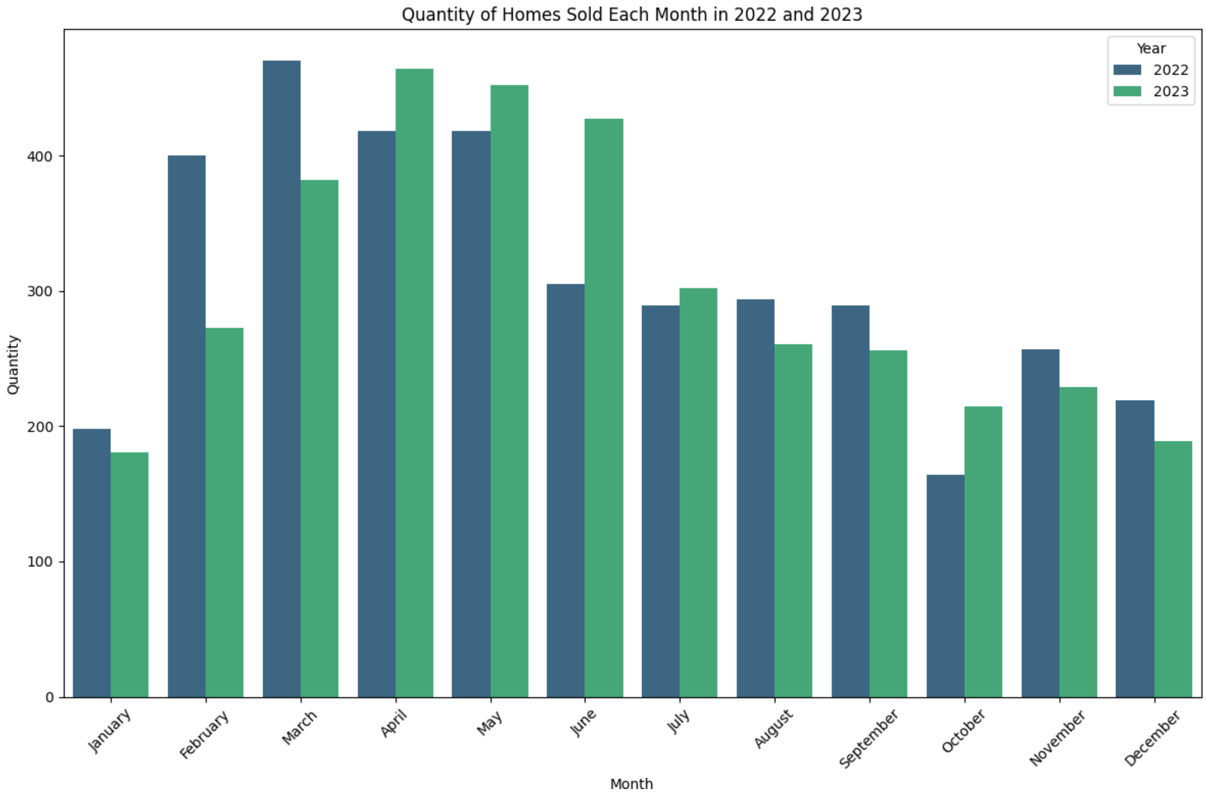

Figure 1: Quantity of Monthly Homes Sold in 2022 and 2023

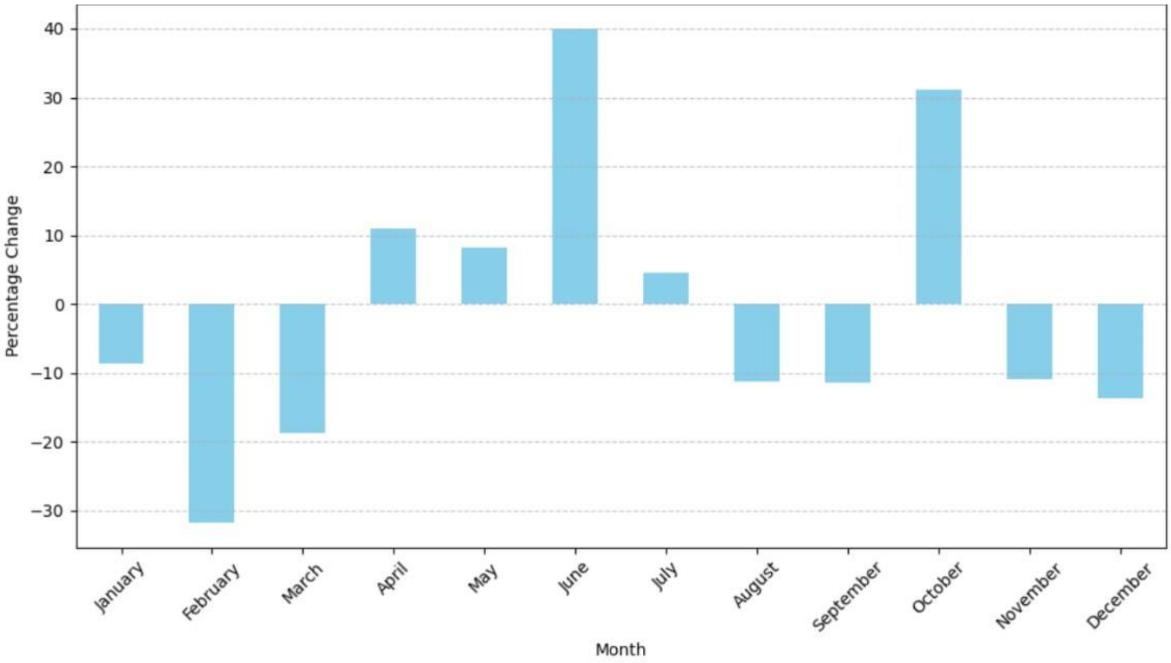

Figure 2: 2022 VS. 2023 Monthly Percentage Change of Real Estate Transactions

Moreover, the analysis extends to explore the broader trends in the real estate market through Fig.3. The histogram in Fig.3 illustrates an overarching downward trajectory in the number of homes sold, showcasing a decline of over 50% from January to December of 2022. Concurrently, the accompanying line plot in Figure 4 illustrates a discernible negative correlation between the escalation of prime rates, starting at 2.5% at the commencement of 2022 and reaching 6.5% by the year's end, and a corresponding 13.75% decrease in average home prices. This observation sheds light on the intricate interplay between macroeconomic factors and real estate dynamics, emphasizing the nuanced relationship between interest rates, market sentiment, and property valuations.

