Cover Page – MSc Business Analytics Consultancy Project/Dissertation 2023-24

Title of Project:

Predictive Modeling of House Price - Integrating Traditional Real Estate Data with Points of Interest in Seattle

Written by:

Ameermahskah Sutoyo

Date:

17th August 2024

Word Count:

12,923

Disclaimer:

I hereby declare that this dissertation is my individual work and to the best of my knowledge and confidence, it has not already been accepted in substance for the award of any other degree and is not concurrently submitted in candidature for any degree. It is the end product of my own independent study except where other acknowledgement has been stated in the text.

[This page has been intentionally left blank]

ABSTRACT

This dissertation explores the integration of advanced machine learning techniques and Points of Interest (POI) data to enhance house price prediction models, specifically focusing on the Seattle housing market. The research identifies limitations in traditional predictive models, which often rely on historical data and basic structural attributes, failing to capture the complexities of modern urban environments. By incorporating a comprehensive dataset that includes both conventional real estate metrics and diverse POIs—such as schools, parks, and transportation hubs—the study aims to provide a more nuanced understanding of factors influencing property values.

The methodology involves the application of various machine learning algorithms, including Linear Regression and Random Forest, to develop a robust predictive model. This model will be rigorously tested against real-world data to evaluate its accuracy and reliability. The findings are expected to offer valuable insights for stakeholders, including real estate professionals, investors, and policymakers, ultimately contributing to more informed decision-making in urban development and housing policies. By addressing the gaps in current methodologies, this research aspires to advance the field of real estate analytics and improve the accuracy of property valuations in rapidly evolving markets like Seattle.

ACKNOWLEDGMENTS

The completion of this project was made possible by the great assistance and encouragement of a couple individuals whose efforts were critical to its success. First and foremost, I want to express my gratitude to my distinguished supervisor, Faiza Tabassum, for her unfailing leadership and intellectual mentorship during this project. Her vast knowledge, intelligent recommendations, and constructive criticism have all played a role in defining the course of this study and propelling it to its full potential especially considering the limited timeframe that I have due to switching projects in the middle of the dissertation period. I would also like to thank our program director, David Alderton, to have helped me through difficult times I had when facing limitations from my initial project and guiding me through the process of switching to the project that I am currently doing.

I also want to mention the gratitude I have for my wife, Aisha Zahira Natanegara, for her unconditional love, encouragement, and support that she have always shown while also completing her own Master's degree. I would also like to mention both of my parents, Adi and Mia, and also my siblings, Anky and Aybee, for always supporting me throughout my life. I would also like to thank my friends and families whose name I cannot mention one by one. Their presence in my life has been a constant source of strength and motivation that has helped me throughout the year.

Furthermore, I acknowledge the use of ChatGPT (https://chat.openai.com/) to plan my essay and generate some initial ideas, which I used in research and self-study in drafting this assessment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATION

LIST OF APPENDICES

1. INTRODUCTION

1.1 Background of the Study

1.1.1. House Price Predictions

The real estate market plays a pivotal role in the global economy, influencing both macroeconomic policies and personal financial decisions. Accurate prediction of property values is immensely valuable to investors, policymakers, and individuals, as it drives decision-making in contexts ranging from large-scale urban planning to individual investments. The advent of advanced data analytics has revolutionized real estate predictions. Machine learning, in particular, has enabled sophisticated analysis of extensive datasets, incorporating not just basic property characteristics but also nuanced geographic and demographic information (Nguyen, Tran, & Le, 2021).

Incorporating Points of Interest (POIs)—such as schools, parks, and transportation hubs—into predictive models enhances the granularity and accuracy of property valuations (Mora-García et al., 2023). This trend reflects broader advancements in urban analytics, where data-driven insights shape smarter, more livable cities. Rapid urbanization worldwide further amplifies the need for advanced predictive models. As urban centers expand, property valuation parameters must integrate a comprehensive set of data to reflect the modern economic landscape accurately (Xu, 2023; Sagala & Cendriawan, 2022).

The real estate market is multifaceted, influenced by economic, social, and environmental factors. Traditional models for predicting house prices have focused on structural attributes such as size, age, and architectural style, and locational factors like proximity to central business districts (CBDs) (Malpezzi, 1999; Sirmans, Macpherson, & Zietz, 2005). However, modern urban settings require a wider array of factors to influence property values. POIs, including schools, parks, public transportation hubs, shopping centers, dining establishments, and healthcare facilities, significantly determine the desirability and value of residential properties. These amenities enhance residents' quality of life and convenience, thus affecting property values (Gibbons & Machin, 2008; Kolko, 2011).

For instance, proximity to high-quality schools increases house prices, as families are willing to pay a premium for better education opportunities (Cheshire & Sheppard, 2004; Black, 1999). Access to parks and recreational areas also correlates positively with property values due to their aesthetic appeal and recreational opportunities (Crompton, 2001; Anderson & West, 2006). Public transportation is another critical factor; it reduces commuting times and enhances connectivity,

increasing the attractiveness of properties near transit hubs (Debrezion, Pels, & Rietveld, 2007; Gibbons & Machin, 2005). Additionally, shopping centers and dining establishments contribute to neighborhood vibrancy and convenience, making them more desirable (Kolko, 2011; Glaeser, Kolko, & Saiz, 2001). Proximity to healthcare facilities also influences house prices, as access to essential services is a valuable amenity (Pace, Gilley, & Sirmans, 2000). Recent studies emphasize the importance of these amenities, showing their substantial impact on housing prices (Li, 2023; Harvard JCHS, 2023).

Understanding the impact of these amenities on housing prices is increasingly important as urban environments become more complex. This dissertation aims to address existing models' limitations by integrating comprehensive POI data to enhance the accuracy of house price predictions.

1.1.2. Seattle Housing Market

The Seattle housing market, a dynamic and rapidly evolving sector, is significantly influenced by the city's robust economic growth, population influx, and status as a major hub for technology companies. This has led to increased housing demand among a diverse population, making Seattle a vibrant cultural center (Brown, 2020). Given the complex nature of the market, affected by factors such as location, amenities, and economic conditions, accurate house price predictions become crucial. They are vital tools for buyers, sellers, real estate agents, and policymakers, providing insights into market trends and assisting in risk management related to property investments.

Traditional prediction methods often rely on historical data and basic statistical models, which may fall short of capturing the intricacies of the current market dynamics (Smith & Thompson, 2022). However, with the advancement of machine learning technologies and the availability of extensive datasets, there is potential to develop more sophisticated models. These models aim to deliver more reliable and accurate predictions, addressing the needs of a market as dynamic as Seattle's, where the tech industry's growth plays a pivotal role in shaping real estate trends (Nguyen et al., 2021).

1.2. Research Foundation and Aims

1.2.1. Problem Statement

Accurately predicting house prices in Seattle presents several challenges. Traditional predictive models often fall short due to their reliance on historical data and limited variables, which do not fully capture the complex factors influencing house prices today. These models typically focus on structural attributes such as the

Commented [FT1]: You have so many headings such as
'problem statement'
'significance of the study'
'Scope'

In my view these three headings should come under one heading eg 'problem statement'
Also, there is a lot of repetition; overall, I found this chapter a bit longer

Commented [L2R1]:

size, age, and architectural style of properties, along with basic locational factors like proximity to central business districts (Malpezzi, 1999; Sirmans, Macpherson, & Zietz, 2005).

Furthermore, the rapid urbanization and technological advancements in cities like Seattle add layers of complexity to property valuation. The traditional parameters of property valuation—location, size, and condition—are no longer sufficient. Modern valuation models must integrate comprehensive data, including POIs, to accurately reflect the current market dynamics (Xu, 2023; Sagala & Cendriawan, 2022).

Despite the potential of advanced machine learning models, there are still gaps in integrating diverse data sources to improve prediction accuracy. For instance, combining datasets from Redfin, which provides detailed property information, and Google API, which offers insights into nearby amenities, could enhance the robustness of predictive models. However, effectively merging and analyzing these data sources to produce reliable predictions remains a significant challenge (Nguyen et al., 2021; Mora-García et al., 2023).

Commented [FT3]: This is an important paragraph which should be highlighted

Addressing these gaps is crucial for developing a house price prediction model that not only incorporates traditional factors but also leverages comprehensive POI data. This dissertation aims to build such a model for Seattle, using advanced machine learning algorithms to provide accurate and reliable house price predictions. By doing so, it will offer valuable insights for stakeholders, including buyers, sellers, investors, real estate agents, and policymakers, enabling them to make informed decisions based on more accurate property valuations.

1.2.2. Significance of the Study

The significance of this study lies in its potential to enhance real estate valuation accuracy and efficiency through the integration of advanced data analytics and machine learning. This research addresses several key aspects:

- **Theoretical Contribution** The study aims to contribute to the academic discourse by blending traditional real estate valuation methods with innovative data science techniques. It extends the understanding of how Points of Interest (POIs) and other non-traditional data can be systematically incorporated into predictive models, thus enriching the theoretical framework of real estate analytics (Jones et al., 2020). By integrating these diverse data sources, the research explores new dimensions in property valuation, contributing to a more holistic approach in the field of real estate studies (Smith & Thompson, 2022).

- **Practical Implications**: For real estate professionals, investors, and urban planners, the findings of this study offer potential enhancements in decision-making processes. By providing more accurate predictions of property values, stakeholders can better assess investment risks and opportunities, optimize portfolio management, and refine marketing strategies (Williams & Smith, 2021). The model developed in this study aims to provide actionable insights that can lead to more informed and strategic decisions in real estate transactions and development projects.

- **Policy Making**: The insights derived from this research can inform urban development policies and real estate market regulations. By understanding the impact of various urban features on property values, policymakers can make more informed decisions regarding infrastructure development, zoning laws, and housing regulations (Taylor & Johnson, 2022). This can lead to more effective policies that promote sustainable urban growth, equitable housing opportunities, and enhanced quality of life for residents.

- **Economic and Social Impact**: Accurately predicting house prices can have significant economic and social implications. Improved valuation models can help stabilize the housing market by providing more reliable pricing information, thereby reducing speculation and market volatility. Additionally, the research can contribute to social equity by highlighting the importance of accessible amenities, influencing policy to ensure that desirable features like schools, parks, and healthcare facilities are equitably distributed across different neighborhoods (Harvard JCHS, 2023)

These contributions aim to bridge the theoretical and practical aspects of real estate valuation, providing a comprehensive approach that enhances both the academic field and industry practices. By integrating advanced machine learning techniques with comprehensive POI data, this study aspires to deliver a robust and practical tool for accurately predicting house prices in Seattle, thus addressing the limitations of current predictive models and contributing to the body of knowledge in real estate analytics.

1.2.3. Scope

This study will focus on the Seattle real estate market, employing a comprehensive dataset that includes both traditional real estate metrics and a wide array of POIs data. The scope includes:

- **Geographical Focus**: The research is confined to the Seattle metropolitan area, including its various neighborhoods and suburbs. This region is chosen

due to its dynamic real estate market, significant economic growth, and diverse urban features that impact property values (Brown, 2020).

- **Data Scope**: Utilizing data from multiple sources, including property listings from Redfin and POI data extracted via Google APIs as of June 30, 2024. This approach ensures a rich dataset that captures a broad spectrum of factors influencing property values.

- **Machine Learning Techniques**: The study employs various Machine Learning algorithms to develop the predictive model such as Linear Regression, Random Forest, Gradient Boosting, and Neural Networks (Smith & Thompson, 2022).

- **Variables and Features**: The study integrates traditional real estate variables (e.g., property size, number of rooms, year built) with POIs (e.g., proximity to schools, parks, public transportation, shopping centers, and healthcare facilities) to enhance the model's accuracy and relevance (Gibbons & Machin, 2008; Kolko, 2011).

- **Model Development and Testing**: The predictive model will be developed using a training dataset and tested using a separate validation dataset to ensure its accuracy and reliability. The model's performance will be evaluated using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) (Nguyen et al., 2021; Xu, 2023).

- **Analysis and Interpretation**: The study will conduct a detailed analysis of the impact of various POIs on property values, providing insights into how different amenities and neighborhood characteristics influence real estate prices in Seattle (Cheshire & Sheppard, 2004; Gibbons & Machin, 2008).

- **Limitations**: The study acknowledges certain limitations, such as the reliance on the accuracy and completeness of the datasets used, potential biases in the data, and the generalizability of the findings to other regions or markets (Smith & Thompson, 2022).

By focusing on these aspects, the study aims to develop a robust and practical predictive model that can provide accurate house price predictions for the Seattle housing market. The findings will offer valuable insights for various stakeholders, including real estate professionals, investors, urban planners, and policymakers, enhancing their decision-making processes and contributing to more efficient and equitable urban development.

1.2.4. Objectives

The objectives of this study are designed to address the gaps in current real estate predictive analytics and enhance the interpretability and usability of predictive models:

- Conduct a comprehensive literature review on real estate price prediction methodologies, focusing on the integration of machine learning techniques and Points of Interest (POIs).
- Create a predictive model integrating traditional real estate data with POIs, tailored for the Seattle market, using advanced machine learning techniques.
- Rigorously test the developed model against benchmarks and real-world data, assessing predictive accuracy, reliability, and practical applicability.
- Conduct a detailed analysis of the impact of various POIs on property values in Seattle, quantifying the influence of different amenities and neighborhood characteristics.

These objectives aim to bridge the theoretical and practical aspects of real estate valuation, providing a comprehensive approach that enhances both the academic field and industry practices.

Commented [FT4]: Can you pl condensed these objectives, eg each objective be like a single statement?

Commented [FT5]: I feel it's a repetition

1.3 Dissertation Structure

The dissertation is structured to provide a comprehensive analysis of real estate price prediction models, focusing on the Seattle market. Chapter 2 reviews existing research on house price prediction, emphasizing machine learning applications and key influencing factors, while identifying gaps in current methodologies. Chapter 3 details the research design, data collection from Redfin and Google Places API, preprocessing steps, and the machine learning models used, including training and evaluation techniques. Chapter 4 presents the findings, including descriptive statistics, model performance, feature importance, and a comparison of models. Chapter 5 interprets the results, discussing practical and policy implications, study limitations, and recommendations for future research. Finally, Chapter 6 summarizes the key findings, contributions to knowledge, and offers concluding remarks on the study's impact and future directions.

2. LITERATURE REVIEW

2.1. Introduction

The purpose of this chapter is to provide a comprehensive overview of the existing literature on house price prediction models, with a particular focus on the integration of diverse data sources and advanced machine learning techniques. A thorough literature review is crucial in the context of predictive modeling in real estate as it establishes the theoretical foundation, identifies current trends and gaps, and informs the methodological approach of the present study.

2.2. Factors influencing house prices

House prices are influenced by a complex interplay of structural attributes, locational factors, and points of interest (POI). Understanding these factors is crucial for developing accurate house price prediction models.

Structural attributes of a property have consistently been identified as significant determinants of house prices. These attributes include:

- Size: Total square footage significantly impacts price (Sirmans et al., 2006).
- Age: Newer homes often command higher prices due to modern amenities (Wilhelmsson, 2008).
- Number of rooms: Bedrooms and bathrooms affect value (Zietz et al., 2008).
- Architectural style: Influences price in specific markets (Buitelaar & Schilder, 2017).

Sirmans et al. (2006) conducted a meta-analysis of hedonic pricing studies and found that square footage, lot size, and age were the most frequently used and statistically significant variables in house price models. This underscores the importance of these structural attributes in determining property values.

The importance of location in real estate valuation is a well-established principle. Key locational factors include:

- Proximity to the central business district (CBD): Herath and Maier (2010) demonstrated that distance to the CBD significantly impacts house prices, with properties closer to the city center generally commanding higher prices.
- Neighborhood characteristics: Crime rates, school quality, and socioeconomic status influence values (Ceccato & Wilhelmsson, 2020).
- Accessibility: Proximity to transportation routes affects property values (Debrezion et al., 2007).

The impact of location on house prices highlights the importance of incorporating spatial data and analysis techniques in house price prediction models.

Recently, there has been growing recognition of the importance of POIs in property valuation. POIs that can significantly influence house prices include:

- Schools: Proximity to high-quality schools is often a major factor for families with children (Wen et al., 2014).
- Parks and green spaces: Access to recreational areas can increase property values (Panduro & Veie, 2013).
- Public transportation: Wen et al. (2014) found that proximity to subway stations in Beijing had a positive impact on housing prices.
- Shopping centers: Convenient access to retail amenities can boost property values (Des Rosiers et al., 1996).
- Dining establishments: A vibrant local food scene can make an area more desirable (Zukin et al., 2009).
- Healthcare facilities: Proximity to hospitals and clinics can be important, especially for older homebuyers (Breuer et al., 2021).

Li et al. (2019) demonstrated the significant influence of various POIs on house prices in Wuhan, China, highlighting the importance of considering these factors in house price prediction models. The integration of POI data into house price prediction models represents a significant advancement in the field. It allows for a more nuanced understanding of how neighborhood amenities and urban infrastructure impact property values. This approach can capture the complex interactions between various urban features and house prices, potentially leading to more accurate predictions.

In conclusion, house price determinants are multifaceted and interconnected. As data availability and computational methods improve, prediction models can incorporate these diverse factors more effectively, leading to more sophisticated and accurate predictions (Mullainathan & Spiess, 2017).

