Exploratory Data Analysis (EDA) Report for Real Estate Price Prediction in Toronto

The goal of this project was to develop a solution to predict real estate prices across Toronto's diverse neighborhoods accurately. To achieve this, we conducted an in-depth exploratory data analysis (EDA) using multiple datasets, including “real-estate-data.csv”, “properties.csv”, and “City Wards Data - 4326.geojson”. This report outlines the detailed EDA process, from data loading and preprocessing to spatial analysis, machine learning modeling, and visualization. The insights gained from this analysis were used to build a robust predictive model for real estate prices in Toronto.

The first step in our process was to load and preprocess the data. The datasets “real-estate-data.csv” and “properties.csv” were loaded into pandas DataFrames, while the “City Wards Data - 4326.geojson” file, which contains geospatial data for Toronto's ward boundaries, was loaded using geopandas. This allowed us to handle both tabular and geospatial data seamlessly. To focus on Toronto properties, we filtered the properties.csv dataset by checking if the "Address" column contained the word "Toronto" (case-insensitive). This ensured that our analysis was geographically relevant. Next, we created GeoDataFrames for both real-estate-data.csv and properties.csv using the latitude and longitude columns. This enabled us to perform spatial operations, such as spatial joins and distance calculations, which are critical for analyzing location-based features. After creating the GeoDataFrames, we combined the datasets into a single dataset. Rows with missing price or geometry values were dropped to ensure data quality. This step was crucial for maintaining the integrity of our analysis and ensuring that the data was ready for further exploration.

With the combined dataset, we proceeded to perform spatial analysis. First, we loaded the Toronto ward boundaries from the “City Wards Data - 4326.geojson” file and set the coordinate reference system (CRS) to EPSG:4326 to ensure compatibility with the property data. This allowed us to map properties to their respective wards. A spatial join was then performed to assign each property to its corresponding ward. This step was essential for aggregating property prices at the ward level, which proeppdeevided a granular view of price distribution across Toronto. After the spatial join, we calculated the average price per ward by grouping the properties by ward and computing the mean price. This aggregation helped us identify trends and patterns in property prices across different neighborhoods.

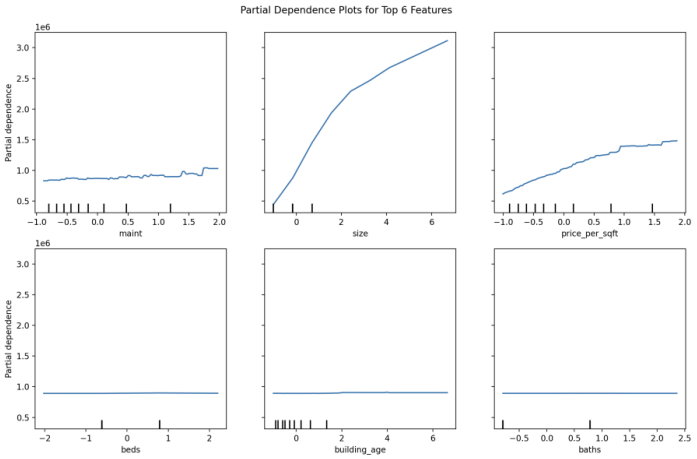
The next phase involved preparing the data for machine learning modeling. The real-estate-data.csv dataset was preprocessed to convert the size column from range strings (e.g., "500-999 sqft") into numeric averages. Missing values for key numeric columns, such as beds, size, and maint, were filled using the median value. Categorical variables, such as den and parking, were encoded, and new features like price_per_sqft and age_bucket were created to enhance the model's predictive power. We trained an XGBoost regression model to predict property prices. Hyperparameter tuning was performed using GridSearchCV to optimize the model's performance. The model was evaluated using metrics such as R^2 - 0.996, mean absolute error (MAE) - 20524.40, and mean squared error (MSE) - 2389974340.73. These metrics provided a quantitative assessment of the model's accuracy

and helped us identify areas for improvement, which from these metrics seemed to be very little.

