A Web Service Tool for Real Estate Price Estimation Powered by Machine Learning

Youssef Roman*, Abdul-Rahman Mawlood-Yunis†
Department of Computer Science and Physics, Wilfred Laurier University, Waterloo, Canada
e-mail: roma0130@mylaurier.ca*, amawloodyunis@wlu.ca†

***Abstract*—Accurate housing price estimation is critical for informed decision-making in the real estate industry. This study develops a Flask-based web platform that integrates machine learning (ML) algorithms to predict housing prices using a dataset of 7,000+ detached home transactions from Ontario's Halton Region (2022–2023). The system applies advanced ML techniques and feature engineering, incorporating economic indicators such as prime rates to enhance prediction accuracy. It enables real-time data exploration, visualization of transaction patterns, and analysis of market shifts driven by interest rate fluctuations. Linear Regression, Random Forest, and XGBoost were evaluated, achieving R² values between 0.93 and 0.997, with Random Forest demonstrating the best balance of predictive performance and overfitting resistance. The models are deployed via a user-friendly web application, allowing users to estimate home prices across the Greater Toronto Area (GTA) based on key property features. By leveraging ML, the tool enhances transparency and efficiency in the real estate market, providing homebuyers, sellers, and investors with a reliable, accessible, and data-driven valuation solution.**

***Keywords- Machine learning*; *Web services*; ; *Feature engineering*; *Real estate*; *House price estimation*.**

I. Introduction

Accurate house price estimation is essential for stakeholders, such as homeowners, developers, investors, and appraisers, as property values are closely tied to economic conditions. Reliable forecasts not only facilitate informed decision-making in real estate transactions but are also critical for ensuring market stability and guiding investment planning.

Previous research has explored various machine learning (ML) techniques, including Linear Regression [1]–[3], Random Forest [2], [4], [5], and Recurrent Neural Networks (RNNs) [6], to forecast housing prices. These studies underscore the importance of feature engineering and the use of diverse algorithms for improved accuracy [7]. For example, approaches like Linear Regression and ensemble methods such as XGBoost have demonstrated promise by incorporating factors like LSTAT (Lower Status of the Population) scores and crime rates per capita [5]. Deep learning methods—such as Logistic Regression, Convolutional Neural Networks, and Long Short-Term Memory (LSTM) networks—have also been employed to predict housing prices using both property characteristics and time-series data [8]. Additionally, time-dependent factors have been analyzed using Auto-Regressive and Moving Average (ARMA) models [8], further highlighting the importance of temporal patterns in price forecasting.

Despite these advances, much of the existing literature does not fully account for external market conditions, such as changes in *prime rate* which are critical for understanding housing price fluctuations. This study aims to address that gap by integrating external economic factors into the price forecasting model, providing a more comprehensive approach. By leveraging both property-specific attributes and broader economic indicators—such as the Bank of Canada's *prime rate*, *localized price per square Foot metrics*, and *walk score* ratings—our model enhances the accuracy and relevance of housing price predictions for the Canadian market.

This research focuses on developing a web-based application using Flask, a Python web framework, to provide real estate information for the Greater Toronto Area (GTA). The key aspects of this application are:

1) Flask Framework: The application is built using Flask, which is lightweight and flexible, making it ideal for creating web services quickly and efficiently.
2) Purpose: The web service aims to provide users with easy access to real estate insights, helping them make informed decisions about property purchases or investments.
3) Data Source: The application utilizes data from various regions within the GTA, ensuring comprehensive coverage of the local real estate market.
4) Functionality:
    - Users can input specific property details. The application provides real-time price estimates based on the input
    - It presents three comparable properties to aid in decision-making
5) Real-time Insights: By connecting property characteristics with current market trends, the application offers valuable, up-to-date information to users.
6) User-Friendly Interface: The platform is designed to be intuitive and easy to use, making it accessible to various real estate stakeholders.
7) Practical Application: The tool bridges the gap between research findings and practical real estate decision-making, allowing users to apply insights directly to their property-related choices.

This application demonstrates the practical application of data science and web development in the real estate sector, providing a valuable resource for both professionals and individuals interested in the GTA property market.