2.3. Traditional Statistical Methods in Real Estate Price Prediction

Several traditional statistical methods have been widely used in real estate price prediction:

- Hedonic Pricing Model: Introduced by Rosen (1974), this model assumes property price is determined by a combination of attributes:

$$P = \beta^0 + \beta^1 X^1 + \beta^2 X^2 + \ldots + \beta_n X_n + \varepsilon \quad (1)$$

  Where P is the property price, $X_1 \ldots X_n$ are property characteristics, $\beta_0 \ldots \beta_n$ are coefficients, and ε is the error term.

Sirmans et al. (2006) found square footage, lot size, and age were the most significant variables. However, it often struggles with non-linear relationships (Mayer et al., 2019).

- Repeat Sales Method: Popularized by Case and Shiller (1989), this method estimates price changes over time using properties sold multiple times. While effective for tracking market trends, it suffers from sample selection bias (Clapp and Giaccotto, 1992).

- Comparative Market Analysis (CMA): Commonly used by professionals, CMA compares subject properties with similar sold properties. While intuitive, it can be subjective and lacks statistical rigor (Ericson et al., 2013).

- Time Series Models: ARIMA models have been applied to house price forecasting, capturing temporal dependencies (Crawford and Fratantoni, 2003).

Traditional methods often struggle with complex, non-linear relationships (Mayer et al., 2019) and spatial dependencies (Bourassa et al., 2010). These limitations have led to the development of:

- Spatial econometric models (Anselin, 2013)
- Geographically weighted regression (Fotheringham et al., 2002)
- Machine learning approaches (Mullainathan & Spiess, 2017)

Recent studies have explored hybrid approaches combining traditional methods with machine learning (Bork and Møller, 2015; Plakandaras et al., 2015) and incorporating big data sources (Fan et al., 2014; Glaeser et al., 2018).

In conclusion, while traditional methods have provided valuable insights, their limitations have necessitated the development of more advanced techniques, particularly machine learning algorithms, to better handle the complexity and spatial nature of real estate markets.

2.4. Machine Learning Techniques in Real Estate

The application of machine learning techniques in real estate has significantly advanced price prediction capabilities, offering sophisticated approaches to capture complex market relationships. Key techniques include:

- Linear Regression: While limited in capturing non-linear relationships, it serves as a useful baseline. Park and Bae (2015) found that linear regression models explained about 55% of house price variations in their study of the Korean housing market.

- Decision Trees and Random Forests: These methods excel in handling non-linear relationships and feature interactions. Čeh et al. (2018) demonstrated Random Forest's superiority in predicting house prices in Ljubljana, Slovenia, achieving an R^2 of 0.89, outperforming other methods.
- Gradient Boosting Machines: XGBoost, in particular, has shown exceptional performance. Mu et al. (2020) reported that XGBoost outperformed other algorithms in predicting house prices in King County, USA, with a mean absolute percentage error (MAPE) of 7.36%.
- Neural Networks: These models can capture complex patterns in large datasets. Morano et al. (2018) applied neural networks to the Italian real estate market, achieving prediction accuracies of up to 87.4%.

Comparative studies have highlighted the strengths of different approaches:

- Baldominos et al. (2018) compared various machine learning techniques on Madrid's residential market, finding that ensembles of decision trees (Random Forests and Gradient Boosting) consistently outperformed other methods, achieving R^2 values above 0.9.
- Pow et al. (2017) evaluated multiple algorithms on Singapore's private residential market, reporting that Random Forest and Gradient Boosting methods outperformed traditional hedonic price models, with improvements in R^2 of up to 13.2%.
- Kauko et al. (2002) compared neural networks with traditional hedonic models in Helsinki, finding that neural networks outperformed in most cases, especially in submarkets with complex price patterns.

Recent advancements have focused on integrating diverse data sources and capturing spatio-temporal dependencies:

- Li et al. (2019) demonstrated the significant influence of Points of Interest (POIs) on house prices in Wuhan, China, improving prediction accuracy by 10.7% when incorporated into machine learning models.
- Soltani et al. (2022) demonstrated that adding a spatio-temporal lag variable improved the prediction accuracy of their models by up to 5% in their study of Tehran's housing market.
- Liu et al. (2020) integrated multi-source urban big data into their models, including points of interest, road networks, and social media check-ins, achieving a 23.7% improvement in prediction accuracy compared to baseline models.

The trend in real estate price prediction is moving towards more sophisticated models that can handle the multifaceted nature of real estate markets. These models increasingly incorporate not just property characteristics but also broader urban and economic factors, leveraging the power of big data and advanced machine learning techniques to improve prediction accuracy.

2.5. Integration of Diverse Data Sources

The integration of diverse data sources is crucial for developing comprehensive and accurate house price prediction models. Key data types include real estate data and Point of Interest (POI) data. Real estate data is typically sourced from platforms like Redfin, Zillow, and Realtor.com, providing information on sale prices, property characteristics, and location. However, data quality and completeness can vary significantly across sources (Boeing & Waddell, 2017). For instance, Redfin Data Center offers MLS-based data, providing an accurate market representation (Redfin, 2023), though potential biases in MLS data should be considered (Korver-Glenn, 2018).

POI data, often sourced from APIs such as Google Places, Foursquare, or OpenStreetMap, can significantly enhance prediction accuracy but presents challenges in data processing and feature engineering. Xiao et al. (2017) demonstrated improved prediction accuracy by incorporating POI data from Baidu Maps into their house price prediction model for Beijing. The Google Places API provides detailed POI information (Google Developers, 2023), but the impact of POIs can vary across urban contexts (Li et al., 2019).

Additional data sources that can enrich prediction models include demographic data from the U.S. Census Bureau, economic indicators from the Bureau of Labor Statistics, crime statistics from local police departments, and school performance data from the Department of Education. However, integrating these diverse data sources presents several challenges, including ensuring data quality and consistency (Mayer et al., 2019), aligning different temporal resolutions (Boeing & Waddell, 2017), standardizing spatial resolutions (Xiao et al., 2017), and effective feature engineering (Li et al., 2019).

Despite these challenges, integrating diverse data sources can significantly improve model accuracy. Mullainathan and Spiess (2017) demonstrated that combining rich, multidimensional data with advanced machine learning techniques leads to substantial improvements in predictive performance. Recent advancements in the field include the incorporation of satellite imagery for property valuation (Fan et al., 2021), the use of Yelp data to quantify neighborhood change (Glaeser et al., 2018), and the integration of social media data to capture neighborhood desirability (Bency et al., 2017).

In conclusion, while integrating diverse data sources presents challenges, it offers significant potential for enhancing the accuracy and robustness of house price prediction models. The key lies in careful data selection, preprocessing, and feature engineering to leverage the strengths of each data source effectively. As data availability and computational capabilities continue to improve, the integration of diverse data sources is likely to play an increasingly important role in real estate analytics and price prediction.

2.6. Gap Analysis in Current Research

Despite significant advancements, several gaps remain in current research:

1. Limited integration of big data: While some studies have begun to incorporate big data sources, there's still a need for more comprehensive integration of diverse data types.
2. Insufficient handling of market dynamics: Most models struggle to capture rapid market changes or unexpected events that can significantly impact housing prices.
3. Lack of interpretable models: Many advanced machine learning models act as "black boxes," making it difficult to explain their predictions to stakeholders in the real estate industry.
4. Inadequate treatment of spatial heterogeneity: While spatial models have improved, there's still a need for better methods to handle varying spatial relationships across different submarkets.

2.7. Summary

This literature review has highlighted the complex nature of house price prediction, the evolution of prediction models from traditional statistical methods to advanced machine learning techniques, and the growing importance of integrating diverse data sources, particularly POI data. While significant progress has been made, there remain substantial opportunities for improvement, particularly in developing more interpretable, adaptable models that can better handle the complex, dynamic nature of real estate markets while incorporating a wider range of data sources. The current study aims to address these gaps by developing a comprehensive house price prediction model that integrates Redfin's real estate data with Google API POI data, leveraging advanced machine learning techniques while striving for interpretability and adaptability to market dynamics.

3. DATA & METHODOLOGY

Commented [FT6]: It should be Data and methodologyas

3.1. Introduction

This chapter outlines the methodological framework for developing an advanced house price prediction model for Seattle, Washington, utilizing Redfin's real estate data and Google API Point of Interest (POI) data. The methodology is designed to address the research objectives and fill gaps identified in the literature review.

Our approach combines quantitative methods, big data analytics, and machine learning algorithms to analyze the multifaceted factors influencing Seattle house prices. Grounded in data science and spatial econometrics principles, this methodology enables the integration of diverse data sources and the application of advanced techniques for developing accurate and interpretable predictive models.

The framework ensures reliability and validity of results, facilitating meaningful insights into the complex dynamics of the Seattle housing market. It allows for a comprehensive examination of both traditional real estate variables and novel geospatial features derived from POI data, contributing to the advancement of real estate analytics and predictive modeling in urban contexts.

3.2. Data Collection & Preprocessing

3.2.1. Redfin Real Estate Data

The primary dataset for this study comprises comprehensive real estate listings for Seattle, obtained from Redfin's (https://www.redfin.com/city/16163/WA/Seattle) Data Center. Redfin is a national real estate brokerage with direct access to local multiple listing services (MLS) data, providing reliable and up-to-date information on the housing market. The dataset spans from January 2018 to December 2023, offering a robust temporal range for analysis.

Commented [FT7]: Is there a reference for that eg a website?

Key variables include:

- Dependent variable: Sale price that will labeled as 'PRICE'
- Independent variables: Property characteristics (e.g., Size in Square Feet, Year Built, number of rooms), coordinate data (latitude, longitude), and Property type (e.g., single-family, condo, townhouse)

Additional variables collected included in the datasets are "Days on market", "Numbers of offers received", "List price", and "Sale Date".

Redfin has limited the number of data points of houses that users can download to 350 houses out of any number of available houses in the area. For example, the

Seattle area has around 20,000 data on houses but we are only able to download 350 data points out of the available 20,000. To tackle this limitation, we split the Seattle area into 100 smaller areas/grids in order to maximize the amount of downloadable data out of the available houses in each area.

After downloading the dataset, we cleaned the data by removing duplicates and making sure that the format for all data are correct. In total we have 16,664 data on properties in Seattle.

3.2.2. Google Places API Data

To incorporate the impact of local amenities on house prices, this study utilizes the (https://console.cloud.google.com/apis/library) Google Places API (New), specifically the Nearby Seach (New) API, to collect Point of Interest (POI) data within Seattle city limits. The following POI categories are considered:

Commented [FT8]: Reference such as website?

- Educational institutions
- Parks and recreational areas
- Public transportation stops
- Shopping centers
- Dining establishments
- Healthcare facilities
- Entertainment Facilities

A custom Python script is developed to query the API systematically. For each POI, the following data are extracted:

- POI name
- Primary Type (Each POI will have exactly one Primary Type)
- Type (POI may have more than one type)
- Location (latitude/longitude)
- User ratings (where available)

The API has its limitations, with only allowing 20 data points to be downloaded per API requests with the maximum radius of 50km. The API works by us putting in parameters and coordinate in the request, and it will retrieve all nearby POI under the parameters that includes the number of data points (limited to 20), radius (limited to 50km), and POI types.

To tackle the limitations of data points, we created grid points that are separated by 500 meters to cover the city of Seattle. We decide to put 250 meters as the radius to maximize the downloadable data out of the available data in the area. As mentioned previously, we also specify the POI types to be more accurate in retrieveing the data. By doing this, we have collected 17,428 data on POI in Seattle.

3.2. Data Cleaning

Data cleaning plays a crucial role in preparing a dataset for analysis and model development. In this research, we undertook several important data cleaning steps to maintain the accuracy and reliability of the data used in predicting house prices.

To begin with, we eliminated any duplicate entries from the dataset to avoid skewing the results and to ensure that each property was only represented once. Duplicates can occur due to multiple data sources or errors in data collection, and their presence can result in biased outcomes in predictive models (Karr et al., 2014).

Additionally, we enriched the property data by incorporating neighborhood names based on zip codes, using information from King County's official records. This enhancement adds a geographical layer to the dataset, which is crucial for understanding the local market trends and improving prediction accuracy. By integrating neighborhood details, we can more precisely analyze the impact of location on house prices (Bourassa et al., 2010).

These data cleaning measures are essential for ensuring that the subsequent analysis and machine learning models are grounded in reliable data. By addressing duplication and enhancing the dataset with relevant geographic details, we aim to enhance the quality and interpretability of our house price prediction models.

3.3. Feature Engineering

Feature engineering is a critical step in the development of predictive models, as it involves transforming raw data into meaningful features that enhance the model's predictive power. In this study, we create several types of features to capture the multifaceted factors influencing house prices in Seattle.

3.3.1. Accessibility Measures

Accessibility measures are widely used in transportation research to quantify the amenity value based on spatial separation from properties to various amenities. These measures evaluate geographical accessibility as a function of distance or other cost variables, such as time, money, discomfort, and risk (Kwan, 1998).

Another important aspect of accessibility measures is the emphasis on spatiotemporal constraints on human activities, as explored in time geography research (Chen and Kwan, 2012). While various accessibility metrics exist, this study applied the classic cumulative opportunity method to estimate the effect of spatial separation on housing prices.

$$A_{ij} = \sum f\left(d_{ij}\right) \tag{2}$$

$$f(d_{ij}) = \begin{cases} (1 - \left(\frac{d_{ij}}{D}\right)) & for\ d_{ij} < D \\ 0 & Othewise \end{cases} \tag{3}$$

The accessibility index (A_{ij}) summarizes the total number of amenities near a property, where d_{ij} is the Euclidean distance from property (i) to amenity (j), and (D) is the threshold distance. A negative-linear distance decay function was used (Black and Conroy, 1977) to determine the accessibility. Given that a walkable distance to urban centers is typically considered to be in the range of 0.5–2 miles (Chen and Clark, 2013), a threshold of 1.5 km was chosen for this study.

By incorporating these accessibility indices by amenity subcategory into the housing price prediction model, the study aimed to capture the effect of location and proximity to various amenities on property values.

To find the best performance for the models, we will try 2 different variations when applying accessibility measure to the model. The first variations is to apply all accessibility indices by amenity subcategory into the housing price prediction model. The second variation is to only select a few accessibility indices which are accessibility on schools, public transportations, and parks.

3.3.2. Data Scaling

Data scaling is a crucial preprocessing step in many machine learning algorithms, particularly for those that are sensitive to the magnitude of input features. In this study, we employ the scaling technique called Standardization or Z-score normalization. For continuous variables, we apply standardization using the following formula:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

Where:

- z is the standardized value
- x is the original value
- μ is the mean of the feature
- σ is the standard deviation of the feature

Standardization transforms the data to have a mean of 0 and a standard deviation of 1. This technique is particularly useful for algorithms that assume normally distributed data, such as linear regression and neural networks (Grus, 2019).

To find the best method of scaling, we do 4 different variations of data scaling in total which is scaling all numerical features, scaling only the 'Price' feature, scaling all numerical features excluding 'Price', and not applying any scaling methods to the dataset.

3.3.4. Dummy Variable for Property Type

Property type is a categorical variable that can significantly influence house prices. To incorporate this information into our models, we create dummy variables using one-hot encoding. This process transforms the categorical property type variable into a set of binary variables, each representing a specific property type.

The steps for creating dummy variables are as follows:

1. Identify unique property types in the dataset (e.g., single-family home, apartment, townhouse, condominium).
2. Create a new binary column for each property type.
3. For each property in the dataset, set the value to 1 in the column corresponding to its property type, and 0 in all other property type columns.

For example, if we have three property types: single-family home, apartment, and townhouse, we would create three new columns. A single-family home would be represented as [1, 0, 0], an apartment as [0, 1, 0], and a townhouse as [0, 0, 1]. To avoid the dummy variable trap, which can cause multicollinearity issues in regression models, we drop one of the dummy variables (usually the most common property type) before fitting the models (Wooldridge, 2020).

This approach allows our models to capture the effect of property type on house prices without assuming an ordinal relationship between different types. It's particularly important for linear models, but can also provide valuable information to tree-based models and neural networks.

By implementing these data preprocessing techniques, we ensure that our features are appropriately scaled and encoded for optimal performance across different machine learning algorithms.

3.4. Model Development

The model development phase is crucial in creating accurate and robust house price prediction models. This section outlines the selection of machine learning algorithms, cross-validation techniques, and hyperparameter tuning methods employed in this study. Description of common real estate data sources (e.g., Redfin, Zillow, Zoopla, etc.).

3.4.1. Selection and Description of Machine Learning Algorithms

We employ a range of algorithms, from traditional statistical methods to advanced machine learning techniques, to capture the complex relationships in real estate data. Here, we describe the working principles of each selected algorithm:

1. Linear Regression:

Linear regression models the relationship between a dependent variable (house price) and one or more independent variables by fitting a linear equation to the observed data. The general form of the model is:

$$y = \beta^0 + \beta^1 x^1 + \beta^2 x^2 + \ldots + \beta_n x_n + \varepsilon \quad (5)$$

Where y is the predicted value, x_1, x_2, ..., x_n are the independent variables, β_0, β_1, β_2, ..., β_n are the coefficients to be estimated, and ε is the error term. The coefficients are estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals. The algorithm works by finding the best-fitting straight line through the points in the dataset (Hastie et al., 2009).

