Thunder Bay House Price Analysis and Recommendation System

GROUP 15

Asma UI Husna - 1276327
Lakehead University

Kinjalben Gherawada - 1279651
Lakehead University

Aye Kyi Kyi Cho - 1276026
Lakehead University

Kunj Patel - 1265570
Lakehead University

Abstract—**Thus, the house price prediction and recommendation system for Thunder Bay using machine learning is an achievement for the authors, in which data was collected through scraping real estate platforms with around 1,951 entries containing critical property features. After adequate preprocessing and feature engineering, several models were put into testing with Random Forest being the one that attained the highest accuracy (R2 = 0.9977). System inputs direct the user to price the property and recommend matching listings based on the Euclidean distance. The key price determinants are property size, number of bedrooms, and location. Although there are some limitations related to either data or geospatial processing that are hurdles, the model provides data-driven insights that serve as a basis for informed decision-making in the property market.**

I. Introduction and Objective

The project intends to develop a house price prediction and recommendation system for Thunder Bay that will assist buyers and sellers in making data-driven parameters while making a decision. In this regard, machine learning techniques were therefore used as analysis to predict the price followed by recommending similar listings after analyzing the features of the properties like size, the number of bedrooms, and location. Hence, to develop this project would bring increased market transparency, fairer pricing, and lower monetary risk by presenting accurate value estimations and personalized recommendations based on something invested by the users' needs.

II. Methods and Approach

A data-driven machine learning methodology has been applied in this project for the prediction of house prices and similar property recommendation. Data was collected using web scraping from huge data sets from real estate platforms, thus needing a preprocessing step followed by feature engineering and normalization by StandardScaler. Other machine learning models were evaluated, but Random Forest outranked all others with high accuracy ($R^2 = 0.9977$). The recommendation algorithm uses Euclidean distance to identify listings akin to the given inputs by the user. Although challenges such as dealing with missing values and geospatial data exist, this approach ensures trustworthy price prediction along with recommendations based on historical data.

A. Dataset Size

The dataset used in this study was collected through web scraping from popular housing listing websites such as Kijiji and tcrealty.ca focused on the Thunder Bay region. It contains 62 entries with 9 original features, later expanded to 12 features through feature engineering.

B. Data Preprocessing

Several preprocessing steps were applied to prepare the data for model training:

- Feature Engineering: New features such as Price per Square Foot, Price per Bedroom and Bathroom, and Total Living Space (sum of bedrooms and bathrooms) were created to enrich the dataset.
- Normalization: The dataset was scaled using StandardScaler, standardizing the features to have a mean of 0 and a standard deviation of 1. This step ensures consistency in scale among the variables and improves model convergence.
- Dimensionality Reduction: This step was intentionally skipped due to the manageable number of features (12), and further reduction was not necessary.
- Handling of Missing Values and Outliers: The team identified these as a challenge and likely took basic steps to mitigate their impact during preprocessing, though exact methods were not detailed.

C. Evaluation Metrics

Several pre-treatments were used in preparation of the data for training the model:

- Feature Engineering: The following new features were introduced: Price per Square Foot, Price per Bedroom, and Bathroom, and Total Living Space (bedrooms plus bathrooms) in order to enhance the dataset.
- Normalization: The data was scaled by StandardScaler, bringing the features into a mean of 0 and a standard deviation of 1. This process ensures scale consistency in the variables and enhances model convergence.
- Dimensionality Reduction: This was bypassed deliberately given the small number of features (12), and additional reduction was not required.
- Dealing with Missing Values and Outliers: The group flagged these as a problem and probably took rudimentary steps to minimize their effect in preprocessing, though precise methods were not explained.

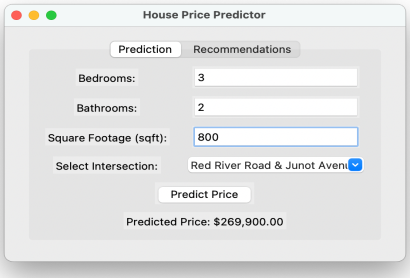
D. Data and Model Limitations

Despite promising results, the project has several limitations:

- **Limited Dataset**: With only 62 entries, the dataset may not generalize well to the broader housing market.
- **Price Variability**: Thunder Bay's diverse housing market causes price variability that can be difficult to model.
- **Geospatial Data Handling**: Features like "Nearest Intersection" were hard to process due to complexity and limited location metadata.
- **Data Availability**: The scarcity of detailed real estate data reduces model accuracy and restricts broader applicability.