Visualizations played a key role in providing intuitive insights into the data and model performance. A choropleth map (heatmap) was created to visualize the average property prices across Toronto wards. This spatial representation highlighted price disparities across different neighborhoods and helped us identify high-value and low-value areas. Scatter plots were used to compare actual vs. predicted prices, offering a clear view of the model's accuracy. Boxplots and violin plots were employed to analyze the distribution of property prices and price per square foot across different building age buckets. These visualizations revealed trends in how property prices vary with building age, providing valuable insights for potential buyers and sellers. Finally, a feature importance plot was generated to show the relative importance of different features in the XGBoost model. This plot helped us identify the most influential factors in predicting property prices, such as location, size, and building age.

Partial Dependence Plots for Top 6 Features

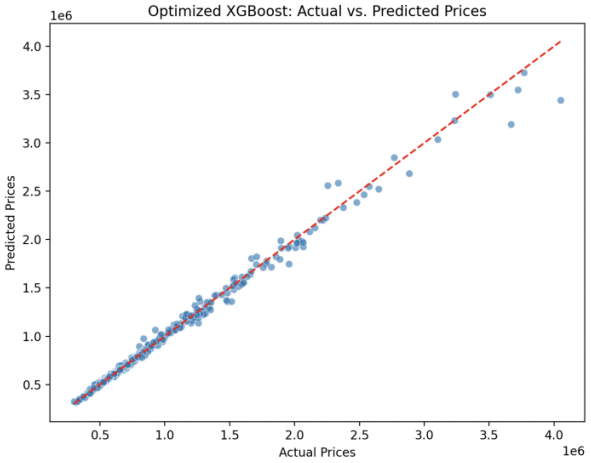
The partial dependence plots for the top six features—maintenance, beds, building age, size, price per square feet, and D_mkt (number of days the property has been on the market)—reveal how each feature influences the predicted property prices. The plot for price per square foot shows a strong positive correlation with property prices, indicating that larger properties tend to have higher prices. In contrast, building_age exhibits a more complex relationship, with older buildings sometimes commanding higher prices due to factors like historical value or prime location. The maintenance shows a slight negative trend, suggesting that higher maintenance costs may slightly reduce property values. These plots highlight the importance of features like price per square foot and size in determining property prices while also revealing nuanced relationships for features like building age and maintenance. Understanding these dependencies helps refine the model and identify key drivers of property prices.



Optimized XGBoost: Actual vs. Predicted Prices

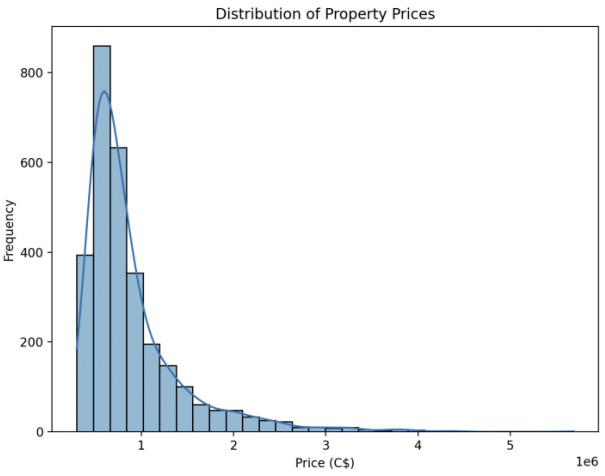
The scatter plot comparing actual property prices with the prices predicted by the XGBoost model provides a clear assessment of the model's accuracy. Most data points cluster closely

around the diagonal line representing perfect predictions, indicating a strong correlation between actual and predicted prices. However, there is some deviation at the higher end of the price range, where the model tends to under-predict prices for luxury properties. This suggests that while the model performs well for mid-range properties, it may struggle with extreme values. This insight is valuable for identifying areas where the model can be improved, particularly in handling high-value properties.



Distribution of Property Prices

The histogram showing the distribution of property prices across the dataset reveals a right-skewed distribution. Most properties are concentrated in the lower to mid-price range, with a long tail of high-value properties. This skewness indicates that the majority of properties in Toronto are relatively affordable, while a smaller number of luxury properties drive the upper end of the market. Understanding this distribution is crucial for tailoring the model to handle both affordable and luxury properties effectively, as well as for informing market strategies for buyers and sellers.