The remainder of the paper is organized as follows: In Section II, we discuss the Data Acquisition and Preprocessing. Section III presents the feature engineering process. Section IV details the ML models's results and evaluation, highlighting their predictive performance. Section V details the web-based application implementation, emphasizing its usability and design. SectionVI details the exploratory data analysis and insights. Finally, Section VII concludes the study and outlines

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-243-2

potential future work.

II. Data Acquisition and Prepossessing

The dataset employed in this study is derived from the REALM MLS Software [9], specifically targeting sold detached homes within the Halton Region for the years 2022 and 2023 [9]. 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 I presents a brief description of the dataset features. 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 Price per square foot (PPSQFT) are mentioned in section III.

TABLE I
Housing Features and their Descriptions

Feature	Description
Address	This column represents the street address or location of the detached home within the Halton Region.
Beds	Indicates the number of bedrooms in the detached home.
Washrooms	Specifies the number of bathrooms (including full and half) in the detached home.
Property 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.

III. Feature Engineering

This section outlines the enhancements made to the real estate dataset to improve insights into property pricing and market trends while optimizing its applicability for machine learning models. These enhancements include:

A. Feature Additions

The following new features are added to the dataset.

- Price Per Square Foot (psqft): Calculated by dividing the total property price by its square footage. Offers valuable insights into property pricing trends across different sizes and locations.
- Prime Rates: Reflects the percentage of the Central Bank of Canada's prime rate at the time of sale. Provides crucial context about the economic environment and interest rate conditions during each transaction.
- Walk Score: Added using the Redfin Walk Score API [10]. Measures property location walkability on a scale from 0 to 100. Higher scores indicate greater accessibility to amenities like shops, schools, parks, and public transportation. Categorized as follows:
  1) Car-Dependent: 0-50
  2) Somewhat Walkable: 50-70
  3) Very Walkable: 70-89
  4) Walker's Paradise: 90-100

B. Feature Transformation

Beyond incorporating new features into the dataset, the following data transformations and feature engineering techniques have been applied:

- Label Encoding: Applied to the 'House Type' and 'City' features to support the integration of categorical data into machine learning algorithms.
- City Name Extraction: A parsing algorithm is used on the 'Address' column to extract city names (Oakville, Milton, Burlington, Halton Hills) and store them in a new 'City' column.
- The 'Sqft Numeric' column is created by converting square footage ranges into average values. Additionally, for properties with missing square footage values, the mean square footage of the dataset is used as an imputed value to ensure completeness.
- Age Feature Transformation: Age ranges are converted into average age values, with 'New' properties assigned a value of 0.

These enhancements collectively contribute to a more robust and informative dataset, enabling more accurate and contextually informed analysis of the real estate market. Table II summarizes the enriched and engineered features, offering a comprehensive overview of the dataset improvements.

TABLE II
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	Total number of bedrooms in the detached home, including basement bedrooms, represented as 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 indicating the walkability of the property's location, influencing its attractiveness and value.

IV. Price Prediction and Model Evaluation

During the ML model development phase, multiple models were built to predict housing prices, each with unique advan-

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-243-2

tages and trade-offs. The models developed included Linear Regression, Lasso Regression, Decision Tree, Random Forest, XGBoost, and Support Vector Machine (SVM). These models represent a diverse range of machine learning techniques, from simple linear models to more complex ensemble and kernel-based methods. Model evaluations were based on a 20% test split, using Root Mean Squared Error (RMSE) and R-squared (R²) as key metrics.

Root Mean Squared Error (RMSE), measures the average magnitude of the prediction errors. It is the square root of the average squared differences between predicted and actual values. A lower RMSE indicates a model with smaller prediction errors. In this context, RMSE provides insight into the typical size of the errors made by the model in predicting housing prices.

R-squared (R²), on the other hand, measures the proportion of the variance in the target variable (housing prices) that is explained by the model. An R² value closer to 1 indicates a better fit, meaning the model explains most of the variance in the data. For example, an R² value of 0.93 suggests that the model explains 93% of the variation in housing prices.

A. Correlation Between Housing Features and Prices

The evaluation process began with an analysis of feature correlation to identify key factors influencing housing prices. Figure 1 presents the correlation matrix heatmap visually represents relationships between different features, using colors to indicate the strength and direction of correlations. It shows the four features most strongly associated with sold price: home square footage (0.84), number of washrooms (0.54), total bedrooms (0.33), and location within the Halton region (0.27).