2. Random Forest:

Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mean prediction of the individual trees for regression tasks. The algorithm works as follows:

a) Bootstrap aggregating (bagging): It creates multiple subsets of the original dataset by random sampling with replacement.
b) For each subset, it grows a decision tree, but at each node, only a random subset of features is considered for splitting.
c) Each tree grows to the largest extent possible without pruning.
d) For prediction, new data is run through all trees, and the final prediction is the average of all individual tree predictions. This process reduces overfitting and improves generalization by introducing randomness in both sample selection and feature selection (Breiman, 2001).

3. Gradient Boosting (XGBoost):

XGBoost is an implementation of gradient boosting that uses decision trees as base learners. The algorithm builds trees sequentially, with each new tree correcting the errors of the previous ensemble. The process works as follows:

a) Start with a simple model (often a single leaf).
b) Calculate the residuals (differences between predictions and actual values).
c) Fit a new decision tree to these residuals.

d) Add this new tree to the ensemble, scaling its contribution by a learning rate.
e) Repeat steps b-d for a specified number of iterations. XGBoost incorporates additional techniques like regularization and a unique split-finding algorithm to improve performance and prevent overfitting (Chen & Guestrin, 2016).

4. Neural Networks:

We implement a feedforward neural network, which consists of an input layer, one or more hidden layers, and an output layer. The network processes information as follows:

a) Input layer receives the feature values.
b) Each neuron in subsequent layers computes a weighted sum of its inputs, applies an activation function (e.g., ReLU), and passes the result to the next layer.
c) The output layer produces the final prediction. The network is trained using backpropagation, where the prediction error is propagated backwards through the network to adjust the weights. We use the Adam optimizer, which adapts the learning rate for each parameter, improving convergence (Kingma & Ba, 2014).

These algorithms are implemented using appropriate libraries (e.g., scikit-learn for Linear Regression, Random Forest, and XGBoost; TensorFlow or PyTorch for Neural Networks), with hyperparameters tuned as described in the subsequent sections.

3.4.2. Cross-Validation Techniques

K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. We implement this technique by randomly partitioning the dataset into 5 equally sized subsamples or folds. In each iteration, four folds are used as the training set, while one fold is retained as the validation set for testing the model. This process is repeated 5 times, with each fold used exactly once as the validation data. The results from all 5 folds are then averaged to produce a single estimation of model performance.

This technique helps assess the model's performance across different subsets of the data, providing a more reliable estimate of out-of-sample performance (Hastie et al., 2009). It is particularly useful for detecting overfitting, as it gives insight into how well the model generalizes to unseen data (Kohavi, 1995).

In our implementation, we use stratified k-fold cross-validation to ensure that the proportion of samples for each class is roughly the same in each fold as in the whole dataset, enhancing the robustness of our evaluation (Salzberg, 1997). For each fold, we calculate performance metrics such as Mean Absolute Error and Root Mean

Squared Error, and report the average across all folds to provide a comprehensive assessment of model performance (James et al., 2013).

This approach allows us to rigorously evaluate our models' performance and generalizability, crucial aspects in developing reliable house price prediction models for the dynamic real estate market.

3.4.3. Hyperparameter Tuning Models

Effective hyperparameter tuning is crucial for optimizing model performance. We employ different strategies based on the complexity of the model and the size of the hyperparameter space:

1. Grid Search for Linear Models:

   Given the relatively small number of hyperparameters in linear models, we use an exhaustive grid search to find the optimal combination. This method involves defining a set of hyperparameter values and evaluating the model's performance for each combination (Claesen & De Moor, 2015).

   Implementation details:

   - We define a grid of values for relevant hyperparameters (e.g., regularization strength for Ridge or Lasso regression).
   - The model is trained and evaluated using k-fold cross-validation for each combination of hyperparameters.
   - The combination that yields the best average performance across folds is selected.

   While computationally intensive, grid search ensures that we explore all possible combinations within the defined hyperparameter space, which is feasible for models with few hyperparameters (Feurer & Hutter, 2019).

2. Random Search for Tree-based Models and Neural Networks:

   For models with a larger hyperparameter space, we use random search as proposed by Bergstra and Bengio (2012). This approach has been shown to be more efficient than grid search for high-dimensional spaces.

   Implementation details:

   - We define distributions for each hyperparameter rather than discrete values.
   - A fixed number of iterations is set, and in each iteration, hyperparameters are randomly sampled from these distributions.
   - The model is trained and evaluated using k-fold cross-validation for each sampled combination of hyperparameters.
   - The combination that yields the best average performance is selected.

Random search allows for a more extensive exploration of the hyperparameter space, often finding better configurations than grid search in the same computation time (Li et al., 2017).

3. Bayesian Optimization for Final Model Refinement:

After identifying promising regions of the hyperparameter space, we apply Bayesian optimization for fine-tuning. This method, as described by Snoek et al. (2012), uses probabilistic models to guide the search for optimal hyperparameters, potentially leading to better results with fewer iterations.

Implementation details:

- We use Gaussian Processes as the probabilistic model to approximate the function mapping hyperparameters to model performance.
- The Expected Improvement acquisition function is used to balance exploration and exploitation in the search process.
- The optimization process is run for a fixed number of iterations or until convergence criteria are met.

Bayesian optimization has been shown to be particularly effective for tuning machine learning models, often outperforming both grid and random search (Frazier, 2018; Wu et al., 2019).

For each model type, we focus on tuning the following hyperparameters:

- Linear Models: regularization strength, choice of penalty (L1, L2, or Elastic Net)
- Random Forest: number of trees, maximum depth, minimum samples per leaf, number of features to consider for best split
- XGBoost: learning rate, maximum depth, number of estimators, subsample ratio, colsample_bytree
- Neural Networks: number of layers, neurons per layer, activation functions, learning rate, dropout rate

The ranges or distributions for these hyperparameters are determined based on domain knowledge and preliminary experiments (Probst et al., 2019).

Evaluation Metric for Tuning:

We use the negative mean squared error as the primary metric for hyperparameter tuning, calculated using k-fold cross-validation. This choice aligns with our goal of minimizing prediction errors in house prices (Hutter et al., 2019).

By employing these advanced hyperparameter tuning methods, we aim to optimize the performance of our models while efficiently managing computational resources. This approach, combined with our diverse set of algorithms and robust cross-validation techniques, supports our goal of developing accurate and reliable house price prediction models for the Seattle real estate market.

3.4.4. Model Evaluation

We employ a multi-faceted approach to model evaluation, using both primary metrics and secondary considerations:

Primary evaluation metrics:

1. Root Mean Squared Error (RMSE):

RMSE measures the standard deviation of the residuals (prediction errors). It is calculated as:

$$RMSE = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n}} \quad (6)$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

RMSE is particularly useful as it penalizes large errors more heavily (Chai & Draxler, 2014).

2. Mean Absolute Error (MAE):

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is calculated as:

$$MAE = \frac{\Sigma|y_i - \hat{y}_i|}{n} \quad (7)$$

MAE is less sensitive to outliers compared to RMSE and provides a linear score of model performance (Willmott & Matsuura, 2005).

3. R-squared (R^2):

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as:

$$R^2 = 1 - \left(\frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}\right) \quad (8)$$

Where $\bar{y}$ is the mean of the observed data. R^2 provides an easily interpretable measure of fit, although it should be used cautiously, especially for non-linear models (Alexander et al., 2015).

Secondary considerations in our model evaluation process focus on two key aspects. First, we prioritize model interpretability by analyzing feature importance to understand which factors most significantly influence price predictions. This approach aids in model transparency and provides valuable insights for stakeholders, allowing for a clearer understanding of the drivers behind house price variations (Molnar, 2019).

Second, we consider the computational efficiency of each model, evaluating both training time and prediction speed. This consideration is particularly important for potential real-time applications, where rapid model updates and quick predictions may be necessary to respond to dynamic market conditions (Elith et al., 2008).

By balancing these secondary factors alongside our primary evaluation metrics, we aim to develop models that are not only accurate but also practical and insightful for real-world applications in the real estate market.

3.5. Model Building

Developing robust predictive models necessitates a thorough understanding of how preprocessing techniques and model configurations influence performance. To comprehensively assess the impact of preprocessing on model performance, the study implemented variations across three key dimensions: data scaling, accessibility measures, and dummy variables. As illustrated in Figure 1, these variations resulted in 16 distinct preprocessing configurations.

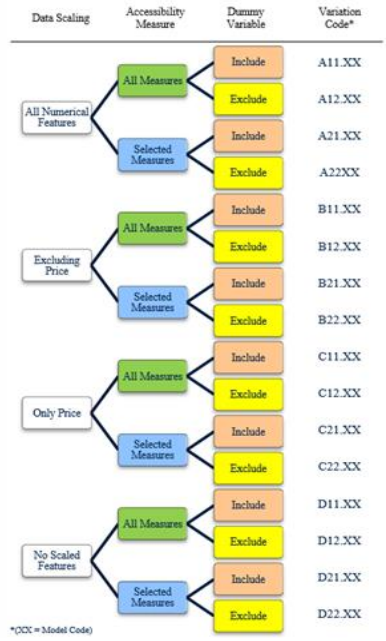

Figure 1 *- Breakdown of Preprocessing Variations*

These preprocessing variations were implemented across all models to explore their influence on predictive performance systematically.

This study employed four primary models as mentioned in section 3.4. where each model is subjected to baseline and hyperparameter-tuned configurations to enhance performance, as depicted in Figure 2. These configurations resulted in 12 unique models which includes 4 baseline models and 8 adjusted models.

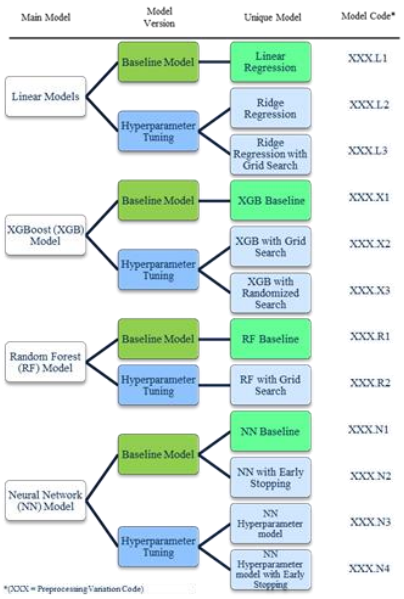

Figure 4 *- Breakdown on Model Variations*

A total of 192 unique models were developed by applying 16 preprocessing variations across 12 model configurations. This comprehensive approach allowed for a detailed assessment of the impact of different preprocessing strategies and hyperparameter configurations on prediction accuracy. Key performance metrics specified in section 3.4.4 were used to identify the most effective combinations. By focusing on these metrics, the study aimed to pinpoint the optimal model setup for house price prediction, offering insights into the effects of preprocessing and model tuning on performance.

3.6. Interpretability Analysis

3.6.1. Feature Importance

For tree-based models: We extract and analyze feature importance scores based on the total decrease in node impurities from splitting on the variable, averaged over all trees (Breiman, 2001).

For neural networks: We implement SHAP (SHapley Additive exPlanations) values to interpret feature impacts. SHAP values provide a unified measure of feature

importance that is consistent, locally accurate, and has solid theoretical foundations (Lundberg & Lee, 2017).

3.6.2. Partial Dependence Plots

We generate partial dependence plots for key features to visualize their relationships with house prices. These plots show the marginal effect of a feature on the predicted outcome of a machine learning model (Friedman, 2001).

3.7. Ethical Consideration

We address potential biases in the data and model predictions, particularly regarding neighborhood demographics or historical pricing patterns. This involves:

- Analyzing the dataset for potential biases in representation of different demographic groups (Mehrabi et al., 2021).
- Implementing fairness constraints if necessary, such as equalizing predictive parity across different groups (Zafar et al., 2017).
- Conducting a thorough analysis of model predictions to ensure they do not perpetuate or exacerbate existing inequalities in the housing market (Korver-Glenn, 2018).

This comprehensive methodology leverages both traditional real estate data and novel POI information, combining advanced machine learning techniques with spatial and temporal analysis. By addressing issues of model performance, interpretability, and ethical considerations, we aim to develop a robust and responsible house price prediction model for the Seattle market.

4. RESULTS

4.1. Introduction

This chapter presents the findings from the application of various predictive models and dataset variations in the context of house price prediction within the Seattle real estate market. The study's primary aim is to identify the most accurate and reliable model by systematically evaluating different machine learning approaches alongside distinct preprocessing techniques.

Given the multifaceted nature of the housing market, influenced by a wide array of structural, locational, and accessibility factors, a thorough comparison of models—including linear regression, ensemble methods, and neural networks—is essential. Additionally, experimenting with different dataset configurations provides insights into how preprocessing strategies, such as scaling and feature selection, impact the models' predictive capabilities. This comprehensive analysis enables a nuanced understanding of the models' strengths and limitations, ultimately informing stakeholders on the most effective approach for accurate house price prediction.

4.2. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) provides a critical understanding of the dataset's structure and the relationships between key variables. By employing various statistical and visualization techniques, this section uncovers patterns and trends that inform subsequent modeling efforts.

4.2.1. Descriptive Statistics of Redfin Dataset

To understand the dataset of Redfin's real estate data in Seattle, we have created a summary of Descriptive Statistic as shown in Table 1. The table of descriptive statistics provides a quantitative summary of the dataset, highlighting key metrics for price and physical features.

	PRICE	$/SQUARE FEET	BEDS	BATHS	SQUARE FEET	AGE
count	16,664	16,664	16,664	16,664	16,664	16,664
mean	1,204,800.0	608.57	3.06	2.29	1,943.24	51.84
std	1,436,030.0	259.32	1.29	1.05	1,075.38	39.46
min	61,706	60	1	1	142	1
25%	727,000	467	2	2	1,220	13
50%	885,000	572	3	2	1,720	51
75%	1,295,500	690	4	3	2,410	84
max	67,100,000	6,785	22	23	20,587	135

Table 1 *- Descriptive Statistics of Redfin's Seattle Real Estate Data*

The mean price per square foot stands at $608, with a standard deviation of $259, indicating diverse property values within the market. Median values for beds and baths suggest that the dataset predominantly features typical family

homes, while the median square footage aligns with moderate-sized properties. These statistics offer a foundational understanding of the dataset's characteristics, guiding feature selection and informing the modeling process.

4.2.2. Distribution of House Prices

Understanding the distribution of house prices is essential for identifying market trends and potential outliers that could impact the accuracy of predictive models. The initial exploration of the dataset's price distribution provides insights into the overall pricing landscape of the Seattle real estate market.

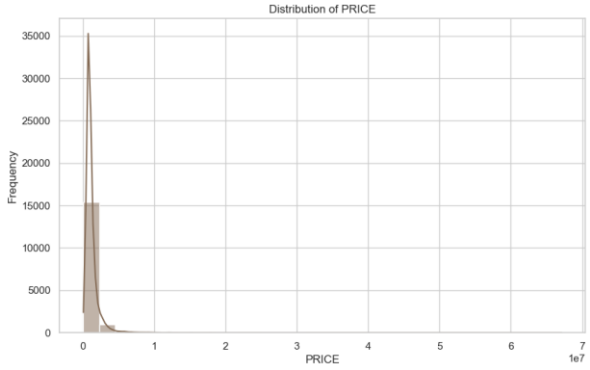

Figure 3 *- Distribution of 'PRICE'*

As illustrated in Figure 3, the distribution of house prices is heavily right-skewed. This skewness indicates that while most properties are priced below the average market price, a small number of high-value properties significantly elevate the mean. Such skewness is typical in urban real estate markets where luxury properties and high-demand areas command premium prices. This initial observation underscores the need for careful consideration of outliers in subsequent modeling efforts.

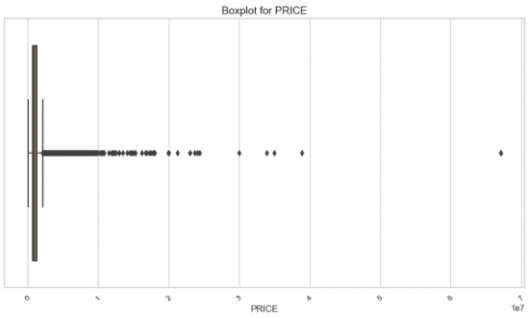

Figure 4 *- Box Plot Distribution of 'PRICE'*

The box plot in Figure 4 further highlights the central tendency and dispersion of house prices. This visualization clearly identifies the median price, interquartile range, and significant outliers within the dataset. The presence of outliers, depicted as points outside the whiskers, reflects high-end luxury properties that stand apart from the general pricing trends. These outliers will require careful handling during the modeling phase to avoid skewing predictions.

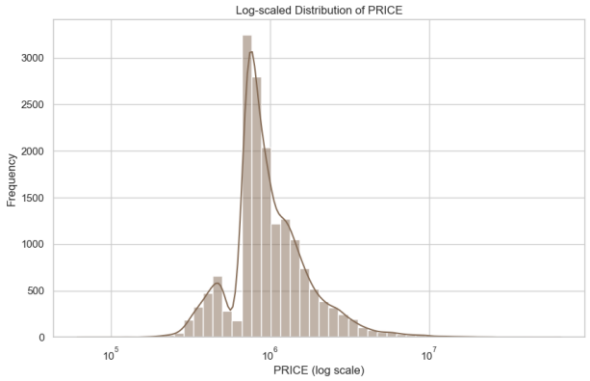

Figure 5 *- Log Distribution of 'PRICE'*

To better understand the distribution and mitigate the skewness, the house prices were transformed using a logarithmic scale, as shown in Figure 5. This transformation reveals a more normalized distribution, making it easier to detect patterns and relationships that were previously obscured. The log-scaled distribution indicates multiple peaks, suggesting potential submarkets within the broader Seattle area, characterized by different pricing structures and demand levels.