III. RESULT AND PERFROMANCE

Yet, the Random Forest model clearly differentiates itself in performing remarkably well, with an R2of 0.9977, achieving very low absolute mean error MAE of 8583.64 in predicting house prices based on the property features. The recommendation engine uses Euclidean distance to identify similar property listings available for end-user suggestion. The analysis has established the property size, number of bedrooms, and location as the complete and absolute cause of the price volatilities. The model may grapple with problems generated by varying markets or small dataset sizes; however, the predictions may be accepted confidently and would provide a useful insight into the processes of buyers and sellers.



Bedrooms: 3

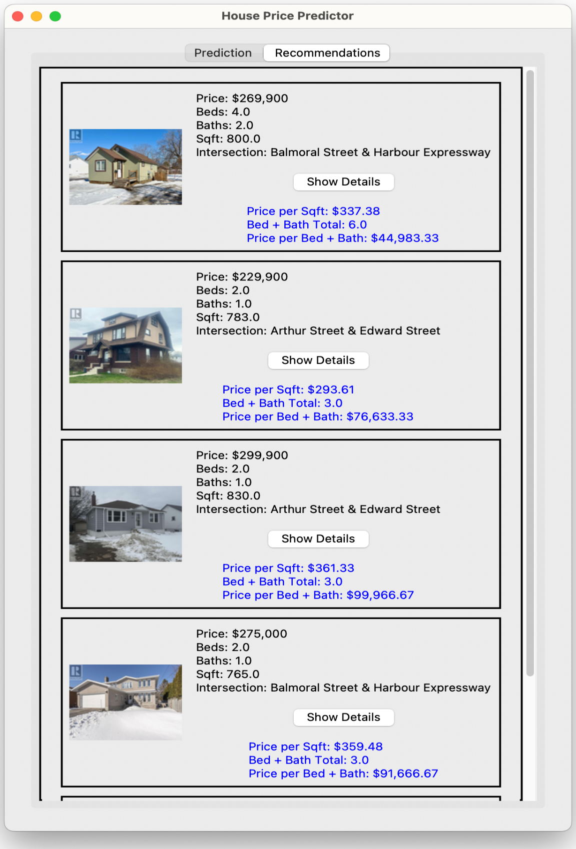
Bathrooms: 2

Square Footage (sqft): 800

Select Intersection: Red River Road & Junot Aven

Predicted Price: $269,900.00

Fig1: Prediction



Price: $269,900
Beds: 4.0
Baths: 2.0
Sqft: 800.0
Intersection: Balmoral Street & Harbour Expressway

Price per Sqft: $337.38
Bed + Bath Total: 6.0
Price per Bed + Bath: $44,983.33

Price: $229,900
Beds: 2.0
Baths: 1.0
Sqft: 783.0
Intersection: Arthur Street & Edward Street

Price per Sqft: $293.61
Bed + Bath Total: 3.0
Price per Bed + Bath: $76,633.33

Price: $299,900
Beds: 2.0
Baths: 1.0
Sqft: 830.0
Intersection: Arthur Street & Edward Street

Price per Sqft: $361.33
Bed + Bath Total: 3.0
Price per Bed + Bath: $99,966.67

Price: $275,000
Beds: 2.0
Baths: 1.0
Sqft: 765.0
Intersection: Balmoral Street & Harbour Expressway

Price per Sqft: $359.48
Bed + Bath Total: 3.0
Price per Bed + Bath: $91,666.67

Fig2: Recommendations

IV. KEY CONTRIBUTIONS

It contributes to the field of real estate market analysis by developing a machine learning-based house price prediction and recommendation system for the city of Thunder Bay. Major contributions include data collection through website scraping, enhancement of the highly predictive and effective features, as well as evaluation of different models where random forest achieved the best success. Apart from the above, this recommendation system can also identify similar properties among several based on their Euclidean distance. Such accurate predictions of prices and data-based recommendations are meant to improve transparency in the market, assist the buyers and the sellers in decision making, and reduce financial risks.