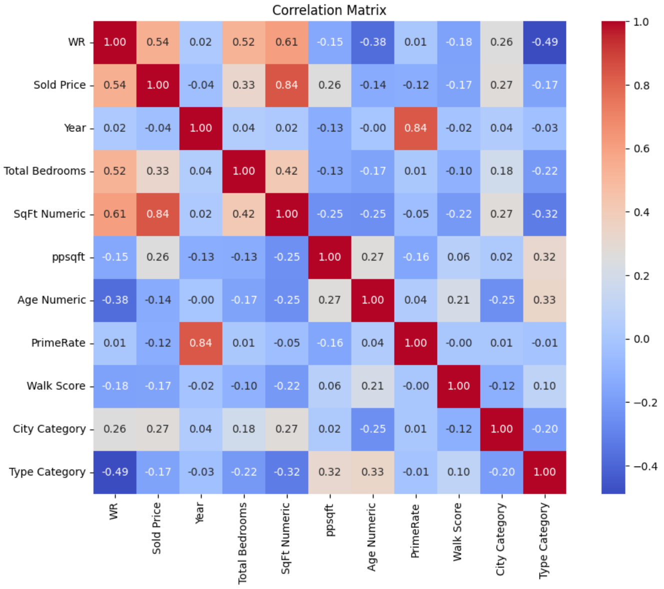


Figure 1. Correlation Matrix for Various Property Features

To understand more the results, a correlation matrix is a table that displays pairwise correlations, showing how strongly two features are related. A heatmap enhances this representation by using colors to illustrate correlation values, making it easier to identify patterns where rows and columns represent the same set of features. For example, "WR" correlates with "WR," "Sold Price" correlates with "Sold Price," and so on. The color bar on the right of figure 1 indicates the strength and direction of the correlation:

- Dark Red (1.0): **Perfect positive** correlation, meaning two features increase together.
- Light Red/Pink (closer to 0.0): **Weak positive** correlation.
- White (0.0): **No correlation**.
- Light Blue/Cyan (closer to 0.0): **Weak negative** correlation.
- Dark Blue (-1.0): **Perfect negative** correlation, meaning one feature decreases as the other increases.

Other findings include a moderately strong positive correlation (0.54) exists between the number of washrooms and sold price, suggesting that homes with more washrooms tend to have higher prices. Home square footage and sold price exhibit a very strong positive correlation (0.84), indicating that larger homes are generally more expensive. Conversely, a weak negative correlation (-0.17) is observed between sold price and Type Category, implying that certain property types may have slightly lower prices. Similarly, the weak negative correlation (-0.25) between home square footage and price per square foot suggests that larger homes tend to have a lower cost per square foot.

This correlation analysis provides valuable insights for identifying influential features and refining predictive models.

B. Model Performance Analysis

Multiple models were built to predict housing prices, including linear regression, ensemble techniques, and tree-based algorithms.

Linear Regression and Lasso Regression models were used to establish a benchmark for model performance, achieving strong R² scores of approximately 0.93. While these models are simple and interpretable, they may not effectively capture 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 R² 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 subject to overfitting, especially with deeper trees.

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.

Random Forest, an ensemble learning method, delivered competitive results with a notable R² value of approximately 0.997. This model combines multiple decision trees to mitigate overfitting and enhance predictive accuracy, making it a popular choice for regression tasks.

XGBoost, a gradient boosting algorithm, displayed strong performance with an R² score of around 0.95. XGBoost excels in optimizing predictive performance by iteratively improving upon weak learners, thus providing superior predictive accuracy.

Table III provides an overview of *model performance*, using Root Mean Squared Error (RMSE) and R-squared (R²) metrics. These metrics are shown in Figures 2 and 3, illustrating the distribution of RMSE and R² across different models.

TABLE III
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

C. Model for an Online Tool

Based on the results obtained, Random Forest was chosen as the most suitable and accurate model for the next phase of this research, which involves developing an online tool using a Flask web application. The tool enables users to input property features—such as Home Type, Number of Bedrooms, Number of Washrooms, Square Footage, Prime Rate, and Walk Score—to estimate their home price, leveraging the robust predictive capabilities of the Random Forest model.