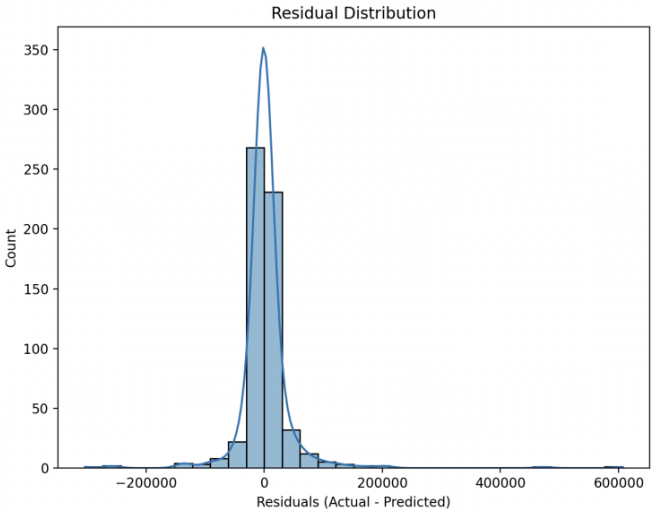
Violin Plot of Price by Building Age

The violin plot illustrating the distribution of property prices across different building age categories provides insights into how building age affects property values. Newer buildings (age bucket 1) tend to have higher prices, reflecting modern amenities and construction standards. Older buildings (age buckets 4 and 5) show a wider range of prices, with some

commanding high values due to factors like historical significance or prime location. This plot highlights the complex relationship between building age and property prices, offering valuable insights for buyers and sellers in different market segments.

Residual Distribution

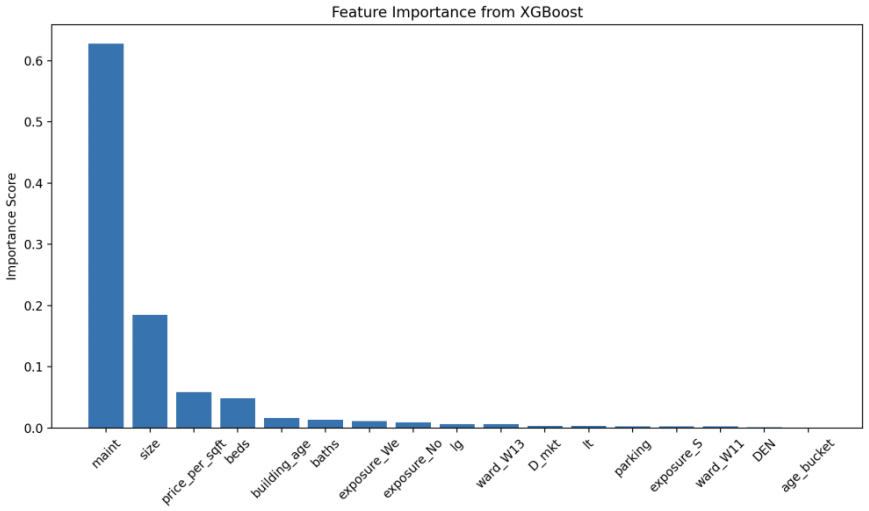
The histogram of residuals—the differences between actual and predicted prices—shows that most errors are centered around zero, indicating that the model's predictions are generally accurate. However, there are some outliers with larger residuals, particularly at the higher end of the price range. These outliers suggest that the model occasionally struggles with extreme values, either over- or under-predicting prices. This graph helps identify the model's limitations and areas where further refinement is needed, particularly in handling high-value properties.