V. STRENGTHS

This project's strengths lie in its data-oriented approach, high model accuracy, and applicability to the real estate market. With web-scraped real estate data applied to model feature engineering and normalization, the system is relevant toward real-world applications. With an exceptional predictive accuracy ($R^2 = 0.9977$), it does create a high reliability of the Random Forest model. In addition, in strengthening the user's personalized recommendations for properties, the system contributes to sound decision making. The ability of the model to handle a variety of property features with the user-friendliness of its insights makes it a powerful tool for buyers, sellers and other stake holders in the housing market.

VI. WEAKNESS

Some positive characteristics are associated with the project, while the presence of some weaknesses impairs its effectiveness. The limited dataset (1,951) might not cover the entire spectrum of diversity of Thunder Bay's housing market, and this reduced dimensionality might affect the generalization of the model as well. Filling in of missing values and geospatial features (neighborhood amenities) becomes problematic. The recommendations are based on Euclidean distance, which might not always be able to properly describe complex property similarities. The fact that the system is not in real-time further implies that market fluctuations are not updated right away. In theory, all these limitations may be addressed through larger datasets, better models, and real-time data integration, to achieve hit accuracy and reliability.

VII. IMPACT

In many ways, this project has far-reaching impact. It opens doorways to transparency in the housing market of Thunder Bay. The model develops data-driven price predictions for housing that could guide potential buyers and sellers in making informed decisions as well as reducing the financial risk involved in dealing with such large amounts of money. Through ma chine learning practices, especially the Random Forest model, this system conveys accurate price estimations and property recommendations thereby increasing market efficiency. It helps in fair pricing and minimizes uncertainty through consideration of the major factors like property size, number of bedrooms, and locations. Despite challenges such as handling geospatial data and the limited information in real estate, this model lays a foundation for future improvements including real-time updates and larger databases, thus creating a tool with great value to the real estate stakeholders.

VIII.CRITICAL ANALYSIS

Thunder Bay House Price Analysis and Recommendation System shows the right trend about property pricing even though it has some limitations. The first reliability issue deals with data obtained from scraping web sources corresponding to listing properties as these sources may not represent a full scope of the market. Added to this is the missing most weighty real estate parameters such as economic trends and mortgage rates besides variations in neighborhood demands. The use of Random Forest under this pretense 2 could also cause overfitting as its generalizability to unseen market conditions is compromised. Furthermore, redundancy or less relevant features that impinge on model efficiency could result from the lack of dimensionality reduction. Future improvements should focus on real-time improvements of the feature set and external economic indicators to strengthen prediction accuracy and recommendation effectiveness. Thunder Bay House Price Analysis Recommendation System is a very good analysis model on property pricing; however, it has certain drawbacks. The listings may not represent the whole market because of the bias that comes with web-scraped data. Also, it misses very significant real estate determinants like economic trends, mortgage rates, and neighborhood-specific demand variations, which compromise the model's ability to pick up the additional price fluctuations accurately. Random Forest algorithm, although highly accurate, is prone to over-fitting and therefore not valid for unseen market conditions. Therefore, the absence of dimensionality reduction can result in non-significant and redundant features reducing the model efficiency. Future improvements should be geared towards real time inputs from feature selection and other economic indicators to improve accuracy in prediction and recommendation efficacy.

IX. SIGNIFICANCE

Thus, the project will change the dynamics of decision-making in real estate in Thunder Bay using data insights. This acceptable access possible due to machine learning, more specifically, the Random Forest model, makes markets much more transparent, leading to fair prices and reduced financial risk from buyers and sellers. By predicting house prices with property features, the user is enabled to make databased decisions, thus simplifying the home buying process and home selling activities. The recommendation system will recommend properties matched to an individual's profile, thus increasing the effectiveness of the property search. The predictive ability of this model can also be applied by realtors and policy makers further to analyze market trends and property valuations toward an informed, data-centric real estate ecosystem.