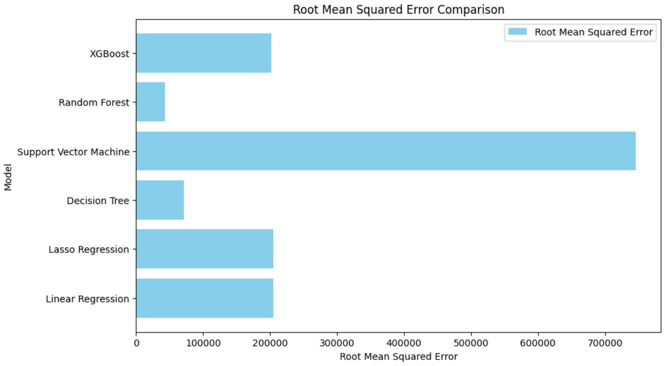


Figure 2. Root Mean Square Error Comparison

V. HOME PRICE ESTIMATION TOOL

The home price estimation tool is a web-based application designed to provide users with real-time predictions of house prices in the GTA, encompassing regions such as Halton, Peel, Toronto, Durham, and Hamilton. This tool is built on a machine learning framework, which dynamically predicts house prices based on user-provided property features, leveraging models tailored for each region to reflect the unique market dynamics and characteristics of that area.

The model development process begins with the collection and preprocessing of historical housing transaction data for each region. Feature engineering is applied to extract relevant attributes, including Home Type, Number of Bedrooms, Number of Washrooms, Square Feet, Prime Rate, and Walk Score, although the exact feature set varies by region. For instance, the Prime Rate is included in areas where interest rate fluctuations heavily influence house prices, while Walk Score may be excluded in regions where it is not a significant factor.

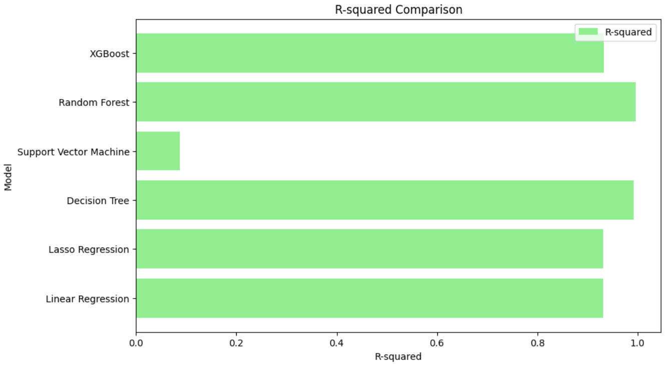


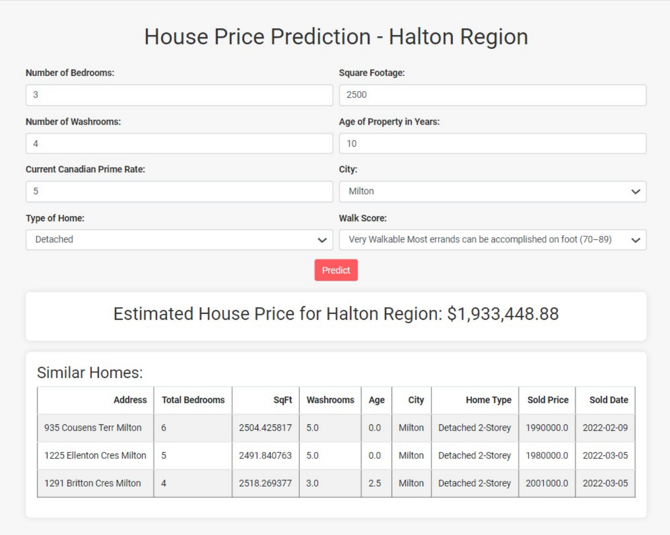
Figure 3. R squared Comparison

Each region's housing data is used to train a Random Forest model, implemented with the `scikit-learn` library. These models are optimized using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), ensuring high accuracy in predictions. Once the best-performing models are identified, they are serialized using the `joblib` library and saved as '.pkl' files. This serialization step enables the models to be stored and loaded efficiently within the web application, without the need for retraining.

