Report 2: Predictive model to estimate property price

Abdiaziz Maalim (202104688) Chihab Agharbi (202208736)
Fahad Ahmed (202207274)

1 Introduction

In this comprehensive report, we embark on a data-driven exploration of the complex dynamics within the real estate market of Connecticut, USA. Our central aim is to equip prospective property buyers with the necessary tools and insights essential for making informed investment decisions amidst the intricacies of this ever-evolving landscape.

A pivotal challenge in the real estate domain revolves around the accurate prediction of house property sale amounts. Addressing this predictive challenge requires a nuanced comprehension of the myriad factors influencing property prices. Thus, our investigation focuses on bridging this predictive gap, providing stakeholders with a dependable tool for anticipating sale amounts based on crucial attributes.

To achieve this objective, we harness the potency of big data and employ cutting-edge data analytics techniques to scrutinize a comprehensive dataset of real estate sales spanning nearly two decades. Encompassing a diverse range of attributes, from property size and location to sales history, the dataset serves as a rich foundation for our analysis. Through meticulous data preprocessing, in-depth exploratory data analysis, and the application of advanced machine learning methodologies, we strive to uncover patterns and trends influencing property prices.

Our commitment to transparency and replicability is underscored by the Python script, 'report_2.py,' crafted specifically for predicting house sale amounts using advanced machine learning techniques. This script addresses the critical challenge of accurate sales forecasts, allowing others to reproduce our results on a dataset subset, fostering collaboration and encouraging further exploration of the data.

Key techniques and tools within our analytical arsenal include data cleaning, feature engineering, regression analysis, clustering, and geospatial visualization. These methodologies not only facilitate the construction of precise price prediction models but also unearth valuable market insights, including regional variations in property pricing, seasonal trends, and the impact of property characteristics on valuation

2 Dataset

2.1 Dataset Source

Link: `https://catalog.data.gov/dataset/real-estate-sales-2001-2018`

2.2 Dataset Description

The Connecticut Real Estate Sales Analysis dataset, available on data.gov, presents a comprehensive collection of information on real estate sales in Connecticut spanning the years 2001 to 2018. This dataset includes crucial details such as the town of the property, property address, date of sale, property type, sales price, and property assessment. Focusing on sales with a price equal to or exceeding $2,000, the dataset covers transactions occurring within the annual period between October 1 and September 30. The data is collected in accordance with specific Connecticut General Statutes and is reported annually, organized by grand list year. This dataset serves as a valuable resource for in-depth analysis and exploration of the dynamics, trends, and patterns within the Connecticut real estate market over the specified timeframe.

3 Task Description

The overarching objective of our project is to leverage machine learning techniques to construct a robust predictive model that accurately estimates Sale Amounts for properties listed and sold in Connecticut, USA. This predictive model incorporates historical data, utilizing features such as List Year, Location, Assessed Value, and additional features derived from existing ones. These derived features are carefully engineered to enhance the model's capacity to discern intricate patterns and relationships that influence property sale prices. The successful implementation of this predictive model holds substantial implications for stakeholders in the real estate market, including buyers, sellers, and investors. By providing valuable insights into the multifaceted factors driving property values, the model facilitates more informed decision-making in the dynamic Connecticut real estate landscape. The inclusion of additional features further refines the model's predictive capabilities, offering a nuanced understanding of the complex interplay of variables that impact property sale amounts. This holistic approach contributes to a more accurate and comprehensive tool for navigating the intricacies of the real estate market in Connecticut

4 Task Design

4.1 Data Cleaning

The data cleaning process before machine learning involved the removal of extraneous features such as Non Use Code, Assessor Remarks, and OPM Remarks to streamline the dataset. A meticulous search for common placeholders for missing values, such as 'NA', 'NAN', 'NaN', and 0, was conducted, and any instances were identified and rectified. Subsequently, rows containing missing information were excluded to enhance dataset integrity. Additionally, a strategic approach was employed to assign Residential Type to instances where Property Type was absent, except for NaN values, ensuring a more complete dataset. These measures collectively contribute to a refined and reliable dataset, laying the foundation for subsequent machine learning tasks and providing a comprehensive basis for analyzing the real estate market dynamics in Connecticut.