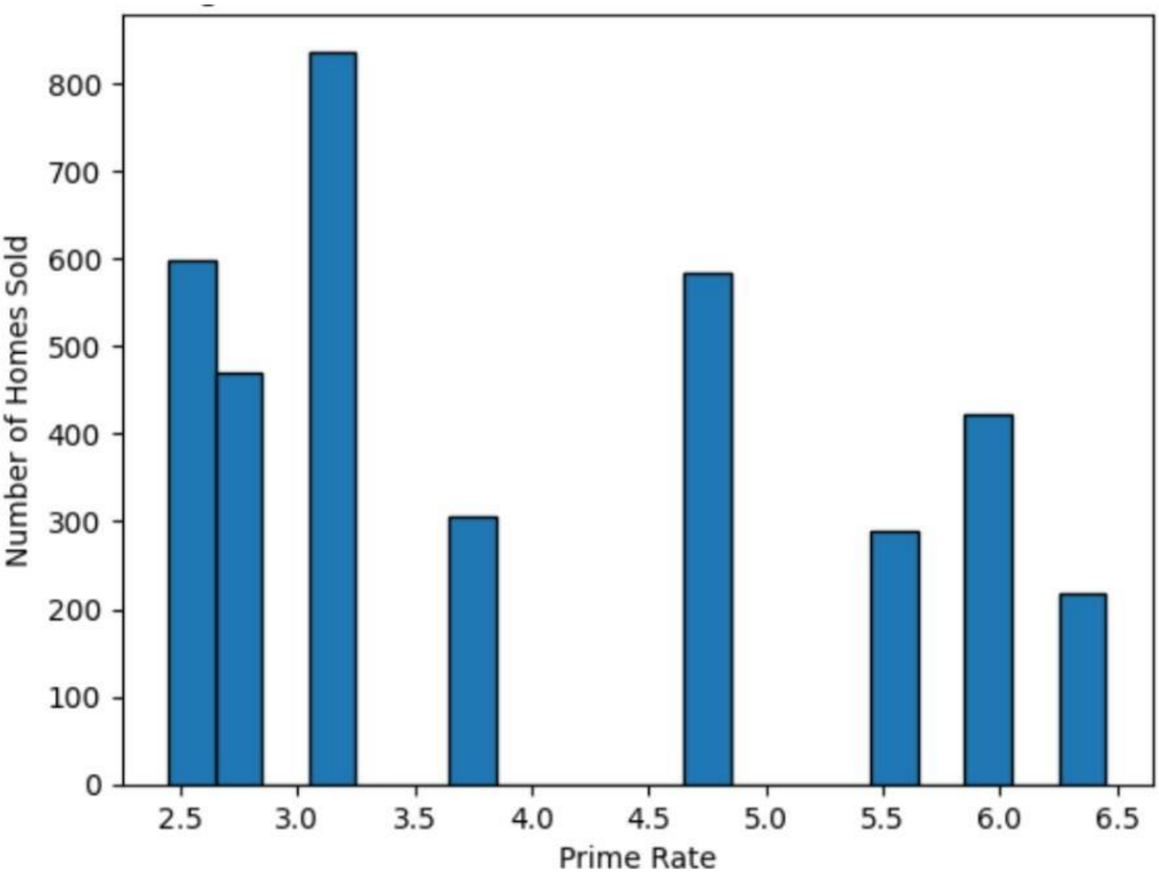

Figure 3: Number of Homes Sold VS. Prime Rate in 2022

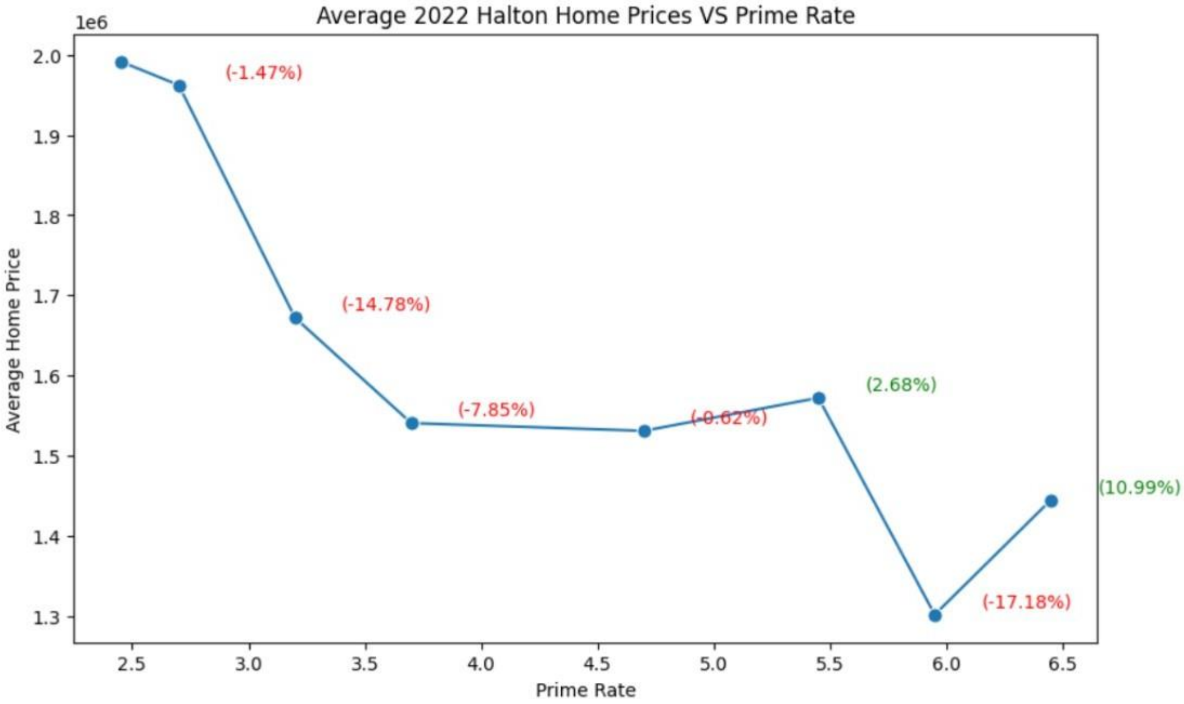

Figure 4: Percentage Trends of Number of Homes Sold VS. Prime Rate in 2022

Figure 5 below illustrates the Average Home Sold Price of 2022 and 2023, revealing a modest decrease of 3.94%. In tandem, Figure 6 showcases the Average Sold Price Per Square Foot of 2022 and 2023, displaying a reduction of -6.39%. These observations suggest that while the increase in interest rates may have exerted some minor influence on home prices, it appears to have been mitigated by other factors. Notably, Canada's low supply levels and the escalating construction costs attributed to heightened inflation and supply chain disruptions likely played significant roles in stabilizing home prices. As a result, despite the uptick in Prime Rates, the impact on the prices of detached homes in the Halton Region remained relatively subdued.

