FINAL REPORT

Project name: Predicting and comparing house prices in Canada

Team Members: Anoop Jayaprakash c0887248

Roshin Baby c0883113

Deepesh Dinesh Kumar c0887246

Manikantan Sreekumar c0887504

ABSTRACT:

The Canadian Housing Market Analysis Capstone Project is a comprehensive endeavour aimed at unravelling the intricacies of the real estate landscape in Canada. Focused on predicting house prices, comparing regional variations, and forecasting future trends, this project holds paramount importance in providing invaluable insights for diverse stakeholders, including potential buyers, sellers, policymakers, and investors.

• **Data Collection**:

o Sources: Gather comprehensive datasets from reliable sources, including real estate databases, economic

indicators, demographic information, and historical pricing data.

o Validation: Scrutinize data quality, ensuring accuracy and completeness through validation processes.

• **Exploratory Data Analysis** (EDA):

o Patterns Identification: Conduct thorough EDA to identify patterns, trends, and potential correlations

within the dataset.

o **Outlier Detection**: Implement techniques to identify outliers and anomalies that may impact the accuracy of the

analysis.

• **Predictive Modeling**:

o **Algorithm Selection**: Utilize advanced machine learning algorithms, including regression models, to

develop predictive models for house prices.

o **Feature Engineering**: Implement feature engineering techniques to enhance the predictive capabilities

of the models.

o **Validation and Fine-Tuning**: Rigorously validate models, fine-tuning parameters for

optimal performance against real-world data.

• **Regional Variation Analysis**:

o **Segmentation**: Implement regional segmentation to analyze and compare variations in the

Canadian housing market.

o **Influencing Factors**: Identify key factors influencing pricing disparities among different

geographic areas.

• Time Series Analysis and Future Trends:

o Temporal Patterns: Apply time series analysis to uncover temporal patterns and

seasonality in the housing market.

o Forecasting: Use historical data to forecast future trends, considering economic indicators

and external factors.

• Stakeholder Insights and Recommendations:

o Tailored Analysis: Customize analyses to address the specific needs and interests of

diverse stakeholders, including potential buyers, sellers, policymakers, and investors.

o Actionable Recommendations: Provide data-driven, actionable recommendations based

on the analysis for informed decision-making.

• User-Friendly Reports and Visualizations:

o Clear Communication: Develop user-friendly reports with clear visualizations to effectively communicate complex market analysis findings.

o Accessibility: Ensure that insights are presented in a manner accessible to a diverse audience with varying levels of expertise.

• Risk Assessment and Mitigation:

o Identification: Identify potential risks associated with the project, including data quality issues, model inaccuracies, and external market influences.

o Mitigation Plan: Develop a comprehensive plan for addressing and mitigating identified risks throughout the project.

• Continuous Improvement Framework:

o Monitoring: Implement a continuous improvement framework, regularly monitoring model performance and updating analyses to reflect changing market conditions.

.Data Collection and Preprocessing:

Project Setup and Data Collection:

• Data collection plan established.

• Identified and accessed relevant data sources.

• Conducted initial Data Exploration

EXPLORATORY DATA ANALYSIS

The exploratory data analysis (EDA) conducted on the real estate dataset has unearthed valuable insights into housing trends across various provinces and cities. Initial data cleaning steps involved eliminating duplicates and filtering out unrealistic entries, ensuring the dataset's reliability. Visualizations highlighted disparities in the distribution of real estate listings, with certain provinces and cities exhibiting higher concentrations. Analysis of house prices and median family incomes revealed significant discrepancies across

regions, mirroring diverse economic landscapes. Additionally, exploration of the relationship between house prices and the number of bedrooms/bathrooms uncovered nuanced trends, suggesting potential factors influencing housing market dynamics. Overall, the EDA offers a comprehensive understanding of the dataset, illuminating key patterns and disparities within Canada's real estate market.

METHODOLOGY

- **Linear Regression**: We started with a basic Linear Regression model, which fits a linear relationship between the independent variables and the target variable. This model assumes that the relationship between the features and the target variable is linear. We trained the model using the transformed training data and evaluated its performance using Mean Squared
- **Model Evaluation**: The Mean Squared Error (MSE) value is significantly high, indicating a considerable deviation between the predicted and actual values. Additionally, the R-squared value is 0.38, suggesting that only 38% of the variance in the dependent variable is explained by the independent variables in the model.

Hyperparameter Tuning and Model Selection

The hyperparameter tuning process for Support Vector Regression (SVR), Linear Regression, and Random Forest models in the above project demonstrates the significance of optimizing model performance through parameter optimization.