The web application is built using the Flask framework, providing a user-friendly interface where users can input specific property features to estimate home prices. The process begins with users selecting their region and city, followed by entering property details such as Home Type, Number of Bedrooms, Number of Washrooms, Square Feet, Prime Rate, and Walk Score. When the form is submitted, the application loads the corresponding region-specific model from its serialized '.pkl' file. The user inputs are then processed, and the model predicts the house price in real time.

Once the price is predicted, it is displayed on the web interface along with up to three comparable property listings from the same city. These comparables are selected based on their similarity to the user-provided property details, using Euclidean distance for proximity, considering factors like location, home type, and number of bedrooms to provide additional context for the predicted price. Figure 4 shows an example of the tool's output for a home price prediction in Milton, located in the Halton Region, alongside comparable listings.

By leveraging pre-trained machine learning models, serialized in '.pkl' format, the home price estimation tool ensures fast, accurate predictions while minimizing computational



House Price Prediction - Halton Region

Number of Bedrooms: 3	Square Footage: 2500
Number of Washrooms: 4	Age of Property in Years: 10
Current Canadian Prime Rate: 5	City: Milton
Type of Home: Detached	Walk Score: Very Walkable Most errands can be accomplished on foot (70–89)

Estimated House Price for Halton Region: $1,933,448.88

Similar Homes:

Address	Total Bedrooms	SqFt	Washrooms	Age	City	Home Type	Sold Price	Sold Date
935 Cousens Terr Milton	6	2504.425817	5.0	0.0	Milton	Detached 2-Storey	1990000.0	2022-02-09
1225 Ellenton Cres Milton	5	2491.840763	5.0	0.0	Milton	Detached 2-Storey	1980000.0	2022-03-05
1291 Britton Cres Milton	4	2518.269377	3.0	2.5	Milton	Detached 2-Storey	2001000.0	2022-03-05

Figure 4. Home Price Tool UI

overhead. This integration of machine learning into a web-based environment offers a practical solution for real estate price estimation, making advanced predictive models accessible to users in real-time. The ML powered Architecture diagram can be shown in figure 5 below. The entire project, including data preprocessing, model training, and the Flask web application, is available on GitHub at https://github.com/Y-Roman/HousePricePredictionModel/tree/master for further exploration and use.

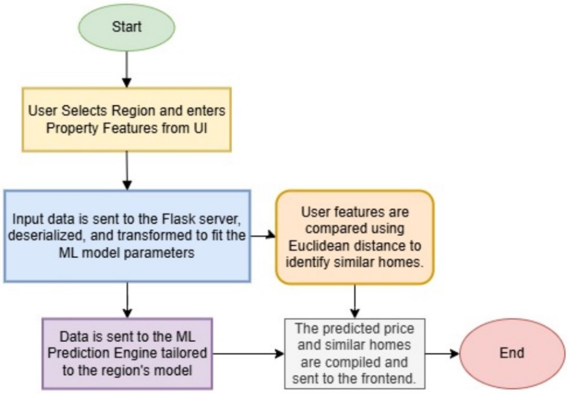

Start
User Selects Region and enters Property Features from UI
Input data is sent to the Flask server, deserialized, and transformed to fit the ML model parameters
User features are compared using Euclidean distance to identify similar homes.
Data is sent to the ML Prediction Engine tailored to the region's model
The predicted price and similar homes are compiled and sent to the frontend.
End

Figure 5. Activity Diagram for the ML Powered Web application

VI. Exploratory Data Analysis

The quantities of homes sold in 2022 and 2023 are illustrated in the bar graphs shown in Fig.6. Notably, the overall trend regarding the number of homes sold remains consistent across both years, with a substantial portion of transactions concentrated in the first quarter, particularly in February, March, and April. However, it is essential to highlight a significant observation regarding the shift in market dynamics between the two years. With the increase in Prime Rates observed from 2022 to 2023, coupled with interest rates surpassing 6.45% in 2023, there is no noticeable decline in the number 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 significant correlation between changes in interest rates and real estate market activity, indicating that financial factors, such as interest rates, may not be a key driver influencing buyer behavior and overall market dynamics during this period for this area.

Furthermore, Fig.7 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 of 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.