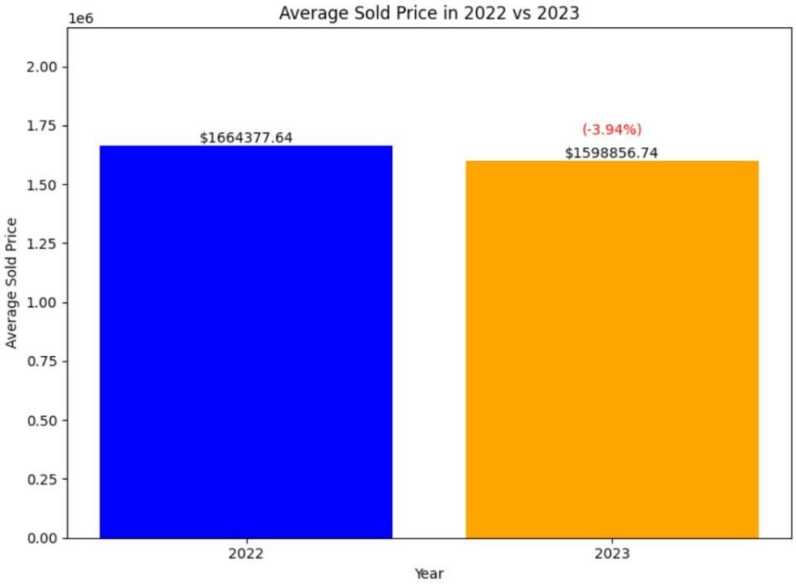

Figure 5: Average Home Price 2022 VS. 2023

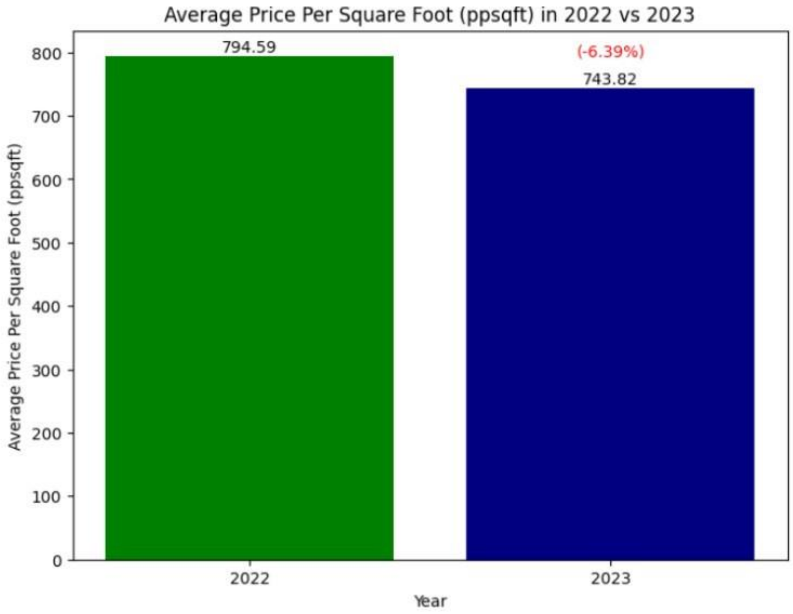

Figure 6: Average Price Per Square Foot 2022 VS. 2023

4. Price Prediction Model and Evaluation

Several models were built to predict housing prices during the ML Model Development phase, each with unique benefits and trade-offs. The correlation matrix is displayed in Figure 7 below, and it identifies the top four characteristics that have a significant correlation with the sold price: the number of bathrooms (0.54), total number of bedrooms (0.33), home square footage (0.84), and city in the Halton region (0.27).