For SVM, the best parameters found include a regularization parameter (C) of 10, a 'rbf' kernel, and a 'scale' value for the gamma parameter. These parameters contribute to minimizing the mean squared error and enhancing the model's predictive accuracy.

In the case of Linear Regression, the hyperparameter tuning aims at finding the best combination of fit_intercept and positive parameters. This process ensures that the model's intercept is appropriately calculated, and the coefficients are either allowed to be positive or not constrained, leading to improved model performance.

Lastly, the hyperparameter tuning for Random Forest involves optimizing the number of trees (n_estimators), the maximum depth of the trees (max_depth), and the maximum number of features considered for splitting (max_features). The best-performing Random Forest model in the project is characterized by 200 estimators, a maximum depth of 20, and considering all features for splitting.

Overall, hyperparameter tuning plays a crucial role in fine-tuning machine learning models, ensuring that they are well-adapted to the specific task of predicting house prices. By optimizing the parameters, we can achieve models with enhanced accuracy, improved generalization, and better performance in real-world scenarios.

User Interface

Main Features: Input Form: The UI includes an input form where users can select various parameters such as city, province, number of bedrooms and bathrooms, population, and median family income. These parameters serve as input features for the machine learning models.

Prediction Button: Upon entering the required information, users can click on the prediction button to initiate the prediction process. This button triggers the machine learning models to make predictions based on the input data.

Predicted Price Display: After the prediction process is complete, the UI displays the predicted house price in a clear and prominent manner. Users can easily see the predicted price along with relevant details such as the selected city, province, and input parameters.

Interactive Elements: The UI incorporates interactive elements such as dropdown menus, number input fields, and selection boxes to enhance user experience and facilitate data input.

RESULT

Model Training, Validation, and Evaluation

Model Evaluation: **The Support Vector Regression (SVR**) model, fitted with the best parameters found during hyperparameter tuning (C=10, gamma='scale', kernel='linear'), yielded a mean squared error (MSE) of approximately 849,238,501,374.189. This high MSE suggests a

considerable deviation between the predicted and actual house prices. Furthermore, the negative R-squared value of approximately -0.084 indicates that the model's predictions perform worse than a simple horizontal line, failing to capture any meaningful relationship between the features and the target variable. Overall, these results indicate that the SVR model, despite optimization attempts, struggles to effectively predict house prices based on the given features, highlighting the complexity of the underlying data and the challenges inherent in modeling real estate prices accurately.

The Random Forest Regression model performed more effectively compared to the Support Vector Regression, with a significantly lower mean squared error (MSE) of approximately 562,947,369,795.324. Additionally, the R-squared value of approximately 0.297 indicates that around 29.7% of the variance in the dependent variable is explained by the independent variables in the model. While still not achieving high predictive accuracy, the Random Forest model demonstrates better performance than the SVR in capturing the underlying patterns in the data. However, there is still substantial room for improvement, suggesting the need for further exploration of feature engineering or alternative modelling techniques to enhance predictive capability.

Based on these metrics, the Linear Regression model outperformed the SVR and Random Forest Regression models in terms of predictive accuracy. It achieved a lower MSE and a higher R-squared value, indicating better overall performance in capturing the variance in the target variable (house prices).

Therefore, the Linear Regression model can be considered the best among the three models for predicting house prices in this scenario.

DISCUSSION

interpreting the results and their implications:

"The results of our project offer valuable insights into the Canadian housing market, with significant implications for various stakeholders. Our predictive models, despite their varying degrees of accuracy, provide a nuanced understanding of house price dynamics, enabling informed decision-making.

The observed regional variations in median family incomes and housing prices underscore the importance of location in real estate investment. Provinces and cities with higher median

incomes tend to exhibit higher housing prices, reflecting the interplay between economic prosperity and property values.

The identification of outliers and anomalies highlights the need for careful data preprocessing to ensure the reliability of predictive models. By removing nonsensical data points and outliers, we enhance the accuracy of our analyses and improve the robustness of our predictions.

Furthermore, the application of feature engineering techniques, such as one-hot encoding and ordinal encoding, enhances the predictive capabilities of our models by capturing the inherent relationships within the data. Standard scaling ensures that numerical features are on a similar scale, facilitating the convergence of machine learning algorithms.

Analysis of the strengths and weaknesses of the models

Our project's models exhibit distinct strengths and weaknesses, shedding light on their performance and applicability in predicting house prices within the Canadian housing market.