X. LIMITATION

On the one hand, the Thunder Bay House Price Analysis and Recommendation System has many competitive advantages, but on the other hand, there exist many limitations. The system model does not provide for the target population because based on data scraped from the web, thus there could be instances when this data is not representative or may not be up to date, thereby introducing valid inaccuracies in price prediction. Considering many external economic factors-including interest rates, inflation, and market trends-are all significant determinants of real estate prices, the individual model follows on being incapable in this regard. Furthermore, the system has difficulty with geospatial complications such as desirability in a neighborhood and proximity to facilities or services, which allow for huge differences in price. Besides, since synthetic data might be utilized for certain elements in different ways, it may not really grasp real world market dynamics, thus impeding the model's overall credibility in some cases. Lastly, despite Random Forest in predictive accuracy, it does have some computational costs, and it likely does not generalize well to rapidly changing market conditions and requires a continuous process of modification and improvement.

XI. FUTURE WORK

The future research direction for the Thunder Bay House Price Analysis and Recommendation System lies in improving data quality, model accuracy, and relevance to real life. Data can be increased by adding other sources of real-time market data such as government property records, mortgage trends, and many more, resulting in a stronger prediction reliability. Additional applications that could take price estimations further include advanced machine learning, such as deep learning or ensemble techniques. Spousal cruise analysis along with other economic indicators such as interest rates and local economic conditions will further assist in market dynamic improvement and their incorporation to the model. It also aims at improving the recommendations according to user preferences, sentiment analysis from real estate reviews, as well as interactive dashboards, enhancing usability realistic recommendation and the suggestion model. Finally, the deployment of this model as a cloud-based or mobile application that is constantly updated on real time would allow it to be accessible and more practical to end users considering these factors: house buyers, sellers, and real estate professionals.

XII. CONCLUSION

This system proves to be an efficient predictive and recommended system for house prices in Thunder Bay, which reflects the function of machine learning in the improvement of real estate. It offers price predictions and personal property recommendations on the basis of the random forest model, which effectively determines the major underlying components in the house prices while making the market transparent and assisting potential buyers or sellers towards better decisions. This also denotes many challenges, such as limited data, geospatial complexities, and economy-related issues for further improvements. Despite that, the model is highly predictive and has strong fundamentals to improve its future installations, such as real-time data inclusion, better algorithms, and an enhanced user experience. The power of development can turn this into a tool for the real estate industry, potential investors, and home-buyers in a few years into the future

XIII. END SECTIONS

A. Appendices

The appendices detail the materials that support the Thunder Bay House Price Analysis and Recommendation System. This includes discussing the sources of raw data and the preprocessing steps, along with feature engineering steps. It presents the data visualization techniques through descriptive tables and charts measuring the trends in the dataset: both numerical and categorical data distributions. They also have various model metrics such as MAE, MSE, RMSE, and R2 scores for different models of machine learning so that Random Forest could be chosen to score for itself. The given snippets have also been data-preprocessing, normalization, and training of the model making reproduction easier. Onto additional resources, such as design mockups for the user interface of the recommendation system, as well as a little bit of thought on dealing with missing values, provide a thorough picture on how this project methodology and implementation details were developed

B. Acknowledgment

We would like to express our utmost gratitude to our instructor, **Mahzabeen Emu**, for every bit of advice, guidance, support, and feedback regarding this project. The way in which data science can shape this project would be because of the great from her supervisor's contribution and knowledge. Thanks also to our team, **Asma Ul Husna**, **Aye Kyi Kyi Cho, Kinjalben Gherawada, and Kunj Amitkumar Patel** for their dedication, teamwork and hard work contributing to data collection, preprocessing, model development and final presentation. Acknowledgment to Kijiji, Houseful.ca and City of Thunder Bay Open Data Portal for fulfilling their promise of making open access real estate information available. Last and most definitely not the least, we thank all our well-wishers who are our friends and families for their support and encouragement throughout the project

XIV. REFERENCES

[1] Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer.

[2] City of Thunder Bay. (n.d.). Open Data Portal. Retrieved from https://opendata.thunderbay.ca

[3] Town & Country (Reality Real Estate Brokerage Inc) retrieved from https://tcrealty.ca/

[4] Han, J., Pei, J., Kamber, M. (2011). Data Mining: Concepts and Techniques. Elsevier.

[5] McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.

[6] Canadian Real Estate Association (CREA). (2023). Housing Market Reports. Retrieved from https://www.crea.ca

[7] Zillow Research. (2023). Housing Market Trends Analysis. Retrieved from https://www.zillow.com/research/

[8] Waskom, M. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.

[9] Raschka, S., Mirjalili, V. (2019). Python Machine Learning. Packt Publishing. 8. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

[10] [8]. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press