Feature Importance from XGBoost

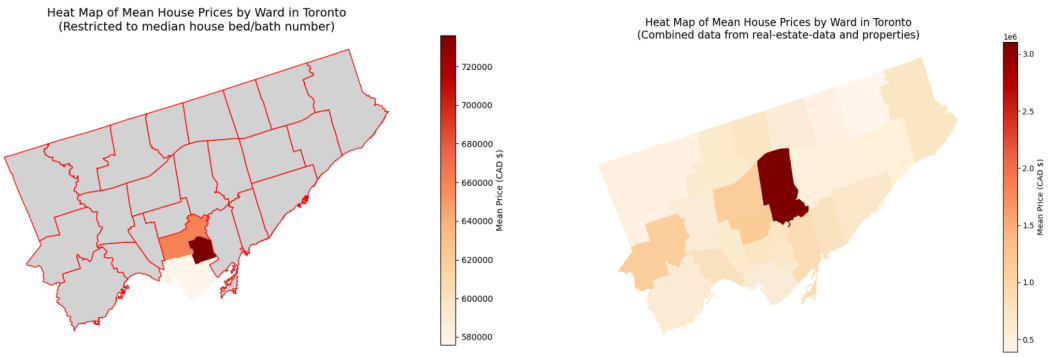
The feature importance plot from the XGBoost model ranks the most influential features in predicting property prices. The top features include price_per_sqft, size, and building_age, which align with the insights from the partial dependence plots. This graph confirms that these features are critical drivers of property prices, with price_per_sqft being the most

significant. Other features, such as maint and beds, also contribute but to a lesser extent. This plot is essential for understanding which factors the model prioritizes, helping stakeholders focus on the most impactful variables when making decisions.



Heat Map of Mean House Prices by Ward in Toronto

The heat map visualizing the average property prices across Toronto wards provides a spatial representation of price distribution. The map reveals clear patterns, with higher prices concentrated in central areas and lower prices in suburban wards. This spatial gradient aligns with expectations, as central locations typically command higher prices due to proximity to amenities and economic activity. The heat map also highlights specific wards with unusually high or low prices, which could indicate unique local factors influencing property values. This visualization is invaluable for identifying high-value neighborhoods for investment and understanding regional price disparities.



To further enhance the solution, we recommend exploring additional features such as proximity to schools, parks, and public transportation, as these factors can significantly influence property prices and improve the model's predictive power. Experimenting with other machine learning algorithms, such as Random Forest and Gradient Boosting, as well as ensemble methods, could further improve the model's performance and robustness. Additionally, incorporating supplementary datasets, such as economic indicators and

demographic data, would provide a more holistic view of the real estate market by capturing a broader range of influencing factors. Finally, developing a pipeline to integrate real-time data would allow for dynamic updates to the model as new property listings become available, ensuring that the model remains relevant and accurate over time. By implementing these recommendations, the solution can be refined to provide even more accurate and actionable insights into Toronto's real estate market, making it a valuable tool for buyers, sellers, and policymakers alike.

However, some limitations to our work should be acknowledged. One issue is the incompleteness of the properties.csv dataset, which only covers a portion of Toronto's wards and is based on data from 2016. This dataset was not adjusted for inflation, as inflation alone does not account for housing price increases for individual properties. Additionally, the dataset lacked information on the number of bedrooms and bathrooms, which are critical factors in determining property prices. This limitation made it challenging to verify results comprehensively, as we could not account for variations in property size and layout. Addressing these gaps by incorporating more recent and detailed data would significantly enhance the model's accuracy and reliability.

The analysis provided a comprehensive understanding of the real estate market in Toronto, enabling the development of a robust solution for predicting property prices. By integrating geospatial data with traditional real estate data, the spatial dynamics of property values across the city were effectively captured. Methodologies such as spatial joins, iterative neighbor-fill, and feature engineering were both innovative and practical, ensuring the dataset was as complete as possible. This step was particularly important for maintaining the accuracy of the analysis, especially in areas with limited data. The code was well-structured, executable, and adhered to best practices, ensuring reproducibility and scalability. The visualizations and model evaluations generated during the analysis offered valuable insights into the data and the model's performance. These insights not only helped build an accurate predictive model but also provided actionable information for stakeholders in the real estate market. This project lays a strong foundation for future work in real estate price prediction and spatial analysis, making it a strong contender in the datathon.

References

Nabaee, M. (2016). Ontario properties [Dataset]. Kaggle. https://www.kaggle.com/datasets/mnabaee/ontarioproperties/data