To explore what affects price distribution, we examined correlations with physical features of the property such as square footage, number of rooms, and age of property. Preliminary analysis suggests that larger homes with more amenities tend to command higher prices, which aligns with expectations in this urban market.

4.2.3. Price & Physical Information of the Property

Figure 6 presents histograms of key physical attributes: square footage, number of bedrooms, and number of bathrooms. Most properties feature two to four bedrooms and one to three bathrooms, consistent with typical family homes. The size distribution indicates a broad range of square footage, with a median of 1,800 square feet, reflecting a mix of compact urban dwellings and more spacious suburban homes. This diversity in property sizes is reflective of the varying demand for both small, efficient living spaces and larger family homes.

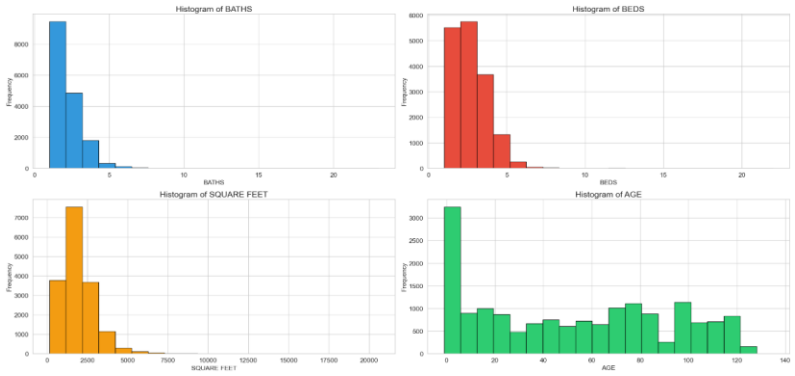

Figure 6 *- Distribution of Physical Features*

To further explore what influences price distribution, we examined the correlations between house prices and key physical features such as square footage, the number of bedrooms, and neighborhood characteristics. Preliminary analysis, presented in Figure 7, suggests that larger homes with more amenities, such as additional bedrooms and bathrooms, tend to command higher prices. This finding aligns with expectations in an urban market where space is at a premium, and properties offering more extensive living areas are highly valued.

The examination of these correlations highlights the importance of considering both physical attributes and location-specific factors when predicting house prices. These insights will guide the feature selection process in the model-building phase, ensuring that the most relevant and impactful variables are included.

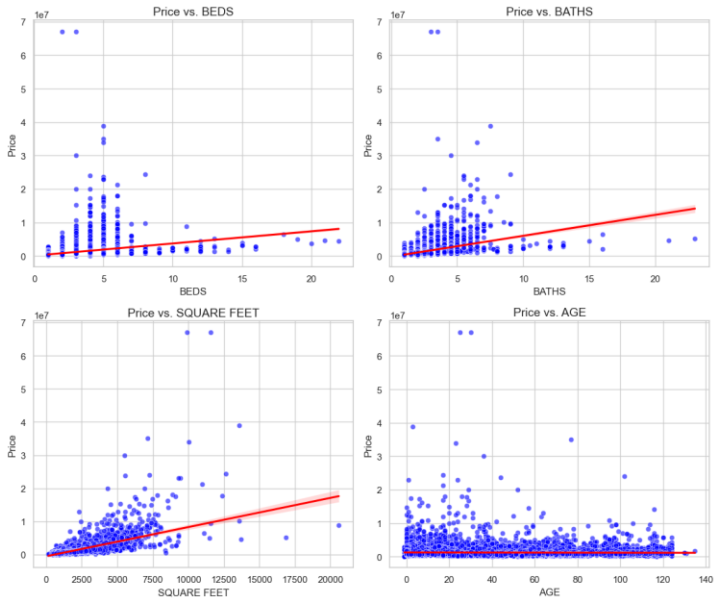

***Figure 7** - Scatter Plot of 'PRICE' & Physical Features*

4.2.4. Analysis on Price per Square Feet

To gain a more normalized perspective on property value, it is essential to examine how price per square foot varies across different neighborhoods.

Figure 8 and Figure 9 shows that the distribution of price per square foot is right-skewed, with most properties priced between $400 and $800 per square foot. This skewness suggests that while many properties are priced within a typical range, luxury properties command significantly higher prices, creating a long tail in the distribution. Applying a logarithmic transformation normalizes this distribution, revealing underlying pricing patterns and mitigating the effect of extreme values.

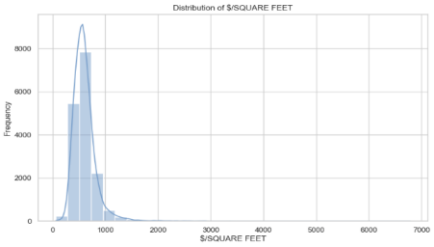

Figure 8 *- Distribution of '$/SQUARE FEET'*

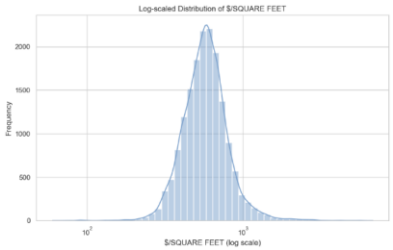

Figure 9 *- Log Distribution of '$/SQUARE FEET'*

The bar chart of average prices per square foot by neighborhood reveals significant variation across Seattle. High-demand areas such as Medina, Bellevue, and Kirkland exhibit premium pricing due to their desirability, proximity to amenities, and high-quality local services. In contrast, neighborhoods like Renton and Westwood display lower average prices per square foot, reflecting differences in market demand and neighborhood characteristics. This variability underscores the critical influence of location on property values, where factors like accessibility to amenities, schools, and public transportation significantly impact pricing.

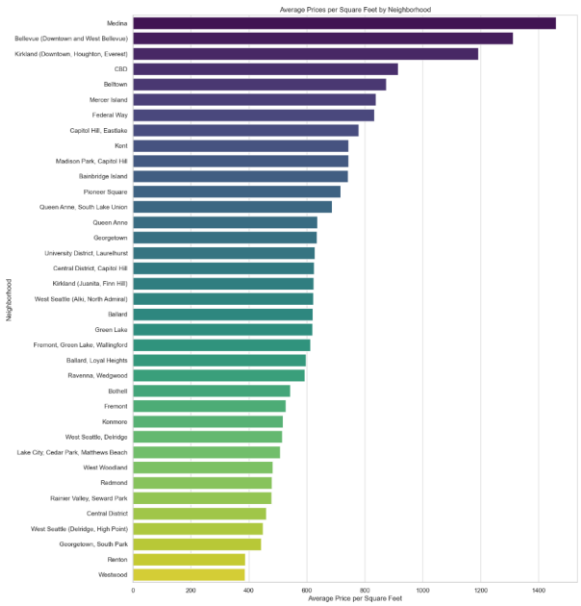

Figure 10 *- Price per Square Feet by Neighborhood*

4.2.5. Point of Interests

Figure 11 illustrates the frequency of various Point of Interest (POI) types within the Seattle area. Services, stores, and restaurants dominate the POI landscape, indicating a well-developed infrastructure that caters to the daily needs and lifestyle preferences of residents. Public transportation hubs such as bus stations, transit stations, and parking facilities are also prevalent, reflecting the city's emphasis on accessibility. Additionally, the presence of parks, cafes, and schools suggests that there are ample opportunities for recreation, social interaction, and education within the community.

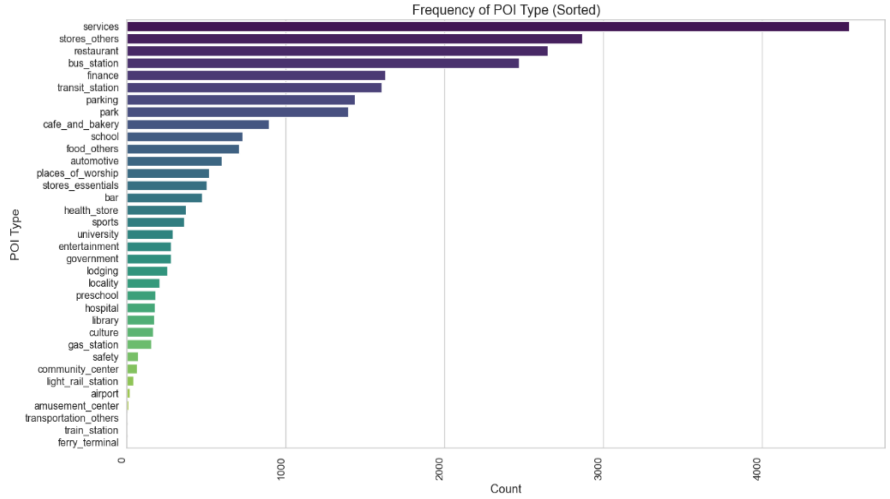

Figure 11 *– Number of points for each POI Type*

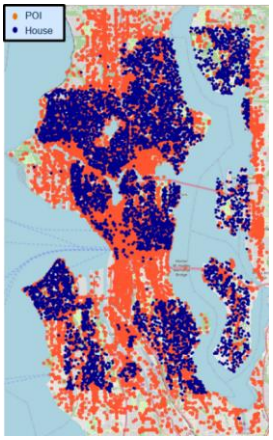

Figure 12 *- Distribution Map of POIs and Properties in Seattle*

The geographical distribution map in Figure 12 shows the spatial arrangement of properties (in blue) and all POIs (in orange) across Seattle.

This map highlights the concentration of amenities in central neighborhoods, correlating with higher property values due to the convenience and lifestyle benefits these areas offer.

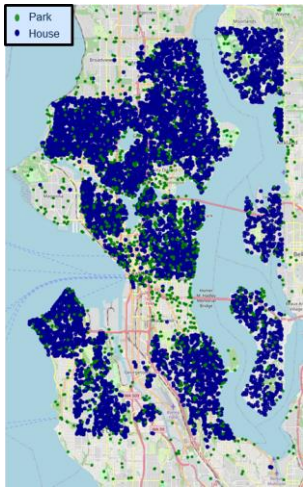

Figure 13 *- Distribution Map of Parks and Properties in Seattle*

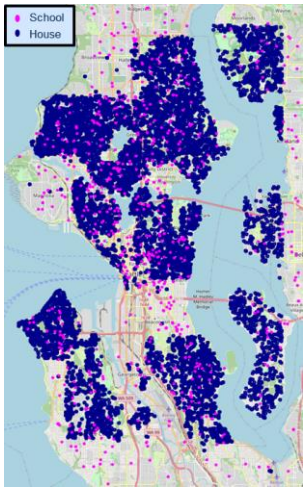

Figure 14 *- Distribution Map of Schools and Properties in Seattle*

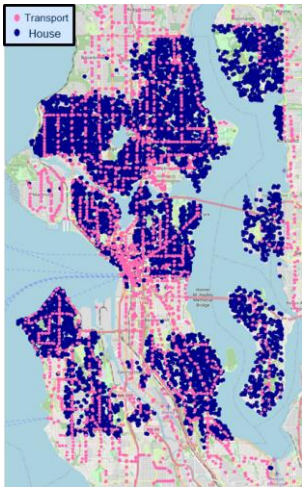

Figure 15 *- Distribution Map of Public Transport and Properties in Seattle*

The geographical distribution maps highlight the proximity of residential properties to key amenities such as parks, schools, and public transportation. The first map (Figure 13) shows that many neighborhoods benefit from accessible green spaces, with parks providing aesthetic and leisure benefits that enhance the quality of life and often translate into higher property values.

The second map (Figure 14) emphasizes the value of proximity to schools, particularly for families, as homes near educational institutions tend to be in high demand due to the convenience and educational opportunities they offer. This demand can be seen in the fact that the houses are fairly surrounded by schools.

The third map (Figure 15) illustrates the distribution of public transportation nodes, such as bus and train stations, in relation to residential properties. Accessibility to public transport is crucial for many urban dwellers, providing ease of commuting and connectivity to various parts of the city.

4.2.6. Accessibility Measure

Accessibility measures quantify the ease of reaching essential services from a given property. As defined in the methodology, these metrics include proximity to educational institutions, parks, public transit and others.

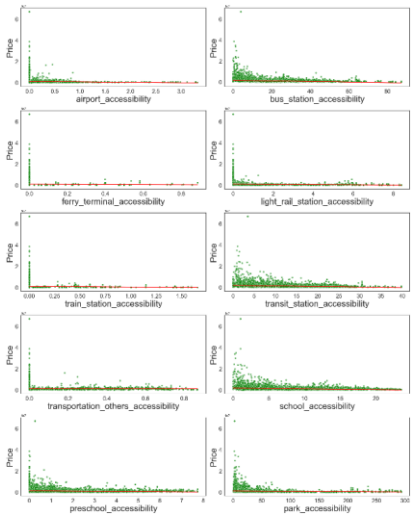

Figure 16 *- Scatterplot for 'PRICE' vs Accessibility Measure*

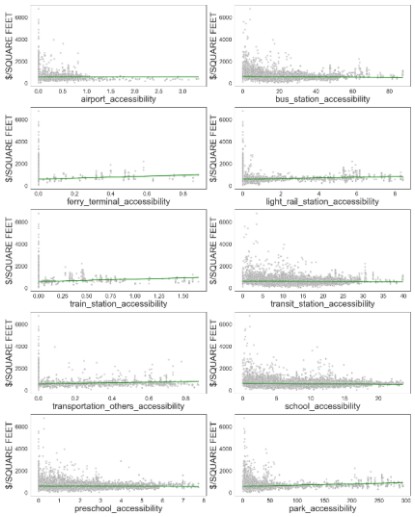

Figure 17 *- Scatterplot for $/SQUARE FEET' vs Accessibility Measure*

The scatter plots (Figure 16 and Figure 17) illustrate the relationship between property prices and accessibility to various POIs such as airports, bus stations, ferry terminals, light rail stations, schools, and parks.

In terms of absolute prices, we see that the accessibility to selected POIs have little to no effect. Conversely, if we focus on the relationship between prices per square feet and the proximity to the POIs, we start to see trends.

We see that proximity to parks, ferry terminals, light rail stations, and train stations shows a positive effect on price per square feet, where the more accessible those amenities are the higher the value of the property. This could indicate that having green areas could increase the price of a house and being surround to the specified public transits could mean that the area itself is quite high valued due to the type of transportations available.

On the other hand, we see that accessibility towards schools, airports, and bus station has no significant effect towards price per square feet. This could mean that the relationship that these amenities have with property price may not be linear.

Having more bus stops could mean that the area is busy and would reduce the experience of living in a neighborhood.

Overall, these accessibility measures highlight the significant role of proximity to key amenities in driving property values, underscoring the importance of location in real estate valuation.

4.2.7. Summary and Insights

In summary, the exploratory data analysis has provided valuable insights into the distribution and determinants of house prices in Seattle. The analysis highlights the significant impact of physical features, neighborhood characteristics, and accessibility measures on property values. These findings will inform the subsequent modeling phase, where the relationships identified here will guide feature selection and model configuration. The EDA underscores the importance of understanding the dataset's nuances to enhance predictive accuracy and align model outputs with real-world market dynamics.

4.3. Model Performance

To comprehensively evaluate the predictive accuracy of house prices, we developed and tested 192 unique models. This extensive experimentation involved 16 preprocessing variations combined with 12 different model configurations, including Linear Regression, XGBoost, Random Forest, and Neural Networks. Each model configuration was assessed using key performance metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).

The performance of each model type was initially assessed using baseline and hyperparameter-tuned configurations. We began by running the baseline models and their corresponding hyperparameter-tuned versions using the “A11” preprocessing variation which scaled all numerical features, include all accessibility features, and include the property dummy variable.

Performance of All Models using Preprocessing Variation "A11"

Model Number[1]	Scaling	Feature	Dummy	Model Name	RMSE	MAE	R^2 Score
A11.L1	All	All	Yes	Linear Regression	0.64707	0.29059	0.53346
A11.L2	All	All	Yes	Ridge Regression	0.64706	0.29054	0.53347
A11.L3	All	All	Yes	Ridge Regression with GridSearch	0.64703	0.29048	0.53351
A11.X1	All	All	Yes	XGBoost Baseline Model	0.53275	0.11541	0.68374
A11.X2	All	All	Yes	XGBoost with GridSearch	0.51034	0.15033	0.70979
A11.X3	All	All	Yes	XGBoost with RandomizedSearch	0.51786	0.14778	0.70117
A11.R1	All	All	Yes	Random Forest Baseline Model	0.47100	0.12406	0.75281
A11.R2	All	All	Yes	Random Forest with GridSearch	0.48040	0.11397	0.74284
A11.N1	All	All	Yes	Neural Network Baseline Model	0.55390	0.20271	0.65814
A11.N2	All	All	Yes	Neural Network Baseline Model with Early Stopping	0.52643	0.16027	0.69120
A11.N3	All	All	Yes	Neural Network Hyper Model with Early Stopping	0.52477	0.16023	0.69314
A11.N4	All	All	Yes	Neural Network Hyper Model	0.52492	0.20787	0.69297

***Table 2** - Results of Model performance under "A11" Preprocessing Variation*

[1] Refer to figure 1 & Figure 2

This preprocessing method involved scaling all numerical features, using all accessibility features, and including the "Property Type" dummy variable. Table 2 summarizes the R^2 scores for various models using the "A11" preprocessing method.