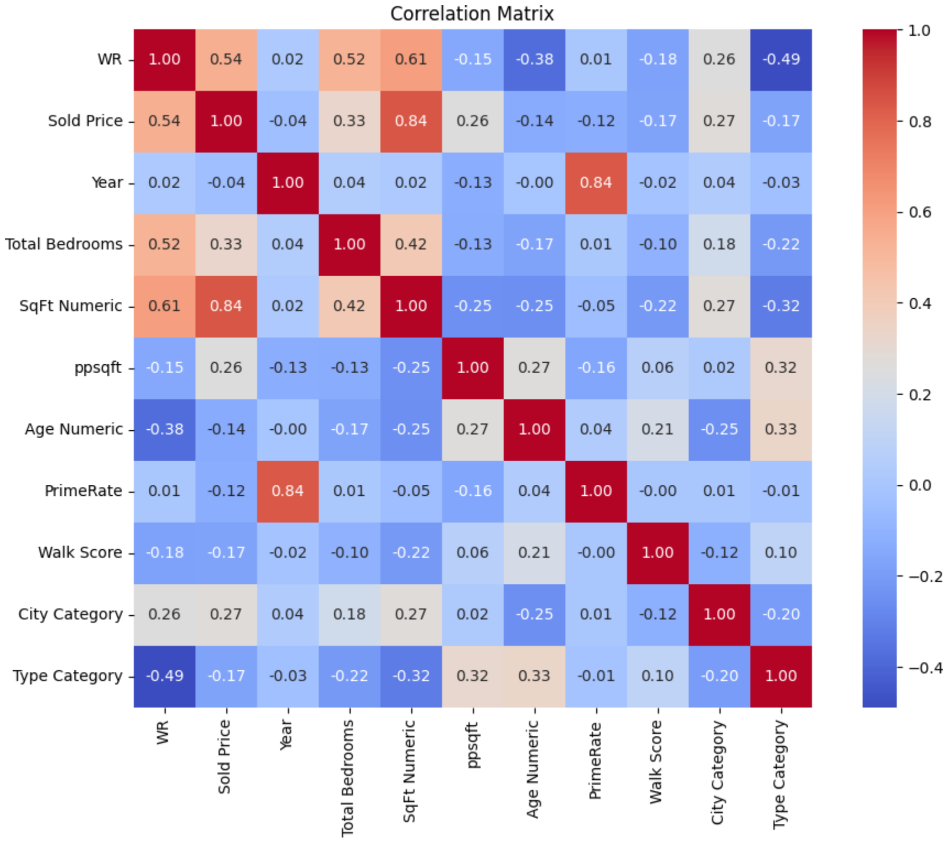

Figure 7: Sold Price Correlation Matrix

Table 3 provides an overview of Model Performance Metrics, highlighting key evaluation criteria such as Root Mean Squared Error (RMSE) and R-squared (Rš) values. These metrics are shown in Figures 8 and 9, illustrating the distribution of RMSE and Rš across different models. The models developed encompassed various methodologies, including linear regression, ensemble techniques, and tree-based algorithms. Linear Regression and Lasso Regression models were employed to establish baseline performance, showcasing strong Rš scores of approximately 0.93. These models offer simplicity and interpretability but may struggle with capturing complex non-linear relationships in the data. Decision Tree, Random Forest, and XGBoost models were subsequently explored to leverage more sophisticated modeling techniques. The Decision Tree model demonstrated exceptional performance with an Rš value nearing 0.99, suggesting an impressive ability to capture intricate patterns within the dataset. However, it's important to note that decision trees can be prone to overfitting, especially with deeper trees.

Table 3
Model Evaluation Metrics

Model Name	Root Mean Squared Error	R-squared
Linear Regression	205345	0.930948
Lasso Regression	205345	0.930948
Decision Tree	71050	0.991733
Support Vector Machine	746374	0.087736
Random Forest	42916	0.996984
XGBoost	202231	0.950595

The Support Vector Machine (SVM) exhibited some predictive capabilities, albeit with a lower R-squared value of 0.0877 compared to the higher values of 0.93 and 0.99 achieved by other models such as Decision Trees and Linear Regression. The ensemble learning technique Random Forest produced competitive results with a noteworthy Rš value of about 0.997. This model is a popular choice for regression tasks because it combines multiple decision trees to improve predictive accuracy and reduce overfitting. The gradient boosting algorithm XGBoost performed well, showing an R2 score of about 0.95. By iteratively improving upon weak learners, XGBoost excels at optimizing predictive performance and produces superior predictive accuracy.

Decision Tree and Random Forest models perform better than others in terms of Rš percentage, demonstrating their effectiveness in capturing the variance in housing prices. For the purpose of creating the online tool with Flask Web Service, Random Forest was determined to be the most accurate and suitable model. The tool's robust predictive capabilities allow users to estimate the price of their home based on property features like walk score, prime rate, square footage, number of beds, number of bathrooms, and home type.