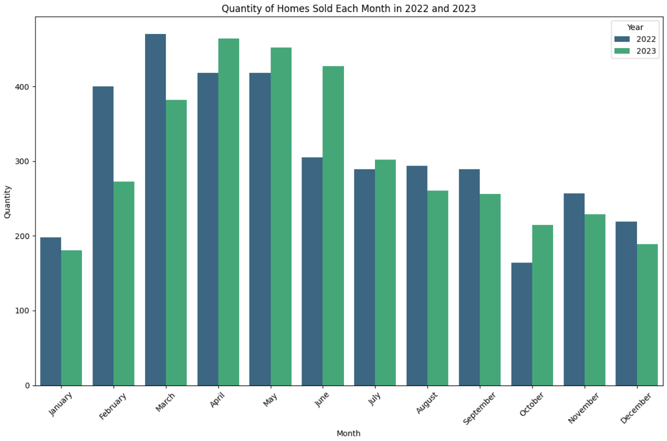


Figure 6. Quantity of Monthly Homes Sold in 2022 and 2023

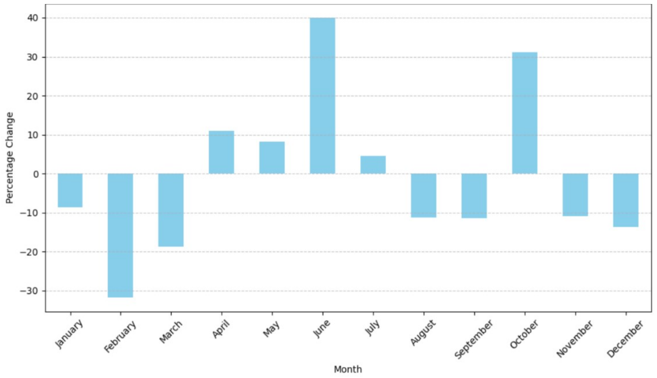


Figure 7. Monthly Percentage Change in Real Estate Transactions between 2022 and 2023

Moreover, the analysis extends to explore the broader trends in the real estate market through Fig. 8. The histogram in Fig.

8 illustrates an overarching downward trajectory in the number of homes sold, displaying a decline of over 50% from January to December of 2022. Concurrently, the accompanying line plot in Figure 9 illustrates a noticeable negative correlation between the escalation of prime rates, starting at 2.5% at the beginning of 2022 and reaching 6.5% by the year's end, and a corresponding 13.75% decrease in average home prices. These observations highlights the complex interaction between macroeconomic factors and the real estate market, emphasizing the subtle relationship between interest rates, market sentiment, and property values.

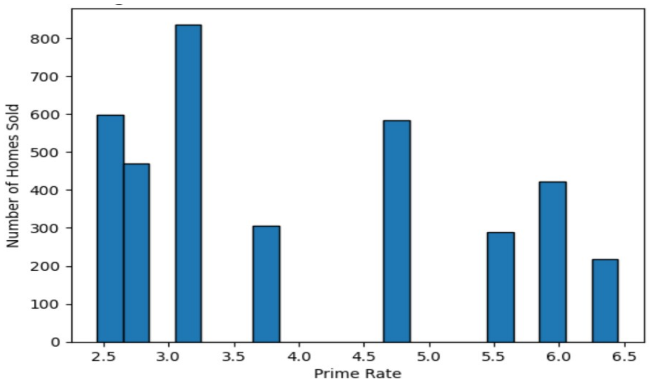


Figure 8. Number of Homes Sold VS. Prime Rate in 2022

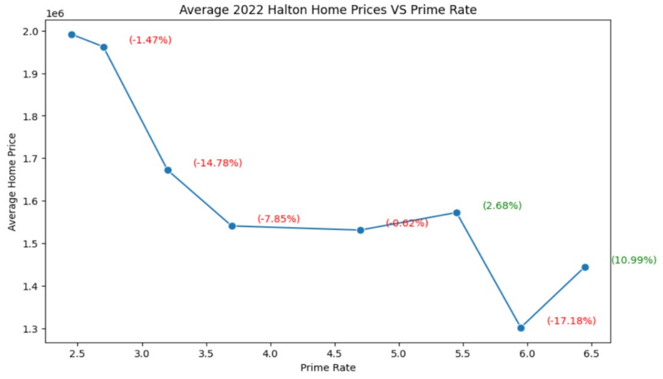


Figure 9. Percentage Trends of Number of Homes Sold VS. Prime Rate in 2022

Figure 10 illustrates the Average Home Sold Price of 2022 and 2023, revealing a modest decrease of 3.94%. In tandem, Figure 11 displays 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.