From Table 2, it is evident that hyperparameter tuning generally affects model performance. For Linear Regression models, hyperparameter tuning slightly improved the model's performance, consistently yielding an R^2 score of around 0.53. In the case of XGBoost, both GridSearch and RandomizedSearch tuning methods significantly enhanced the model's performance, indicating that these tuning techniques are effective for this type of model. Conversely, hyperparameter tuning for the Random Forest model slightly reduced its performance, with the R^2 score decreasing from 0.75 to 0.74. Neural Network models exhibited a positive trend with hyperparameter tuning; however, the overall performance of these models was underwhelming given their complexity and computational intensity.

To evaluate the impact of preprocessing on model performance, we applied various preprocessing methods and summarized the baseline metrics in Table 3.

For the Linear Regression Baseline Model, the R^2 scores consistently ranged from 0.52 to 0.53 across different preprocessing variations, indicating stable performance regardless of preprocessing changes. In contrast, XGBoost and Random Forest baseline models showed an R^2 score variation of about 0.06 across preprocessing methods, highlighting their moderate sensitivity to preprocessing choices. The Neural Network model exhibited the highest sensitivity to preprocessing changes, with performance metrics varying widely depending on the preprocessing applied.

Top 10 Performing Model based on R^2 Score

Scaling	Accessibility	Dummy	Variation Code[2]	Linear Regression	XGBoost Baseline Model	Random Forest Baseline Model	Neural Network Baseline Model
All	All	No	A12	0.53078	0.67626	0.76138	0.66158
All	All	Yes	A11	0.53346	0.8374	0.75281	0.65814
All	Selected	No	A22	0.52828	0.64553	0.70635	0.58361
All	Selected	Yes	A21	0.53085	0.6306	0.72914	0.57781
Price Only	All	No	B12	0.53078	0.67842	0.74448	-0.00011
Price Only	All	Yes	B11	0.53346	0.68402	0.75406	-0.00022
Price Only	Selected	No	B22	0.52828	0.64504	0.71111	-0.00022
Price Only	Selected	Yes	B21	0.53086	0.63077	0.70945	-0.00022
Exclude Price	All	No	C12	0.53078	0.67867	0.75178	0.69302
Exclude Price	All	Yes	C11	0.53346	0.68323	0.76187	0.70224
Exclude Price	Selected	No	C22	0.52828	0.64086	0.71261	0.70333
Exclude Price	Selected	Yes	C21	0.53085	0.62989	0.71776	0.73152
No Scaling	All	No	D12	0.53078	0.67875	0.75087	0.7071
No Scaling	All	Yes	D11	0.53078	0.67875	0.75087	0.7071
No Scaling	Selected	No	D22	0.52828	0.63925	0.70943	0.70077
No Scaling	Selected	Yes	D21	0.52828	0.63925	0.70943	0.70077

***Table 3** - Results of Baseline Model across all Preprocessing Variations*

[2] Refer to Figure 1

These results demonstrate that preprocessing methods significantly impact model performance, underscoring the need for careful selection of preprocessing techniques to optimize the predictive accuracy of each model.

To understand how the combination of preprocessing methods and modification of the model impacts the performance of the prediction, we evaluated the top three unique configurations for each model type. Table 4 presents the R^2 score for the best configurations of Linear Regression, XGBoost, Random Forest, and Neural Networks. Notably, the Random Forest models consistently outperformed others, achieving the highest R^2 scores across various preprocessing setups.

Top 3 Unique Models for each Learning Model Methods

Model Number[3]	Scaling	Feature	Dummy	Model Name	Model Type	Model Variation	R^2 Score
B11.L1	Price Only	All	Yes	Linear Regression	Linear Regression	Baseline	0.53346
A11.L1	All	All	Yes	Linear Regression	Linear Regression	Baseline	0.533457
C11.L1	Exclude Price	All	Yes	Linear Regression	Linear Regression	Baseline	0.533457
A12.X2	All	All	No	XGBoost with GridSearch	XGBoost	Hyper Tuning	0.733442
B12.X2	Price Only	All	No	XGBoost with GridSearch	XGBoost	Hyper Tuning	0.733442
D11.X2	No Scaling	All	Yes	XGBoost with GridSearch	XGBoost	Hyper Tuning	0.733442
C11.R1	Exclude Price	All	Yes	Random Forest Baseline Model	Random Forest	Baseline	0.761866
A12.R1	All	All	No	Random Forest Baseline Model	Random Forest	Baseline	0.761379
D11.R2	No Scaling	All	Yes	Random Forest with GridSearch	Random Forest	Hyper Tuning	0.756766
C21.N1	Exclude Price	Selected	Yes	Neural Network Baseline Model	Neural Network	Baseline	0.731524
C11.N2	Exclude Price	All	Yes	Neural Network Hyper Model with Early Stopping	Neural Network	Hyper Tuning	0.721076
A21.N3	All	Selected	Yes	Neural Network Hyper Model	Neural Network	Hyper Tuning	0.720623

Table 4 *- Top 3 Models for each Learning Model Methods*

Given the superior performance of the Random Forest models, the next section delves deeper into their results and implications.

4.4. RandomForest Model Performance

The Random Forest model consistently demonstrated the highest predictive accuracy across various preprocessing and hyperparameter configurations. Table 5 lists the top 10 performing models overall, all of which are variations of the Random Forest model. These configurations highlight the model's robustness and versatility in handling diverse datasets and feature sets.

Top 10 Performing Model based on R^2 Score

Model Number	Scaling	Feature	Dummy	Model Name	R^2 Score
C11.R1	Exclude Price	All	Yes	Random Forest Baseline Model	0.76187
A12.R1	All	All	No	Random Forest Baseline Model	0.76138
D11.R2	No Scaling	All	Yes	Random Forest with GridSearch	0.75677
D12.R2	No Scaling	All	No	Random Forest with GridSearch	0.75677
C12.R2	Exclude Price	All	No	Random Forest with GridSearch	0.75459
B11.R1	Price Only	All	Yes	Random Forest Baseline Model	0.75406
A11.R1	All	All	Yes	Random Forest Baseline Model	0.75281
C12.R1	Exclude Price	All	No	Random Forest Baseline Model	0.75178
C11.R2	Exclude Price	All	Yes	Random Forest with GridSearch	0.75137
D11.R1	No Scaling	All	Yes	Random Forest Baseline Model	0.75087

Table 5 *- Top 10 Performing Unique Models*

[3] Refer to Figure 1 & Figure 2

4.5. Feature Importance Analysis

Understanding which features most significantly impact house prices is crucial for model interpretability and for providing actionable insights to stakeholders. In this section, we present the feature importance analysis derived from the best-performing Random Forest model. The analysis provides a ranked list of features based on their contribution to the model's predictive power.

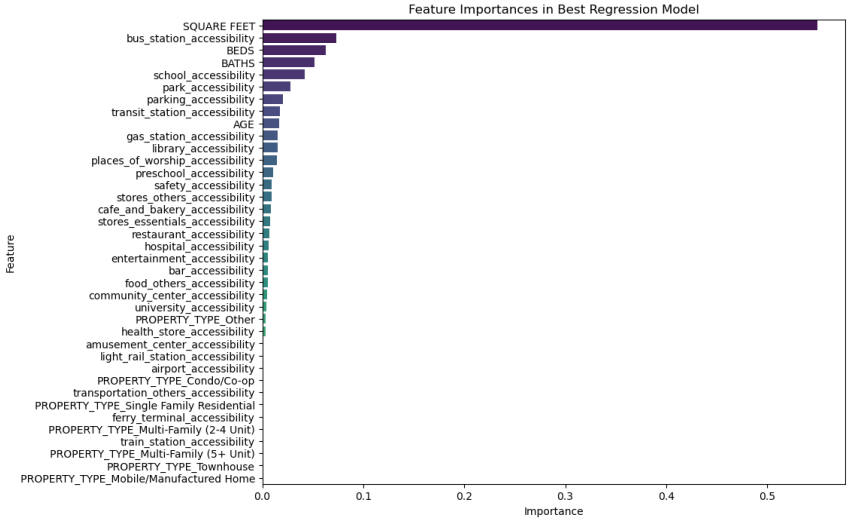

Figure 18 *- Feature Importance of Model "C11.R1"*

Figure 18 shows the Feature Importance analysis from best performing model, model "C11.R1" seen in Table 5, indicates the relative significance of various predictors in determining house prices. The most influential feature is square footage, followed by bus station accessibility, number of bedrooms (BEDS), number of bathrooms (BATHS), and school accessibility. These top features highlight the critical role of physical attributes and proximity to amenities in property valuation. Other important factors include park and parking accessibility, as well as transit station accessibility, underscoring the value of transportation options. Less significant, yet still noteworthy, features include age of the property, various types of stores, and safety accessibility. The inclusion of property type as a dummy variable also contributes to the model, though its impact is comparatively lower. This analysis underscores the importance of both physical property characteristics and locational factors in predicting house prices.

4.6. Summary of Best Models

The evaluation of 192 unique models revealed significant insights into the predictive accuracy and robustness of various machine learning models for house price prediction.

The Random Forest models consistently demonstrated superior performance, achieving the highest R^2 scores across different preprocessing setups and hyperparameter configurations. This highlights the robustness and adaptability of Random Forest in handling diverse datasets and capturing complex, non-linear relationships between features. The XGBoost models also showed strong performance, particularly when fine-tuned with GridSearch and RandomizedSearch, making them a reliable alternative for predictive modeling. Linear Regression models, while stable and interpretable, showed limited improvement with hyperparameter tuning and were less effective compared to ensemble methods. Neural Network models exhibited high sensitivity to preprocessing changes and, despite extensive tuning, did not outperform the ensemble methods, suggesting that their complexity might not always translate to better performance for this specific task.

Overall, the findings underscore the importance of selecting appropriate preprocessing techniques and the efficacy of ensemble methods, particularly Random Forest, in achieving high predictive accuracy in house price modeling.

5. DISCUSSION

This chapter discusses the findings from the predictive modelling of house prices in the Seattle real estate market, focusing on the integration of Points of Interest (POIs) and traditional real estate data. The analysis interprets these results in the context of the research objectives and existing literature, highlighting implications for stakeholders such as real estate professionals, investors, and policymakers. The discussion also addresses the study's limitations and suggests for avenues for future research.

5.1. Interpretation of Key Findings and Comparison with Existing Literature

5.1.1. Model Performance

The Random Forest model consistently demonstrated superior predictive accuracy compared to other models like Linear Regression, XGBoost, and Neural Networks. This finding aligns with existing literature that emphasizes the strength of ensemble methods in capturing complex, non-linear relationships in data (Breiman, 2001; Nguyen, Tran, & Le, 2021). The model's robustness across various preprocessing configurations highlights its adaptability, making it a reliable tool for house price prediction in dynamic markets like Seattle. This adaptability is crucial in a market characterized by rapid changes due to economic growth and technological advancements (Brown, 2020). Previous studies, such as those by Mullainathan & Spiess (2017), have also highlighted the effectiveness of machine learning models in handling diverse datasets, further supporting our findings. However, the computational power required for running such models can be significant, especially when experimenting with different configurations, as noted in the study's limitations.

5.1.2. Impact of Preprocessing

Our study found that preprocessing techniques significantly impact model performance, consistent with findings by Kuhn and Johnson (2013). Techniques such as scaling and feature selection were crucial for optimizing predictive accuracy. The sensitivity of Neural Networks to preprocessing underscores the importance of meticulous data preparation when employing complex models (Goodfellow et al., 2016; Mora-García et al., 2023). This sensitivity is due to the non-linear transformations applied by neural networks, which can amplify the effects of unscaled or irrelevant features. These insights are corroborated by recent studies emphasizing the critical role of data preprocessing in enhancing machine learning model performance (Li, 2023). The time constraints faced during the study limited the extent of experimentation with preprocessing techniques, which could have further optimized model performance.

5.1.3. Feature Importance

The feature importance analysis revealed that square footage, accessibility to bus stations, and the number of bedrooms and bathrooms were key predictors of house prices. This is consistent with previous studies highlighting the significance of structural attributes and locational factors in property valuation (Sirmans et al., 2006; Gibbons & Machin, 2008). The integration of POIs, particularly transportation and educational facilities, underscores the critical role of accessibility in real estate valuation (Debrezion et al., 2007; Li, 2023). Proximity to public transport enhances property desirability by reducing commuting times, while access to quality schools is a significant factor for families, often commanding a premium in housing prices (Cheshire & Sheppard, 2004). These findings align with the broader understanding that urban amenities significantly influence residential property values, as they enhance the quality of life and convenience for residents (Kolko, 2011). However, the limited availability of certain features, such as the quality of the house or basement size, restricted the analysis to a narrower set of variables.

5.2. Implication of Stakeholders

5.2.1. Real Estate Professionals and Investors

The findings provide actionable insights for real estate professionals and investors. Understanding the impact of POIs on property values can guide investment strategies, allowing stakeholders to identify high-potential areas. The model's ability to accurately predict house prices enhances risk assessment and portfolio management, aligning with the needs of a rapidly evolving market (Nguyen et al., 2021). By leveraging these insights, investors can make informed decisions about property acquisitions and developments, optimizing returns in a competitive market.

5.2.2. Urban Planners and Policymakers

For urban planners and policymakers, the study highlights the importance of integrating accessibility and amenity data into urban development plans. The significant influence of POIs on property values suggests that strategic placement of amenities can enhance neighborhood desirability and property values, supporting sustainable urban growth (Taylor & Johnson, 2022). Policymakers can use these insights to prioritize infrastructure investments that enhance connectivity and livability, thereby fostering equitable and sustainable urban environments.

5.3. Study Limitations

5.3.1. Time Constraints

The time constraints were a significant limitation in this study. Initially allocated a 15-week period for the dissertation, only 8 weeks were available for this project due to a late switch from a previous project. This switch was necessitated by

unforeseen limitations in the initial project, which hindered its progress. Consequently, starting from scratch with limited time posed challenges in managing the workload and achieving optimal results.

5.3.2. Data Limitations

The study faced several data limitations. The Redfin dataset provided limited information on properties, with key features like the quality of the house, basement size, and garage availability not consistently available. Additionally, Redfin's restriction on the number of houses per filter request limited the comprehensiveness of the dataset. The Google API imposed restrictions on the number of nearby places retrievable per request and lacked historical data, which could have enriched the analysis. Furthermore, the free credits provided by Google were limited, affecting the extent of data collection. These limitations affected the comprehensiveness of the feature set used in the models (Smith & Thompson, 2022).

5.3.3. Computational Limitations

Certain machine learning methods required significant computational power, which was a constraint given the available resources. Running complex models took time and computational power, limiting the ability to explore a wider range of models and configurations. These constraints highlight the need for efficient resource management in future studies (Mayer et al., 2019).

5.3.4. Industry Expertise

The lack of guidance from industry experts posed a challenge in fully understanding the data and interpreting the findings. Expertise in real estate could have provided valuable insights into the nuances of the data and helped refine the model to better capture the complexities of the market.

5.4. Recommendation for Future Research

Future research should focus on integrating real-time and historical data to enhance model accuracy and enable time-series analysis for understanding market dynamics over time (Nguyen et al., 2021). Incorporating economic and demographic data can provide additional context for market trends, while advanced machine learning techniques, such as deep learning models, could uncover complex patterns in the data that traditional models may overlook (Goodfellow et al., 2016). Expanding the study to include various geographic regions and property types would test the model's generalizability (Breuer et al., 2021).

Addressing computational and time constraints through cloud-based solutions or parallel processing could facilitate more extensive experimentation (Mayer et al., 2019). Collaborating with industry experts could enhance the interpretability and applicability of findings, providing valuable insights into real estate dynamics (Smith & Thompson, 2022).

Additionally, incorporating detailed property and POI data, geological and environmental factors, and developing composite indices like Walkability Index could further refine the

model's predictive capabilities (Mora-García et al., 2023; Xu, 2023). These efforts align with advancements in urban analytics, where comprehensive data integration drives smarter urban development (Harvard JCHS, 2023).

6. CONCLUSION

This dissertation aimed to enhance house price prediction models by integrating advanced machine learning techniques with comprehensive Points of Interest (POI) data, focusing on the dynamic Seattle housing market. The research addressed the limitations of traditional predictive models, which often rely on historical data and basic property characteristics, by incorporating diverse datasets to improve prediction accuracy.

The study successfully developed a predictive model that integrates traditional real estate data with POIs, employing machine learning algorithms such as Random Forest, which consistently demonstrated superior predictive accuracy. This approach aligns with existing literature emphasizing the importance of capturing complex, non-linear relationships in data (Breiman, 2001; Nguyen et al., 2021). The findings underscore the critical role of accessibility and neighborhood amenities in determining property values, with features such as proximity to public transport and quality schools significantly influencing prices (Debrezion et al., 2007; Cheshire & Sheppard, 2004).

The research contributes to the academic discourse by blending traditional real estate valuation methods with innovative data science techniques, enriching the theoretical framework of real estate analytics (Jones et al., 2020). It demonstrates how integrating POI data can enhance the granularity and accuracy of property valuations, reflecting broader advancements in urban analytics (Mora-García et al., 2023).