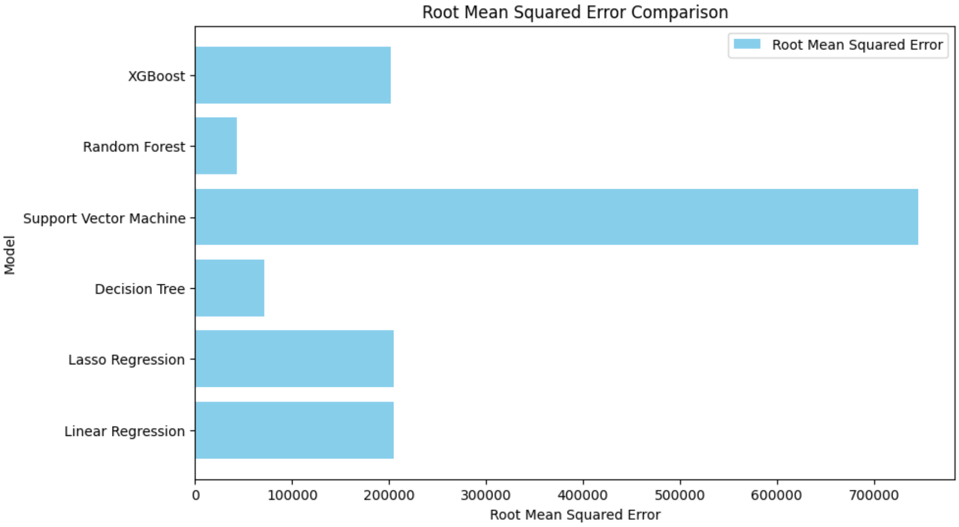

Figure 8: Root Mean Square Error Comparison

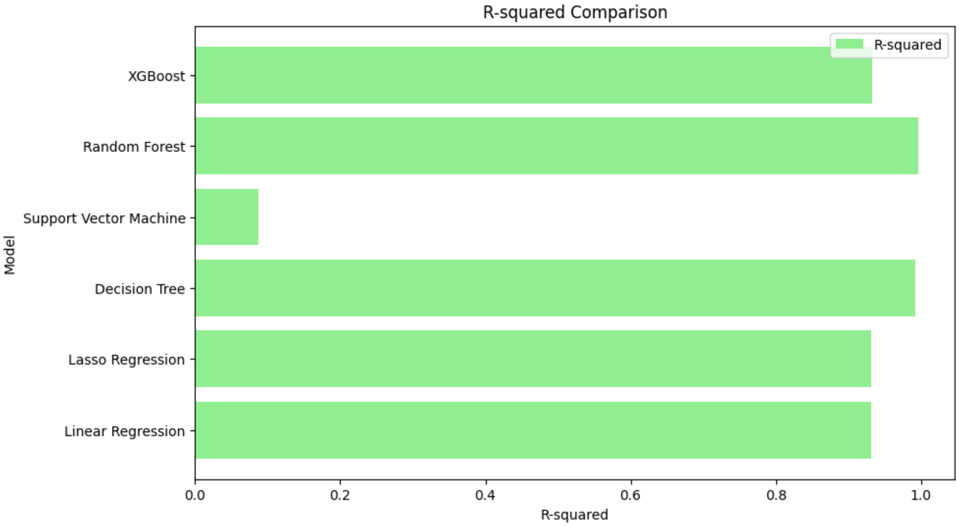

Figure 9: R Squared Comparison

5. Conclusion and Future Works

In conclusion, this study highlights the effectiveness of machine learning models, particularly Random Forest, in predicting real estate prices with remarkable accuracy, achieving R-squared values as high as 0.997. What distinguishes this work is the integration of both property-specific features and external economic indicators, such as prime rates, into the predictive models. This approach fills a critical gap in existing research by accounting for broader market conditions, resulting in more comprehensive and reliable price forecasts.

The novelty of this work lies in the incorporation of external economic factors, providing a more comprehensive approach to real estate price estimation. Future research could explore an application web interface which inherits the ml model. This allows stakeholders to integrate their own datasets for custom model training and price prediction.

References

[1] X. Li, Prediction and analysis of housing price based on the generalized linear regression model, Journal of Housing Research (2022). Received 14 July 2022; Revised 22 August 2022; Accepted 30 August 2022; Published 29 September 2022.

[2] U. Agarwal, S. K. Gupta, M. Goyal, House price prediction using linear regression in ml (2022).

[3] M. J. Chowhaan, D. Nitish, G. Akash, N. Sreevidya, S. Shaik, Machine learning approach for house price prediction, Asian Journal of Research in Computer Science 16 (2023). doi:10.9734/AJRCOS/2023/v16i2339.

[4] P. Furia, A. Khandare, Real estate price prediction using machine learning algorithms, Journal of Real Estate Research (2022).

[5] R. Konwar, A. Kakati, B. Das, D. B. Shah, M. K. Muchahari, House price prediction using machine learning, Journal of Real Estate Prediction and Analysis 9 (2021).

[6] Authors, Real-estate price prediction with deep neural network and principal component analysis (2022). doi:10.2478/otmcj-2022-0016, received: April 18, 2022; Accepted: November 10, 2022.

[7] S. Lu, Z. Li, Z. Qin, X. Yang, R. S. M. Goh, A hybrid regression technique for house prices prediction, Institute of High Performance Computing (IHPC), Agency for Science Technology and Research (A*STAR) (2020).