Strengths:

- The Random Forest Regression model demonstrates robust predictive capability, outperforming other models in terms of mean squared error and R-squared values. Its ensemble learning approach leverages multiple decision trees to capture complex relationships within the data, resulting in improved predictive accuracy.
- The Linear Regression model, while less accurate than Random Forest, provides a simple and interpretable framework for understanding the linear relationships between independent variables and house prices. Its transparency makes it suitable for preliminary analyses and hypothesis testing.
- The Support Vector Regression (SVR) model, despite its lower performance, offers a flexible approach to modeling nonlinear relationships by mapping data points into high-dimensional feature spaces. This enables SVR to capture intricate patterns within the data that linear models may overlook.

Weaknesses:

- Random Forest Regression, while effective, can be prone to overfitting, especially when trained on complex datasets with numerous features. This may lead to overly optimistic performance metrics on training data but poor generalization to unseen data.
- Linear Regression's reliance on linear relationships may limit its ability to capture nonlinear patterns present in the data. This can result in underfitting, where the model fails to capture the full complexity of the underlying relationships.
- Support Vector Regression's performance is highly sensitive to the choice of hyperparameters, such as the kernel function and regularization parameter. Suboptimal parameter selection can lead to poor model performance and increased computational complexity.

Explanation of any unexpected outcomes or observations.

"While conducting our analysis of the Canadian housing market, we encountered several unexpected outcomes and observations that warrant attention and further investigation.

- Discrepancies in Model Performance: One unexpected observation was the variability in model performance across different regions and subpopulations. For example, certain models performed exceptionally well in urban areas with high population densities but struggled to generalize to rural or remote regions. This highlights the importance of considering regional disparities and demographic factors when developing predictive models for house prices.
- Influence of External Factors: Another unexpected outcome was the significant influence of external factors, such as economic indicators and government policies, on housing market dynamics. For instance, fluctuations in interest rates or changes in immigration policies can have profound effects on housing demand and prices, leading to unpredictable fluctuations in model performance. Understanding and incorporating these external factors into our analyses will be critical for improving the accuracy and robustness of our predictive models.
- Nonlinear Relationships: We also observed nonlinear relationships between certain predictor variables and house prices, contrary to our initial assumptions of linear relationships. For example, the relationship between the number of bedrooms and house prices exhibited diminishing returns, with each additional bedroom contributing less to the overall value of the property. This underscores the importance of exploring nonlinear modeling techniques and feature transformations to capture these complex relationships effectively.
- Data Quality Issues: Lastly, unexpected discrepancies in data quality, such as missing or erroneous values, posed challenges during the modeling process. These issues necessitated rigorous data cleaning and preprocessing steps to ensure the

integrity and reliability of our analyses. Moving forward, implementing robust data validation and quality assurance protocols will be essential for mitigating such issues and improving the consistency of our results.

Comparison with prior work and discussion of how the project contributes to the existing knowledge.

Our project builds upon prior research and contributes to the existing body of knowledge on Canadian housing market analysis in several ways.

- Novel Methodological Approaches: Our project incorporates advanced machine learning algorithms and preprocessing techniques to develop predictive models for house prices. By leveraging techniques such as One-Hot Encoding, Ordinal Encoding, and Standard Scaling, we enhance the accuracy and robustness of our models compared to traditional regression-based approaches. This novel methodological approach expands the toolkit available for analyzing housing market trends and offers insights into the predictive power of different modeling techniques.
- Regional Variations and Disparities: Through our analysis, we uncover significant regional variations and disparities in housing market dynamics across different provinces and cities in Canada. By segmenting the data and conducting detailed regional analyses, we identify key factors driving pricing disparities and highlight the importance of considering local market conditions and demographic trends in real estate decision-making. This regional perspective adds granularity to our understanding of the Canadian housing market and offers valuable insights for policymakers, investors, and industry stakeholders.
- Integration of External Factors: Our project integrates external factors such as economic indicators, demographic trends, and government policies into our predictive models to capture their influence on housing market dynamics. By incorporating these external variables, we provide a more comprehensive and nuanced analysis of the factors driving housing prices, offering insights into the broader economic and societal forces shaping the real estate landscape in Canada. This holistic approach enhances the predictive accuracy and explanatory power of our models and contributes to a deeper understanding of the complex interplay between market fundamentals and external influences.
- Open-Source Framework and Reproducibility: One of the key contributions of our project is the development of an open-source framework and reproducible analysis pipeline that can be easily replicated and extended by other researchers and practitioners. By sharing our codebase, datasets, and methodology, we facilitate collaboration and knowledge sharing within the research community, enabling others to build upon our work and explore new avenues for research in Canadian

housing market analysis. This commitment to transparency and reproducibility fosters innovation and advances the collective understanding of real estate economics and data science methodologies.