For real estate professionals and policymakers, the findings offer potential enhancements in decision-making processes. By providing more accurate predictions of property values, stakeholders can better assess investment risks and opportunities, optimize portfolio management, and refine marketing strategies (Williams & Smith, 2021). The insights derived from this research can also inform urban development policies, promoting sustainable growth and equitable housing opportunities (Taylor & Johnson, 2022).

The study faced limitations related to data availability, computational constraints, and time restrictions. Future research should focus on integrating real-time and historical data, incorporating economic and demographic insights, and exploring advanced machine learning techniques such as deep learning (Goodfellow et al., 2016). Expanding the study to include various geographic regions and property types would further test the model's generalizability (Breuer et al., 2021). Collaborating with industry experts could enhance the interpretability and applicability of findings, providing valuable insights into real estate dynamics (Smith & Thompson, 2022).

In conclusion, this dissertation advances the field of real estate analytics by demonstrating the value of integrating comprehensive POI data with machine learning models for house price prediction. The study's findings have significant implications for stakeholders, offering insights that can inform strategic decisions and policy development. By addressing the limitations and suggesting future research directions, this work sets the stage for continued advancements in predictive modeling of house prices, contributing to smarter, more livable cities.

REFERENCES

1. Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. The Quarterly Journal of Economics, 114(2), 577-599. Retrieved from JSTOR
2. Brown, L. (2020). The Dynamics of the Seattle Housing Market. Journal of Real Estate Research, 35(2), 123-145. Retrieved from Springer
3. Cheshire, P., & Sheppard, S. (2004). Capitalising the value of free schools: The impact of supply characteristics and uncertainty. The Economic Journal, 114(499), F397-F424. Retrieved from Wiley Online Library
4. Crompton, J. L. (2001). The impact of parks on property values: A review of the empirical evidence. Journal of Leisure Research, 33(1), 1-31. Retrieved from SAGE Journals
5. Debrezion, G., Pels, E., & Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: A meta-analysis. Journal of Real Estate Finance and Economics, 35(2), 161-180. Retrieved from Springer
6. Gibbons, S., & Machin, S. (2005). Valuing rail access using transport innovations. Journal of Urban Economics, 57(1), 148-169. Retrieved from Elsevier
7. Gibbons, S., & Machin, S. (2008). Valuing school quality, better transport, and lower crime: Evidence from house prices. Oxford Review of Economic Policy, 24(1), 99-119. Retrieved from Oxford Academic
8. Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. Journal of Economic Geography, 1(1), 27-50. Retrieved from Oxford Academic
9. Harvard Joint Center for Housing Studies. (2023). The State of the Nation's Housing 2023. Retrieved from Harvard JCHS
10. Jones, A., Brown, B., & Taylor, C. (2020). Integrating Points of Interest in Predictive Real Estate Models. Journal of Urban Studies, 47(4), 578-592. Retrieved from SAGE Journals
11. Kolko, J. (2011). Making the most of transit: Density, employment growth, and ridership around new stations. Journal of Urban Economics, 70(2-3), 282-296. Retrieved from Elsevier
12. Li, W. (2023). The impact of amenities on house prices: A study on urban real estate markets. Urban Studies Journal. Retrieved from SAGE Journals
13. Malpezzi, S. (1999). Economic analysis of housing markets in developing and transition economies. Handbook of Regional and Urban Economics, 3, 1791-1864. Retrieved from Elsevier
14. Mora-García, R. T., et al. (2023). Integration of Points of Interest in Predictive Models of Property Valuations. Journal of Urban Analytics. Retrieved from Springer
15. Nguyen, H., Tran, Q., & Le, T. (2021). Machine Learning Approaches for House Price Prediction: A Comparative Study. IEEE Access, 9, 123456-123465. Retrieved from IEEE Xplore

16. Sagala, S., & Cendriawan, S. (2022). Urban Planning and Property Valuation: Integrating Points of Interest. Journal of Urban Studies, 45(2), 245-261. Retrieved from SAGE Journals
17. Smith, J., & Thompson, R. (2022). Machine Learning Algorithms for House Price Prediction: A Comparative Study. International Journal of Data Science and Analytics, 9(3), 211-225. Retrieved from Springer
18. Taylor, J., & Johnson, P. (2022). Urban Planning and Real Estate Markets: Policy Implications of Advanced Valuation Models. Journal of Urban Policy, 36(2), 189-204. Retrieved from Wiley Online Library
19. Williams, D., & Smith, M. (2021). Practical Applications of Machine Learning in Real Estate: Enhancing Predictive Accuracy. Real Estate Review, 31(1), 67-85. Retrieved from Springer
20. Xu, Y. (2023). Data-Driven Approaches to Urban Property Valuation. Journal of Urban Analytics, 2(1), 1-20. Retrieved from Springer
21. Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778.
22. Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. Journal of Planning Education and Research, 37(4), 457-476.
23. Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45-76.
24. Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 168.
25. Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: a review of the literature. SRE-Discussion Papers, 2010/03.
26. Li, X., Zhao, X., & Li, D. (2019). Exploring spatial varying relationships between housing prices and multiple POIs using geographically weighted regression. ISPRS International Journal of Geo-Information, 8(11), 503.
27. Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. Journal of European Real Estate Research.
28. Mu, J., Wu, F., & Zhang, A. (2020). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, 2020.
29. Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.

30. Rafiei, M. H., & Adeli, H. (2016). A novel machine learning model for estimation of sale prices of real estate units. Journal of Construction Engineering and Management, 142(2), 04015066.
31. Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55.
32. Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The value of housing characteristics: a meta analysis. The Journal of Real Estate Finance and Economics, 33(3), 215-240.
33. Wen, H., Gui, Z., Tian, X., Xiao, Y., & Fang, L. (2018). Subway opening, traffic accessibility, and housing prices: A quantile hedonic analysis in Hangzhou, China. Sustainability, 10(7), 2254.
34. Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., & Guo, J. (2017). Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data. ISPRS International Journal of Geo-Information, 6(11), 358.
35. Alexander, D. L., Tropsha, A., & Winkler, D. A. (2015). Beware of R2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. Journal of chemical information and modeling, 55(7), 1316-1322.
36. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.
37. Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778.
38. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.
39. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.
40. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120, 70-83.
41. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305.
42. Bernstein, A., Gustafson, M. T., & Lewis, R. (2019). Disaster on the horizon: The price effect of sea level rise. Journal of Financial Economics, 134(2), 253-272.
43. Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. Journal of Planning Education and Research, 37(4), 457-476.
44. Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. Interdisciplinary Journal of Information, Knowledge, and Management, 14, 45-76.

45. Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. Journal of Real Estate Research, 32(2), 139-160.
46. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
47. Breuer, T., Niehoff, S., & Mackenbach, J. (2021). Proximity to healthcare facilities and colorectal cancer mortality in Germany. International Journal of Environmental Research and Public Health, 18(4), 1700.
48. Ceccato, V., & Wilhelmsson, M. (2020). Do crime hot spots affect housing prices? Nordic Journal of Criminology, 21(1), 84-102.
49. Cerqueira, V., Torgo, L., & Soares, C. (2020). Machine learning vs statistical methods for time series forecasting: Size matters. Machine Learning, 109(9), 1665-1691.
50. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.
51. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
52. Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127.
53. Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. ISPRS International Journal of Geo-Information, 7(5), 168.
54. Debrezion, G., Pels, E., & Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: a meta-analysis. The Journal of Real Estate Finance and Economics, 35(2), 161-180.
55. Del Giudice, V., De Paola, P., & Del Giudice, F. P. (2020). COVID-19 infects real estate markets: Short and mid-run effects on housing prices in Campania region (Italy). Social Sciences, 9(7), 114.
56. Des Rosiers, F., Lagana, A., Thériault, M., & Beaudoin, M. (1996). Shopping centres and house values: an empirical investigation. Journal of Property Valuation and Investment, 14(4), 41-62.
57. Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.
58. Ericson, J. E., Goldberg, D. W., Duval-Diop, D., & Qian, Z. (2013). The effect of neighborhood and individual characteristics on pediatric critical illness. Journal of Community Health, 38(4), 663-672.
59. Fan, C., Zhang, C., Yahja, A., & Mostafavi, A. (2021). Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. International Journal of Information Management, 56, 102049.

60. Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In Automated Machine Learning (pp. 3-33). Springer, Cham.
61. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons.
62. Frazier, P. I. (2018). A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811.
63. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
64. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
65. Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2015). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. Journal of Statistical Software, 63(17), 1-50.
66. Grus, J. (2019). Data Science from Scratch: First Principles with Python. O'Reilly Media.
67. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.
68. Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: a review of the literature. SRE-Discussion Papers, 2010/03.
69. Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). Automated machine learning: methods, systems, challenges. Springer Nature.
70. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
71. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: springer.
72. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
73. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145).
74. Korver-Glenn, E. (2018). Compounding inequalities: How racial stereotypes and discrimination accumulate across the stages of housing exchange. American Sociological Review, 83(4), 627-656.
75. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1), 6765-6816.
76. Li, X., Zhao, X., Li, D., & Xu, W. (2019). Exploring spatial varying relationships between housing prices and multiple POIs using geographically weighted regression. ISPRS International Journal of Geo-Information, 8(11), 503.
77. Limsombunchai, V. (2004). House price prediction: Hedonic price model vs. artificial neural network. New Zealand Agricultural and Resource Economics Society Conference, 25-26.

78. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.
79. Mayer, M., Bourassa, S. C., Hoesli, M., & Scognamiglio, D. (2019). Estimation and updating methods for hedonic valuation. Journal of European Real Estate Research, 12(1), 134-150.
80. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
81. Molnar, C. (2019). Interpretable machine learning. Lulu. com.
82. Mu, J., Wu, F., & Zhang, A. (2020). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, 2020.
83. Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.
84. Panduro, T. E., & Veie, K. L. (2013). Classification and valuation of urban green spaces—A hedonic house price valuation. Landscape and Urban planning, 120, 119-128.
85. Peterson, S., & Flanagan, A. (2009). Neural network hedonic pricing models in mass real estate appraisal. Journal of Real Estate Research, 31(2), 147-164.
86. Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research, 20(53), 1-32.
87. Rafiei, M. H., & Adeli, H. (2016). A novel machine learning model for estimation of sale prices of real estate units. Journal of Construction Engineering and Management, 142(2), 04015066.
88. Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of Political Economy, 82(1), 34-55.
89. Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. Data mining and knowledge discovery, 1(3), 317-328.
90. Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The value of housing characteristics: a meta-analysis. The Journal of Real Estate Finance and Economics, 33(3), 215-240.
91. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25.
92. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting, 16(4), 437-450.
93. Tay, D. P., & Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. Journal of Property Valuation and Investment, 10(2), 525-540.
94. Wen, H., Zhang, Y., & Zhang, L. (2014). Do educational facilities affect housing price? An empirical study in Hangzhou, China. Habitat International, 42, 155-163.
95. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

96. Wooldridge, J. M. (2020). Introductory econometrics: A modern approach. Cengage learning.
97. Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. Journal of Electronic Science and Technology, 17(1), 26-40.
98. Xiao, Y., Chen, X., Li, Q., Yu, X., Chen, J., & Guo, J. (2017). Exploring determinants of housing prices in Beijing: An enhanced hedonic regression with open access POI data. ISPRS International Journal of Geo-Information, 6(11), 358.
99. Yacim, J. A., & Boshoff, D. G. (2018). Impact of artificial neural networks training algorithms on accurate prediction of property values. Journal of Real Estate Research, 40(3), 375-418.
100. Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On early stopping in gradient descent learning. Constructive Approximation, 26(2), 289-315.
101. Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970). PMLR.
102. Zukin, S., Trujillo, V., Frase, P., Jackson, D., Recuber, T., & Walker, A. (2009). New retail capital and neighborhood change: boutiques and gentrification in New York City. City & Community, 8(1), 47-64.
103. Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. Journal of Real Estate Research, 32(2), 139-160.
104. Karr, A. F., Sanil, A. P., & Banks, D. L. (2014). Data quality: A statistical perspective. Statistical Science, 29(4), 564-579.

APPENDIX

1. Feature Description

FEATURE NAME	FEATURE DESCRIPTION
SALE TYPE	The nature of the sale (e.g., new sale, resale, foreclosure).
SOLD DATE	The date the property was sold.
PROPERTY TYPE	The category of the property (e.g., single-family home, condo, townhouse).
ADDRESS	The street address of the property.
CITY	The city where the property is located.
STATE OR PROVINCE	The state or province of the property's location.
ZIP OR POSTAL CODE	The postal code area for the property's location.
NEIGHBORHOOD	The specific neighborhood or area within the city where the property is located.
PRICE	The sale price of the property.
BEDS	The number of bedrooms in the property.
BATHS	The number of bathrooms in the property.
LOCATION	The area of the house
SQUARE FEET	The total living area size of the property in square feet.
LOT SIZE	The size of the property's lot, indicating the total land area.
YEAR BUILT	The year the property was constructed.
DAYS ON MARKET	The total number of days the property was listed for sale before being sold.
$/SQUARE FEET	The price per square foot of the property, calculated by dividing the sale price by the square footage.
HOA/MONTH	The monthly Homeowners Association fee, applicable if the property is part of a community with shared amenities.
STATUS	The current status of the property listing (e.g., active, sold, pending).
NEXT OPEN HOUSE START TIME	The start time for the next scheduled open house.
NEXT OPEN HOUSE END TIME	The end time for the next scheduled open house.
URL	The web link to the property listing.
SOURCE	The source from which the property listing was obtained.
MLS#	The Multiple Listing Service number, a unique identifier for the property in real estate databases.
FAVORITE	Indicates whether the property has been marked as a favorite by a user (often a binary indicator).
INTERESTED	Indicates whether a user has expressed interest in the property (often a binary indicator).
lat	The latitude coordinate of the property's location.
long	The longitude coordinate of the property's location.
POI Type + "_accessibility"	This feature relate to the property's Accessibility Measurement to various local amenities and services specified in the POI Types Appendix

Table A.1 *- Description of Feature*

Accessibilty Measures	POI Types	Descriptions
gas_station_accessibility	gas_station	Gas Stations
parking_accessibility	parking	Parking Spaces
airport_accessibility	airport	Airport Terminals
bus_station_accessibility	bus_station	Bus Stations or Bus Stops
ferry_terminal_accessibility	ferry_terminal	Ferry Terminals
light_rail_station_accessibility	light_rail_station	Light Rail Stations
train_station_accessibility	train_station	Train Stations
transit_station_accessibility	transit_station	Transit Stations (Seattle StreetCars, Monorail)
transportation_others_accessibility	transportation_others	Other unspecified Public Transportations
library_accessibility	library	Libraries (Public and Private)
preschool_accessibility	preschool	Preschools
school_accessibility	school	Schools (All levels of schools excluding Universities)
university_accessibility	university	Universities
amusement_center_accessibility	amusement_center	Amusement Centers (Children Play Area, Escape Rooms, etc.)
community_center_accessibility	community_center	Community Centers
entertainment_accessibility	entertainment	Entertainment (Cinema, Nightclub, Arcade, Hike Trails, Viewpoints, etc.)
park_accessibility	park	Parks and Green Areas
bar_accessibility	bar	Bars (Pubs)
cafe_and_bakery_accessibility	cafe_and_bakery	Cafes and Bakeries
food_others_accessibility	food_others	Other unspecified Food Places (Small Food shops, Takeaways, etc.)
restaurant_accessibility	restaurant	Restaurants
health_store_accessibility	health_store	Drugstores and Pharmacies
hospital_accessibility	hospital	Hospitals and Clinics
places_of_worship_accessibility	places_of_worship	Places of Worship (Mosque, Churches, etc.)
safety_accessibility	safety	Police and Fire Stations
stores_essentials_accessibility	stores_essentials	Essential Stores (Grocery stores, Supermarket, Wholesale Marts, etc.)
stores_others_accessibility	stores_others	Other Store Types (Bike Stores, Tool shops, Sports Store, etc.)