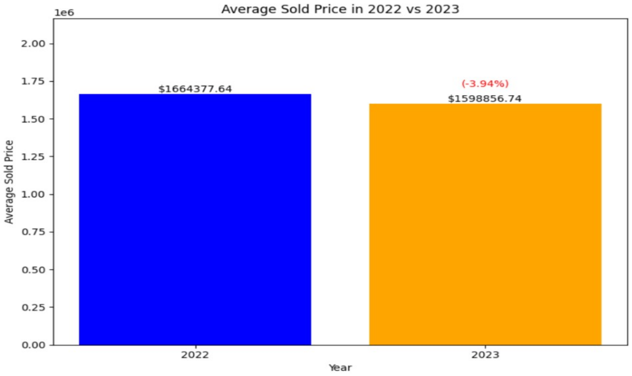


Figure 10. Average Home Price 2022 VS. 2023

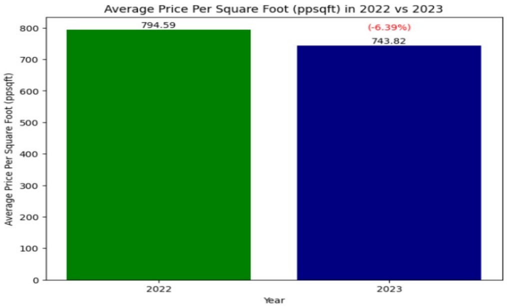


Figure 11. Average Price Per Square Foot 2022 vs. 2023

VII. Conclusion and Future Work

This study highlights the development of a Flask-based web application for home price estimation, which transforms complex machine learning predictions into an accessible, user-friendly tool. This web application allows users to input property details and receive precise price predictions, along with information on three comparable properties. It serves as a valuable resource for homebuyers, sellers, real estate agents, and investors, effectively bridging the gap between sophisticated machine learning models and practical real-world applications.

The integration of property-specific attributes with external economic indicators, such as prime rates, adds depth to the predictive capabilities of the model. This approach addresses existing research gaps by incorporating broader market conditions, resulting in more comprehensive and reliable price forecasts.

Looking ahead, future work could expand the tool's capabilities to include features allowing users to upload and analyze their own datasets, facilitating direct model training and tailored predictions through the web interface. These enhancements would provide users with greater flexibility and further broaden the tool's application to meet diverse needs in the real estate market.

References

[1] X. Li, "Prediction and analysis of housing price based on the generalized linear regression model", *Journal of Housing*

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-243-2

Research, 2022, Received 14 July 2022; Revised 22 August 2022; Accepted 30 August 2022; Published 29 September 2022.

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

[3] M. J. Chowhaan, D. Nitish, G. Akash, N. Sreevidya, and S. Shaik, “Machine learning approach for house price prediction”, *Asian Journal of Research in Computer Science*, vol. 16, no. 23, 2023. DOI: 10.9734/AJRCOS/2023/v16i2339.

[4] P. Furia and 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, and M. K. Muchahari, “House price prediction using machine learning”, *Journal of Real Estate Prediction and Analysis*, vol. 9, 2021.

[6] Authors, “Real-estate price prediction with deep neural network and principal component analysis”, 2022, Received: April 18, 2022; Accepted: November 10, 2022. DOI: 10.2478/otmcj-2022-0016.

[7] A.-R. Mawlood-Yunis and S. Yu, “Applying machine learning and optimization algorithms to perform feature selection”, in *Proceedings of the 7th International Conference on the Dynamics of Information Systems (DIS 2024)*, 2024.

[8] S. Lu, Z. Li, Z. Qin, X. Yang, and 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.

[9] R. MLS, *Realm mls software package*, https://thenewrealm.ca/, Accessed: 2025-02-01.

[10] Redfin Corporation, *Redfin data center*, Accessed: 2024-11-30, 2024.