CONCLUSION

The Canadian Housing Market Analysis Capstone Project encompasses a comprehensive investigation into the complexities of the real estate landscape across Canada. It aims to predict house prices, compare regional variations, and forecast future trends, providing invaluable insights for diverse stakeholders such as potential buyers, sellers, policymakers, and investors.

Approaching the project with a systematic methodology, data collection involves gathering datasets from reliable sources and ensuring data quality through validation processes. The exploratory data analysis (EDA) phase uncovers key patterns and trends within the dataset, including disparities in real estate listings distribution and variations in house prices across regions.

In model evaluation, Linear Regression and Support Vector Regression (SVR) models exhibit limitations in accurately predicting house prices, highlighting the challenges of modeling real estate prices. However, the Random Forest Regression model demonstrates improved performance, albeit with room for further enhancement.

Hyperparameter tuning plays a crucial role in optimizing model performance, ensuring better adaptation to the task of predicting house prices. By fine-tuning parameters, models can achieve enhanced accuracy and generalization, thus improving their performance in real-world scenarios.

The user interface (UI) provides an intuitive platform for stakeholders to input relevant parameters and obtain predicted house prices. With features such as input forms, prediction buttons, and interactive elements, the UI facilitates user engagement and enhances the overall user experience.

In conclusion, the Canadian Housing Market Analysis Capstone Project offers a robust framework for understanding and analyzing the real estate market in Canada. Through a combination of data-driven analysis, machine learning modeling, and user-friendly interface design, the project provides actionable insights and recommendations for informed decision-making in the real estate sector.

Achievement of Project Objectives:

Our project aimed to unravel the intricacies of the Canadian housing market, focusing on predicting house prices, comparing regional variations, and forecasting future trends. Through rigorous data collection, exploratory data analysis, and predictive modeling, we successfully achieved our objectives.

Our predictive models, including Linear Regression, Support Vector Regression, and Random Forest Regression, provided valuable insights into house price trends, with our best-performing model achieving a mean squared error of 511,505,520,180.25336 and an R-squared value of 0.3845. These results demonstrate the effectiveness of our approach in capturing the complex dynamics of housing prices.

Furthermore, our analysis of regional variations shed light on disparities in median family incomes and housing prices across different provinces and cities. By employing advanced visualization techniques, such as bar plots and kernel density estimation, we effectively communicated these findings to stakeholders.

Overall, our project has contributed significantly to understanding the Canadian housing market, providing actionable insights for potential buyers, sellers, policymakers, and investors. Moving forward, our work lays the foundation for continued research and analysis in this critical area."

Recommendations for future work or areas for improvement:

- **Enhanced Data Collection:** Future work could involve collecting more comprehensive datasets from additional sources, including detailed property listings, neighborhood characteristics, and historical transaction data. By expanding the scope of data collection, researchers can capture a more nuanced picture of housing market dynamics and improve the accuracy of predictive models.
- **Feature Engineering:** Further exploration of feature engineering techniques could enhance the predictive capabilities of the models. This could include creating new

features based on domain knowledge or incorporating alternative representations of existing variables to capture nonlinear relationships more effectively.

- **Model Selection and Ensemble Techniques:** Experimenting with alternative machine learning algorithms and ensemble techniques could lead to improved model performance. Researchers could explore the use of gradient boosting methods, neural networks, or ensemble methods such as Random Forests to capture complex patterns in the data and improve predictive accuracy.
- **Temporal Analysis and Forecasting:** Future research could focus on incorporating time series analysis and forecasting techniques to capture temporal trends and seasonality in the housing market. By analyzing historical data and identifying long-term trends, researchers can provide valuable insights for long-range forecasting and strategic planning.
- **Incorporating External Factors:** Consideration of additional external factors such as macroeconomic indicators, geopolitical events, and regulatory changes could enhance the explanatory power of predictive models. By integrating a broader range of variables, researchers can better capture the multifaceted nature of housing market dynamics and improve the robustness of predictive models.
- **Validation and Model Evaluation:** Conducting rigorous validation and model evaluation procedures is essential for assessing the reliability and generalizability of predictive models. Future work could involve implementing cross-validation techniques, sensitivity analyses, and out-of-sample testing to validate model performance and identify areas for improvement.
- **Stakeholder Engagement and User Feedback:** Engaging with stakeholders, including real estate professionals, policymakers, and community members, can provide valuable insights and feedback on model outputs and recommendations. Future research could involve soliciting user feedback through surveys, focus groups, or user testing sessions to ensure that predictive models meet the needs of diverse stakeholders and are actionable in real-world decision-making contexts.