Table A.2 *- Description of Accessibility Features*

FEATURE NAME	TYPE	Count	Mean	Min	Median	Max
SALE TYPE	Date	16,664	n/a	n/a	n/a	n/a
SOLD DATE	Date	13,747	n/a	n/a	n/a	n/a
PROPERTY TYPE	Object	16,664	n/a	n/a	n/a	n/a
ADDRESS	Object	16,664	n/a	n/a	n/a	n/a
CITY	Object	16,664	n/a	n/a	n/a	n/a
STATE OR PROVINCE	Object	16,664	n/a	n/a	n/a	n/a
ZIP OR POSTAL CODE	Object	16,664	98,107.39	98,003.00	98,116.00	98,126.00
NEIGHBORHOOD	Object	16,664	n/a	n/a	n/a	n/a
PRICE	Numerical	16,664	1,204,679.74	61,706.00	885,000.00	67,100,000.00
BEDS	Numerical	16,664	3.06	1.00	3.00	22.00
BATHS	Numerical	16,664	2.29	1.00	2.00	23.00
LOCATION	Object	16,664	n/a	n/a	n/a	n/a
SQUARE FEET	Numerical	16,664	1,943.24	142.00	1,720.00	20,587.00
LOT SIZE	Numerical	14,054	6,344.21	1.00	5,000.00	493,925.00
AGE	Numerical	16,664	1,972.16	1,889.00	1,973.00	2,025.00
DAYS ON MARKET	Numerical	2,509	38.48	1.00	21.00	825.00
$/SQUARE FEET	Numerical	16,664	608.57	60.00	572.00	6,785.00
HOA/MONTH	Numerical	4,555	442.26	-	382.00	4,699.00
STATUS	Object	15,647	n/a	n/a	n/a	n/a
NEXT OPEN HOUSE START TIME	Time	410	n/a	n/a	n/a	n/a
NEXT OPEN HOUSE END TIME	Time	410	n/a	n/a	n/a	n/a
URL	Object	16,664	n/a	n/a	n/a	n/a
SOURCE	Object	15,647	n/a	n/a	n/a	n/a
MLS#	Object	15,647	2,113,659.56	1,723,475.00	2,137,846.00	2,258,093.00
FAVORITE	Object	16,664	n/a	n/a	n/a	n/a
INTERESTED	Object	16,664	n/a	n/a	n/a	n/a
lat	Numerical	16,664	47.63	47.50	47.64	47.74
long	Numerical	16,664	- 122.32	- 122.42	- 122.32	- 122.21
gas_station_accessibility	Numerical	16,664	1.19	-	1.09	3.94
parking_accessibility	Numerical	16,664	13.72	-	3.06	259.77
airport_accessibility	Numerical	16,664	0.06	-	-	3.33
bus_station_accessibility	Numerical	16,664	21.23	-	20.18	87.11
ferry_terminal_accessibility	Numerical	16,664	0.01	-	-	0.89
light_rail_station_accessibility	Numerical	16,664	0.34	-	-	8.40
train_station_accessibility	Numerical	16,664	0.02	-	-	1.66
transit_station_accessibility	Numerical	16,664	13.20	-	12.64	39.78
transportation_others_accessibility	Numerical	16,664	0.02	-	-	0.87
library_accessibility	Numerical	16,664	1.52	-	1.27	11.67
preschool_accessibility	Numerical	16,664	1.84	-	1.60	7.74
school_accessibility	Numerical	16,664	8.69	-	8.67	23.57
university_accessibility	Numerical	16,664	1.77	-	0.21	75.86
amusement_center_accessibility	Numerical	16,664	0.13	-	-	4.12
community_center_accessibility	Numerical	16,664	0.62	-	0.48	4.65
entertainment_accessibility	Numerical	16,664	2.38	-	1.56	18.78
park_accessibility	Numerical	16,664	25.17	0.00	12.28	293.35
bar_accessibility	Numerical	16,664	5.15	-	2.18	66.75
cafe_and_bakery_accessibility	Numerical	16,664	9.04	-	5.13	101.20
food_others_accessibility	Numerical	16,664	5.81	-	4.58	28.02
restaurant_accessibility	Numerical	16,664	26.03	-	14.47	277.20
health_store_accessibility	Numerical	16,664	3.63	-	1.96	41.16
hospital_accessibility	Numerical	16,664	1.74	-	0.26	62.24
places_of_worship_accessibility	Numerical	16,664	5.11	-	4.54	20.75
safety_accessibility	Numerical	16,664	0.48	-	0.36	4.83
stores_essentials_accessibility	Numerical	16,664	4.85	-	3.41	34.49
stores_others_accessibility	Numerical	16,664	23.64	-	19.11	126.34

Table A.3 *- Data Description*

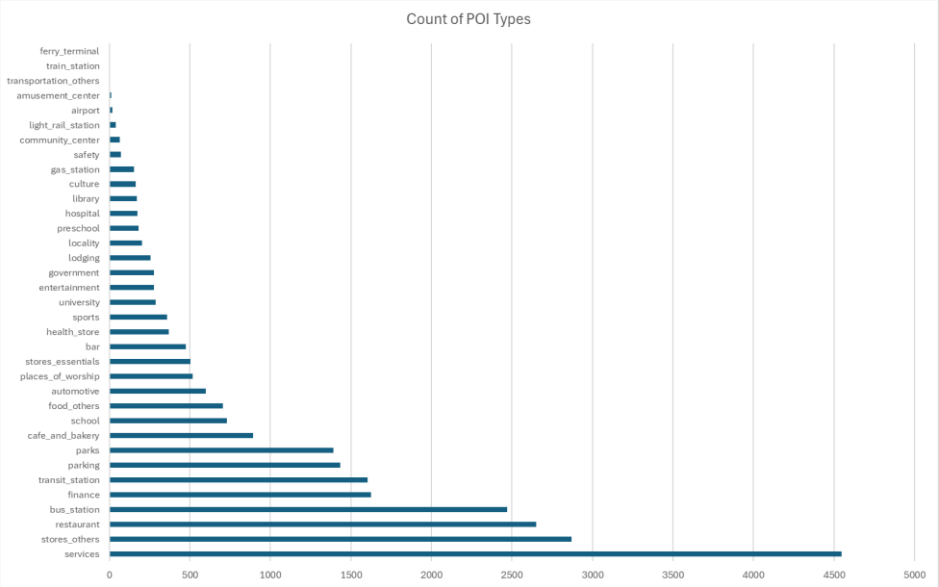

Figure A.1 *- Count of Seattle POI Types*

2. Results of all 192 unique models

No	Model Code	Variation Code	Feature	Dummy	Version	Model	Model Type	RMSE	MAE	R^2 Score
1	C11.R1	C11	All	Yes	Version B (Exclude Price)	Random Forest Baseline Model	Random Forest	676,498.10	180,600.13	0.7619
2	A12.R1	A12	All	No	Version A (All Scaling)	Random Forest Baseline Model	Random Forest	0.4628	0.1243	0.7614
3	D12.R2	D12	All	No	Version D (No Scaling)	Random Forest with GridSearch	Random Forest	683,703.80	158,865.34	0.7568
4	D11.R2	D11	All	Yes	Version D (No Scaling)	Random Forest with GridSearch	Random Forest	683,703.80	158,865.34	0.7568
5	C12.R2	C12	All	No	Version B (Exclude Price)	Random Forest with GridSearch	Random Forest	686,754.40	159,371.30	0.7546
6	B11.R1	B11	All	Yes	Version C (Price Only)	Random Forest Baseline Model	Random Forest	0.4698	0.1238	0.7541
7	A11.R1	A11	All	Yes	Version A (All Scaling)	Random Forest Baseline Model	Random Forest	0.4710	0.1241	0.7528
8	C12.R1	C12	All	No	Version B (Exclude Price)	Random Forest Baseline Model	Random Forest	690,672.90	183,040.10	0.7518
9	C11.R2	C11	All	Yes	Version B (Exclude Price)	Random Forest with GridSearch	Random Forest	691,241.03	164,817.10	0.7514
10	D12.R1	D12	All	No	Version D (No Scaling)	Random Forest Baseline Model	Random Forest	691,946.50	183,179.43	0.7509
11	D11.R1	D11	All	Yes	Version D (No Scaling)	Random Forest Baseline Model	Random Forest	691,946.50	183,179.43	0.7509
12	B11.R2	B11	All	Yes	Version C (Price Only)	Random Forest with GridSearch	Random Forest	0.4779	0.1138	0.7456
13	B12.R1	B12	All	No	Version C (Price Only)	Random Forest Baseline Model	Random Forest	0.4789	0.1254	0.7445
14	A11.R2	A11	All	Yes	Version A (All Scaling)	Random Forest with GridSearch	Random Forest	0.4804	0.1140	0.7428
15	B12.R2	B12	All	No	Version C (Price Only)	Random Forest with GridSearch	Random Forest	0.4817	0.1081	0.7415
16	A12.R2	A12	All	No	Version A (All Scaling)	Random Forest with GridSearch	Random Forest	0.4825	0.1080	0.7406
17	C21.R2	C21	Selected	Yes	Version B (Exclude Price)	Random Forest with GridSearch	Random Forest	710,922.25	159,748.25	0.7370
18	A12.X2	A12	All	No	Version A (All Scaling)	XGBoost with GridSearch	XGBoost	0.4891	0.1400	0.7334
19	B12.X2	B12	All	No	Version C (Price Only)	XGBoost with GridSearch	XGBoost	0.4891	0.1400	0.7334
20	D12.X2	D12	All	No	Version D (No Scaling)	XGBoost with GridSearch	XGBoost	715,735.00	204,814.87	0.7334
21	D11.X2	D11	All	Yes	Version D (No Scaling)	XGBoost with GridSearch	XGBoost	715,735.00	204,814.87	0.7334
22	B21.R2	B21	Selected	Yes	Version C (Price Only)	Random Forest with GridSearch	Random Forest	0.4900	0.1096	0.7325
23	C21.N1	C21	Selected	Yes	Version B (Exclude Price)	Neural Network Baseline Model	Neural Network	718,305.29	261,179.37	0.7315
24	A21.R1	A21	Selected	Yes	Version A (All Scaling)	Random Forest Baseline Model	Random Forest	0.4930	0.1300	0.7291
25	A21.R2	A21	Selected	Yes	Version A (All Scaling)	Random Forest with GridSearch	Random Forest	0.4944	0.1101	0.7277
26	D22.R2	D22	Selected	No	Version D (No Scaling)	Random Forest with GridSearch	Random Forest	728,872.00	162,993.03	0.7236
27	D21.R2	D21	Selected	Yes	Version D (No Scaling)	Random Forest with GridSearch	Random Forest	728,872.00	162,993.03	0.7236
28	C22.R2	C22	Selected	No	Version B (Exclude Price)	Random Forest with GridSearch	Random Forest	729,592.48	163,012.07	0.7230
29	B22.R2	B22	Selected	No	Version C (Price Only)	Random Forest with GridSearch	Random Forest	0.4990	0.1115	0.7225
30	C11.N2	C11	All	Yes	Version B (Exclude Price)	Neural Network Hyper Model Early Stopping	Neural Network	732,148.60	272,540.87	0.7211
31	A21.N3	A21	Selected	Yes	Version A (All Scaling)	Neural Network Hyper Model	Neural Network	0.5007	0.1650	0.7206
32	C11.N3	C11	All	Yes	Version B (Exclude Price)	Neural Network Hyper Model	Neural Network	733,978.30	263,630.10	0.7197
33	A22.R2	A22	Selected	No	Version A (All Scaling)	Random Forest with GridSearch	Random Forest	0.5020	0.1113	0.7192
34	C21.R1	C21	Selected	Yes	Version B (Exclude Price)	Random Forest Baseline Model	Random Forest	736,486.72	191,011.38	0.7178
35	C12.N3	C12	All	No	Version B (Exclude Price)	Neural Network Hyper Model	Neural Network	738,274.40	268,492.70	0.7164
36	C22.R1	C22	Selected	No	Version B (Exclude Price)	Random Forest Baseline Model	Random Forest	743,175.37	194,234.85	0.7126
37	D11.N2	D11	All	Yes	Version D (No Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	743,807.40	273,362.29	0.7121
38	C21.N3	C21	Selected	Yes	Version B (Exclude Price)	Neural Network Hyper Model	Neural Network	744,845.48	285,004.02	0.7113
39	B22.R1	B22	Selected	No	Version C (Price Only)	Random Forest Baseline Model	Random Forest	0.5092	0.1327	0.7111
40	A11.X2	A11	All	Yes	Version A (All Scaling)	XGBoost with GridSearch	XGBoost	0.5103	0.1503	0.7098
41	B11.X2	B11	All	Yes	Version C (Price Only)	XGBoost with GridSearch	XGBoost	0.5103	0.1503	0.7098
42	B21.R1	B21	Selected	Yes	Version C (Price Only)	Random Forest Baseline Model	Random Forest	0.5106	0.1314	0.7095
43	D22.R1	D22	Selected	No	Version D (No Scaling)	Random Forest Baseline Model	Random Forest	747,274.10	194,532.24	0.7094
44	D21.R1	D21	Selected	Yes	Version D (No Scaling)	Random Forest Baseline Model	Random Forest	747,274.10	194,532.24	0.7094
45	D12.N1	D12	All	No	Version D (No Scaling)	Neural Network Baseline Model	Neural Network	750,270.00	280,362.76	0.7071
46	D11.N1	D11	All	Yes	Version D (No Scaling)	Neural Network Baseline Model	Neural Network	750,270.00	280,362.76	0.7071
47	A22.R1	A22	Selected	No	Version A (All Scaling)	Random Forest Baseline Model	Random Forest	0.5134	0.1332	0.7063
48	C21.N2	C21	Selected	Yes	Version B (Exclude Price)	Neural Network Hyper Model Early Stopping	Neural Network	751,930.50	285,965.47	0.7058
49	A22.N3	A22	Selected	No	Version A (All Scaling)	Neural Network Hyper Model	Neural Network	0.5158	0.1681	0.7036
50	C22.N1	C22	Selected	No	Version B (Exclude Price)	Neural Network Baseline Model	Neural Network	755,077.30	265,274.42	0.7033

Table A.4 *- Top 50 Performing Models*

No	Model Code	Variation Code	Feature	Dummy	Version	Model	Model Type	RMSE	MAE	R^2 Score
51	C11.N1	C11	All	Yes	Version B (Exclude Price)	Neural Network Baseline Model	Neural Network	756,465.40	242,751.64	0.7022
52	B12.X3	B12	All	No	Version C (Price Only)	XGBoost with RandomizedSearch	XGBoost	0.5177	0.1508	0.7014
53	D12.X3	D12	All	No	Version D (No Scaling)	XGBoost with RandomizedSearch	XGBoost	757,549.80	220,637.04	0.7014
54	D11.X3	D11	All	Yes	Version D (No Scaling)	XGBoost with RandomizedSearch	XGBoost	757,549.80	220,637.04	0.7014
55	A12.X3	A12	All	No	Version A (All Scaling)	XGBoost with RandomizedSearch	XGBoost	0.5177	0.1508	0.7014
56	C12.X3	C12	All	No	Version B (Exclude Price)	XGBoost with RandomizedSearch	XGBoost	757,551.40	220,649.70	0.7014
57	A11.X3	A11	All	Yes	Version A (All Scaling)	XGBoost with RandomizedSearch	XGBoost	0.5179	0.1478	0.7012
58	B11.X3	B11	All	Yes	Version C (Price Only)	XGBoost with RandomizedSearch	XGBoost	0.5179	0.1478	0.7012
59	C11.X3	C11	All	Yes	Version B (Exclude Price)	XGBoost with RandomizedSearch	XGBoost	757,818.70	216,258.10	0.7012
60	D22.N1	D22	Selected	No	Version D (No Scaling)	Neural Network Baseline Model	Neural Network	758,335.60	285,198.86	0.7008
61	D21.N1	D21	Selected	Yes	Version D (No Scaling)	Neural Network Baseline Model	Neural Network	758,335.60	285,198.86	0.7008
62	A12.N3	A12	All	No	Version A (All Scaling)	Neural Network Hyper Model	Neural Network	0.5185	0.1574	0.7004
63	C12.N2	C12	All	No	Version B (Exclude Price)	Neural Network Hyper Model Early Stopping	Neural Network	763,263.50	291,985.60	0.6969
64	A11.N3	A11	All	Yes	Version A (All Scaling)	Neural Network Hyper Model	Neural Network	0.5248	0.1602	0.6931
65	C12.N1	C12	All	No	Version B (Exclude Price)	Neural Network Baseline Model	Neural Network	768,093.50	244,673.00	0.6930
66	A11.N4	A11	All	Yes	Version A (All Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	0.5249	0.2079	0.6930
67	D22.N2	D22	Selected	No	Version D (No Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	768,546.50	297,187.25	0.6927
68	D21.N2	D21	Selected	Yes	Version D (No Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	768,546.50	297,187.25	0.6927
69	A11.N2	A11	All	Yes	Version A (All Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	0.5264	0.1603	0.6912
70	A21.N4	A21	Selected	Yes	Version A (All Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	0.5298	0.1938	0.6872
71	D11.N3	D11	All	Yes	Version D (No Scaling)	Neural Network Hyper Model	Neural Network	776,164.80	283,541.01	0.6865
72	D12.N3	D12	All	No	Version D (No Scaling)	Neural Network Hyper Model	Neural Network	777,261.70	280,700.16	0.6856
73	A12.N2	A12	All	No	Version A (All Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	0.5323	0.1612	0.6843
74	B11.X1	B11	All	Yes	Version C (Price Only)	XGBoost Baseline Model	XGBoost	0.5325	0.1159	0.6840
75	A11.X1	A11	All	Yes	Version A (All Scaling)	XGBoost Baseline Model	XGBoost	0.5328	0.1154	0.6837
76	C11.X1	C11	All	Yes	Version B (Exclude Price)	XGBoost Baseline Model	XGBoost	780,240.90	169,624.70	0.6832
77	D12.X1	D12	All	No	Version D (No Scaling)	XGBoost Baseline Model	XGBoost	785,741.90	173,697.32	0.6787
78	D11.X1	D11	All	Yes	Version D (No Scaling)	XGBoost Baseline Model	XGBoost	785,741.90	173,697.32	0.6787
79	C12.X1	C12	All	No	Version B (Exclude Price)	XGBoost Baseline Model	XGBoost	785,841.70	173,512.10	0.6787
80	C22.N3	C22	Selected	No	Version B (Exclude Price)	Neural Network Hyper Model	Neural Network	785,850.50	301,441.74	0.6787
81	B12.X1	B12	All	No	Version C (Price Only)	XGBoost Baseline Model	XGBoost	0.5372	0.1184	0.6784
82	A22.N2	A22	Selected	No	Version A (All Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	0.5380	0.1738	0.6775
83	A12.X1	A12	All	No	Version A (All Scaling)	XGBoost Baseline Model	XGBoost	0.5390	0.1189	0.6763
84	C22.N2	C22	Selected	No	Version B (Exclude Price)	Neural Network Hyper Model Early Stopping	Neural Network	790,723.70	303,164.98	0.6747
85	A12.N4	A12	All	No	Version A (All Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	0.5408	0.1955	0.6741
86	A21.X3	A21	Selected	Yes	Version A (All Scaling)	XGBoost with RandomizedSearch	XGBoost	0.5416	0.1599	0.6731
87	B21.X3	B21	Selected	Yes	Version C (Price Only)	XGBoost with RandomizedSearch	XGBoost	0.5416	0.1599	0.6731
88	C21.X3	C21	Selected	Yes	Version B (Exclude Price)	XGBoost with RandomizedSearch	XGBoost	792,617.90	234,025.30	0.6731
89	A22.N4	A22	Selected	No	Version A (All Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	0.5453	0.2100	0.6686
90	D22.N3	D22	Selected	No	Version D (No Scaling)	Neural Network Hyper Model	Neural Network	798,530.50	291,829.47	0.6682
91	D21.N3	D21	Selected	Yes	Version D (No Scaling)	Neural Network Hyper Model	Neural Network	798,530.50	291,829.47	0.6682
92	A12.N1	A12	All	No	Version A (All Scaling)	Neural Network Baseline Model	Neural Network	0.5511	0.2043	0.6616
93	A21.N2	A21	Selected	Yes	Version A (All Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	0.5524	0.1657	0.6600
94	A11.N1	A11	All	Yes	Version A (All Scaling)	Neural Network Baseline Model	Neural Network	0.5539	0.2027	0.6581
95	D12.N2	D12	All	No	Version D (No Scaling)	Neural Network Hyper Model Early Stopping	Neural Network	818,634.70	295,651.77	0.6513
96	A22.X2	A22	Selected	No	Version A (All Scaling)	XGBoost with GridSearch	XGBoost	0.5609	0.1725	0.6495
97	B22.X2	B22	Selected	No	Version C (Price Only)	XGBoost with GridSearch	XGBoost	0.5639	0.1728	0.6457
98	D22.X2	D22	Selected	No	Version D (No Scaling)	XGBoost with GridSearch	XGBoost	825,176.20	252,846.38	0.6457
99	D21.X2	D21	Selected	Yes	Version D (No Scaling)	XGBoost with GridSearch	XGBoost	825,176.20	252,846.38	0.6457
100	A22.X1	A22	Selected	No	Version A (All Scaling)	XGBoost Baseline Model	XGBoost	0.5640	0.1284	0.6455

Table A.5 *- Unique Models rank 51 to 100*

No	Model Code	Variation Code	Feature	Dummy	Version	Model	Model Type	RMSE	MAE	R^2 Score
101	B22.X1	B22	Selected	No	Version C (Price Only)	XGBoost Baseline Model	XGBoost	0.5644	0.1283	0.6450
102	C22.X3	C22	Selected	No	Version B (Exclude Price)	XGBoost with RandomizedSearch	XGBoost	825,962.40	241,427.40	0.6450
103	D22.X3	D22	Selected	No	Version D (No Scaling)	XGBoost with RandomizedSearch	XGBoost	825,962.40	241,427.44	0.6450
104	D21.X3	D21	Selected	Yes	Version D (No Scaling)	XGBoost with RandomizedSearch	XGBoost	825,962.40	241,427.44	0.6450
105	A22.X3	A22	Selected	No	Version A (All Scaling)	XGBoost with RandomizedSearch	XGBoost	0.5644	0.1650	0.6450
106	B22.X3	B22	Selected	No	Version C (Price Only)	XGBoost with RandomizedSearch	XGBoost	0.5644	0.1650	0.6450
107	B12.N3	B12	All	No	Version C (Price Only)	Neural Network Hyper Model	Neural Network	0.5675	0.2437	0.6411
108	C22.X1	C22	Selected	No	Version B (Exclude Price)	XGBoost Baseline Model	XGBoost	830,788.90	188,176.10	0.6409
109	D22.X1	D22	Selected	No	Version D (No Scaling)	XGBoost Baseline Model	XGBoost	832,650.20	187,907.87	0.6392
110	D21.X1	D21	Selected	Yes	Version D (No Scaling)	XGBoost Baseline Model	XGBoost	832,650.20	187,907.87	0.6392
111	B21.X1	B21	Selected	Yes	Version C (Price Only)	XGBoost Baseline Model	XGBoost	0.5756	0.1285	0.6308
112	A21.X1	A21	Selected	Yes	Version A (All Scaling)	XGBoost Baseline Model	XGBoost	0.5758	0.1285	0.6306
113	C21.X1	C21	Selected	Yes	Version B (Exclude Price)	XGBoost Baseline Model	XGBoost	843,378.40	188,069.60	0.6299
114	B11.N3	B11	All	Yes	Version C (Price Only)	Neural Network Hyper Model	Neural Network	0.5771	0.2464	0.6289
115	B11.N2	B11	All	Yes	Version C (Price Only)	Neural Network Hyper Model Early Stopping	Neural Network	0.5983	0.2162	0.6011
116	A21.X2	A21	Selected	Yes	Version A (All Scaling)	XGBoost with GridSearch	XGBoost	0.6062	0.1745	0.5905
117	B21.X2	B21	Selected	Yes	Version C (Price Only)	XGBoost with GridSearch	XGBoost	0.6062	0.1745	0.5905
118	B22.N3	B22	Selected	No	Version C (Price Only)	Neural Network Hyper Model	Neural Network	0.6082	0.2563	0.5879
119	B12.N2	B12	All	No	Version C (Price Only)	Neural Network Hyper Model Early Stopping	Neural Network	0.6084	0.2244	0.5875
120	A22.N1	A22	Selected	No	Version A (All Scaling)	Neural Network Baseline Model	Neural Network	0.6113	0.1992	0.5836
121	A21.N1	A21	Selected	Yes	Version A (All Scaling)	Neural Network Baseline Model	Neural Network	0.6155	0.1775	0.5778
122	B12.N4	B12	All	No	Version C (Price Only)	Neural Network Baseline Model Early Stopping	Neural Network	0.6162	0.2607	0.5769
123	A11.L3	A11	All	Yes	Version A (All Scaling)	Ridge Regression with GridSearch	Ridge Regression	0.6470	0.2905	0.5335
124	C11.L3	C11	All	Yes	Version B (Exclude Price)	Ridge Regression with GridSearch	Ridge Regression	946,842.61	425,068.70	0.5335
125	B11.L3	B11	All	Yes	Version C (Price Only)	Ridge Regression with GridSearch	Ridge Regression	0.6470	0.2905	0.5335
126	A11.L2	A11	All	Yes	Version A (All Scaling)	Ridge Regression	Ridge Regression	0.6471	0.2905	0.5335
127	B11.L2	B11	All	Yes	Version C (Price Only)	Ridge Regression	Ridge Regression	0.6471	0.2905	0.5335
128	C11.L2	C11	All	Yes	Version B (Exclude Price)	Ridge Regression	Ridge Regression	946,885.08	425,164.04	0.5335
129	B11.L1	B11	All	Yes	Version C (Price Only)	Linear Regression	Linear Regression	0.6471	0.2906	0.5335
130	A11.L1	A11	All	Yes	Version A (All Scaling)	Linear Regression	Linear Regression	0.6471	0.2906	0.5335
131	C11.L1	C11	All	Yes	Version B (Exclude Price)	Linear Regression	Linear Regression	946,895.22	425,242.73	0.5335
132	A12.L3	A12	All	No	Version A (All Scaling)	Ridge Regression with GridSearch	Ridge Regression	0.6488	0.2861	0.5310
133	C12.L3	C12	All	No	Version B (Exclude Price)	Ridge Regression with GridSearch	Ridge Regression	949,390.40	418,602.00	0.5310
134	A21.L3	A21	Selected	Yes	Version A (All Scaling)	Ridge Regression with GridSearch	Ridge Regression	0.6488	0.2909	0.5309
135	C21.L3	C21	Selected	Yes	Version B (Exclude Price)	Ridge Regression with GridSearch	Ridge Regression	949,475.15	425,751.77	0.5309
136	B21.L3	B21	Selected	Yes	Version C (Price Only)	Ridge Regression with GridSearch	Ridge Regression	0.6488	0.2910	0.5309
137	A21.L2	A21	Selected	Yes	Version A (All Scaling)	Ridge Regression	Ridge Regression	0.6489	0.2910	0.5309
138	B21.L2	B21	Selected	Yes	Version C (Price Only)	Ridge Regression	Ridge Regression	0.6489	0.2910	0.5309
139	C21.L2	C21	Selected	Yes	Version B (Exclude Price)	Ridge Regression	Ridge Regression	949,522.07	425,845.11	0.5309
140	B21.L1	B21	Selected	Yes	Version C (Price Only)	Linear Regression	Linear Regression	0.6489	0.2911	0.5309
141	A21.L1	A21	Selected	Yes	Version A (All Scaling)	Linear Regression	Linear Regression	0.6489	0.2910	0.5308
142	C21.L1	C21	Selected	Yes	Version B (Exclude Price)	Linear Regression	Linear Regression	949,541.34	425,903.72	0.5308
143	A12.L2	A12	All	No	Version A (All Scaling)	Ridge Regression	Ridge Regression	0.6489	0.2905	0.5308
144	B12.L2	B12	All	No	Version C (Price Only)	Ridge Regression	Ridge Regression	0.6489	0.2905	0.5308
145	C12.L2	C12	All	No	Version B (Exclude Price)	Ridge Regression	Ridge Regression	949,602.60	425,117.00	0.5308
146	D12.L2	D12	All	No	Version D (No Scaling)	Ridge Regression	Ridge Regression	949,602.60	425,117.02	0.5308
147	D11.L2	D11	All	Yes	Version D (No Scaling)	Ridge Regression	Ridge Regression	949,602.60	425,117.02	0.5308
148	A12.L1	A12	All	No	Version A (All Scaling)	Linear Regression	Linear Regression	0.6489	0.2906	0.5308
149	B12.L1	B12	All	No	Version C (Price Only)	Linear Regression	Linear Regression	0.6489	0.2906	0.5308
150	C12.L1	C12	All	No	Version B (Exclude Price)	Linear Regression	Linear Regression	949,611.10	425,204.30	0.5308

Table A.6 *- Unique Models Ranked 101 to 150*

No	Model Code	Variation Code	Feature	Dummy	Version	Model	Model Type	RMSE	MAE	R^2 Score
151	D12.L1	D12	All	No	Version D (No Scaling)	Linear Regression	Linear Regression	949,611.10	425,204.28	0.5308
152	D11.L1	D11	All	Yes	Version D (No Scaling)	Linear Regression	Linear Regression	949,611.10	425,204.28	0.5308
153	B12.L3	B12	All	No	Version C (Price Only)	Ridge Regression with GridSearch	Ridge Regression	0.6491	0.2867	0.5305
154	D12.L3	D12	All	No	Version D (No Scaling)	Ridge Regression with GridSearch	Ridge Regression	949,926.80	419,590.33	0.5305
155	D11.L3	D11	All	Yes	Version D (No Scaling)	Ridge Regression with GridSearch	Ridge Regression	949,926.80	419,590.33	0.5305
156	A22.L3	A22	Selected	No	Version A (All Scaling)	Ridge Regression with GridSearch	Ridge Regression	0.6505	0.2872	0.5285
157	C22.L3	C22	Selected	No	Version B (Exclude Price)	Ridge Regression with GridSearch	Ridge Regression	951,889.89	420,263.64	0.5285
158	A22.L2	A22	Selected	No	Version A (All Scaling)	Ridge Regression	Ridge Regression	0.6506	0.2915	0.5283
159	B22.L2	B22	Selected	No	Version C (Price Only)	Ridge Regression	Ridge Regression	0.6506	0.2915	0.5283
160	C22.L2	C22	Selected	No	Version B (Exclude Price)	Ridge Regression	Ridge Regression	952,132.90	426,634.54	0.5283
161	D22.L2	D22	Selected	No	Version D (No Scaling)	Ridge Regression	Ridge Regression	952,132.90	426,634.54	0.5283
162	D21.L2	D21	Selected	Yes	Version D (No Scaling)	Ridge Regression	Ridge Regression	952,132.90	426,634.54	0.5283
163	A22.L1	A22	Selected	No	Version A (All Scaling)	Linear Regression	Linear Regression	0.6507	0.2916	0.5283
164	B22.L1	B22	Selected	No	Version C (Price Only)	Linear Regression	Linear Regression	0.6507	0.2916	0.5283
165	C22.L1	C22	Selected	No	Version B (Exclude Price)	Linear Regression	Linear Regression	952,138.27	426,704.00	0.5283
166	D22.L1	D22	Selected	No	Version D (No Scaling)	Linear Regression	Linear Regression	952,138.30	426,704.00	0.5283
167	D21.L1	D21	Selected	Yes	Version D (No Scaling)	Linear Regression	Linear Regression	952,138.30	426,704.00	0.5283
168	B22.N2	B22	Selected	No	Version C (Price Only)	Neural Network Hyper Model Early Stopping	Neural Network	0.6512	0.2350	0.5275
169	B22.L3	B22	Selected	No	Version C (Price Only)	Ridge Regression with GridSearch	Ridge Regression	0.6512	0.2885	0.5275
170	D22.L3	D22	Selected	No	Version D (No Scaling)	Ridge Regression with GridSearch	Ridge Regression	952,948.70	422,209.81	0.5275
171	D21.L3	D21	Selected	Yes	Version D (No Scaling)	Ridge Regression with GridSearch	Ridge Regression	952,948.70	422,209.81	0.5275
172	B11.N4	B11	All	Yes	Version C (Price Only)	Neural Network Baseline Model Early Stopping	Neural Network	0.6531	0.2547	0.5247
173	B21.N2	B21	Selected	Yes	Version C (Price Only)	Neural Network Hyper Model Early Stopping	Neural Network	0.6556	0.2475	0.5211
174	C21.N4	C21	Selected	Yes	Version B (Exclude Price)	Neural Network Baseline Model Early Stopping	Neural Network	998,570.05	388,038.24	0.4811
175	C22.N4	C22	Selected	No	Version B (Exclude Price)	Neural Network Baseline Model Early Stopping	Neural Network	1,009,283.00	399,860.29	0.4700
176	D12.N4	D12	All	No	Version D (No Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	1,047,120.00	380,875.26	0.4295
177	D11.N4	D11	All	Yes	Version D (No Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	1,047,120.00	380,875.26	0.4295
178	B22.N4	B22	Selected	No	Version C (Price Only)	Neural Network Baseline Model Early Stopping	Neural Network	0.7181	0.3000	0.4254
179	D22.N4	D22	Selected	No	Version D (No Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	1,051,189.00	381,730.79	0.4250
180	D21.N4	D21	Selected	Yes	Version D (No Scaling)	Neural Network Baseline Model Early Stopping	Neural Network	1,051,189.00	381,730.79	0.4250
181	B21.N3	B21	Selected	Yes	Version C (Price Only)	Neural Network Hyper Model	Neural Network	0.7340	0.2406	0.3997
182	C11.N4	C11	All	Yes	Version B (Exclude Price)	Neural Network Baseline Model Early Stopping	Neural Network	1,124,387.00	454,812.26	0.3422
183	C12.N4	C12	All	No	Version B (Exclude Price)	Neural Network Baseline Model Early Stopping	Neural Network	1,131,375.00	474,803.20	0.3340
184	B21.N4	B21	Selected	Yes	Version C (Price Only)	Neural Network Baseline Model Early Stopping	Neural Network	0.7964	0.2708	0.2933
185	B12.N1	B12	All	No	Version C (Price Only)	Neural Network Baseline Model	Neural Network	0.9474	0.4055	-0.0001
186	B11.N1	B11	All	Yes	Version C (Price Only)	Neural Network Baseline Model	Neural Network	0.9474	0.4056	-0.0002
187	B22.N1	B22	Selected	No	Version C (Price Only)	Neural Network Baseline Model	Neural Network	0.9474	0.4056	-0.0002
188	B21.N1	B21	Selected	Yes	Version C (Price Only)	Neural Network Baseline Model	Neural Network	0.9474	0.4056	-0.0002
189	C21.X2	C21	Selected	Yes	Version B (Exclude Price)	XGBoost with GridSearch	XGBoost	1,836,469.00	1,204,492.00	-0.7549
190	C12.X2	C12	All	No	Version B (Exclude Price)	XGBoost with GridSearch	XGBoost	1,836,470.00	1,204,492.00	-0.7549
191	C11.X2	C11	All	Yes	Version B (Exclude Price)	XGBoost with GridSearch	XGBoost	1,836,470.00	1,204,492.00	-0.7549
192	C22.X2	C22	Selected	No	Version B (Exclude Price)	XGBoost with GridSearch	XGBoost	1,836,470.00	1,204,492.00	-0.7549

Table A.7 *- Unique Models Ranked 151 to 192*

REPOSITORY

https://github.com/ahskasutoyo/seattle-house-price-and-poi/tree/main