4.2 Removing Outliers

As a crucial step in enhancing the reliability and accuracy of our project, we conducted a comprehensive outlier removal process. Initially, we organized our dataset in ascending order, facilitating a clear understanding of the distribution from the smallest to the largest values. Subsequently, we employed the Interquartile Range (IQR) as a robust statistical measure to gauge the spread of values within our data. Utilizing the IQR, we established boundaries that delineate a "normal" range for values in specific columns. Any data point falling significantly below or above these boundaries was identified as an outlier—an unusual or extreme value. Recognizing outliers as potential errors or anomalies, we systematically removed them from the dataset. This meticulous process ensures that our subsequent analyses and machine learning models are grounded in a more representative set of data, free from the influence of aberrant values that could otherwise skew results or compromise the integrity of our findings.

4.3 Finding Geolocation

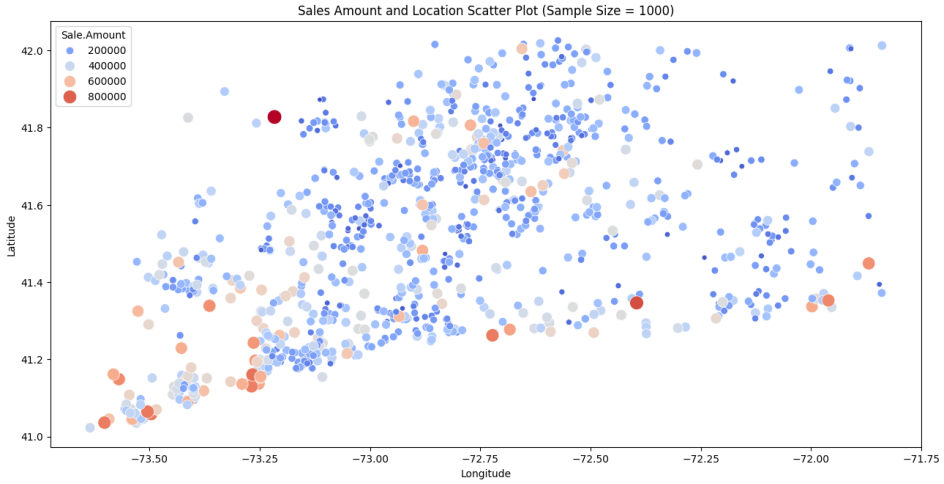

Figure 1: Relation between Geolocation and Sale Amount

The task involved enhancing the property dataset for Connecticut, USA, by incorporating geolocation data using the Google Geolocation API. A custom Python function was crafted to seamlessly integrate latitude and longitude information for each property based on its address and town. The function utilized the 'Address' and 'Town' columns from the dataset to create a consolidated list of addresses, merging them with the respective towns and appending the common string "CT, USA." Subsequently, the Google Geolocation API, powered by a provided API key, was employed to geocode each detailed address, extracting accurate latitude and longitude coordinates. The obtained geolocation data was then added as new columns, namely 'Longitude' and 'Latitude,' to the original dataset. This geocoding process significantly enriches the property dataset, allowing for spatial analysis and location-based insights in the subsequent stages of our project. The seamless integration of geolocation

data contributes to a more comprehensive understanding of the real estate landscape in Connecticut, facilitating diverse analytical perspectives and potential applications.

4.4 Clustering

In our property dataset, we have strategically employed k-means clustering to identify spatial clusters of houses based on their geolocation. The clustering process is driven by the understanding that houses located in proximity tend to form clusters, and the distance of a house to the centroid of its respective cluster correlates with its sale price. To determine the optimal number of clusters, we applied the elbow method, a statistical technique that identifies a point where the reduction in within-cluster sum of squares begins to diminish, indicating an optimal balance between precision and simplicity. After obtaining the optimal number of clusters, each house in the dataset was assigned to its respective cluster. The distance from each house to its cluster centroid was then calculated, providing a novel feature for our machine learning model. This distance feature serves as a valuable indicator of a property's proximity to a cluster centroid, which, as observed, influences its sale price. By integrating these cluster-based distance features, our machine learning model gains enhanced predictive capabilities, capturing nuanced spatial relationships within the Connecticut real estate market and facilitating more accurate estimations of property sale amounts.

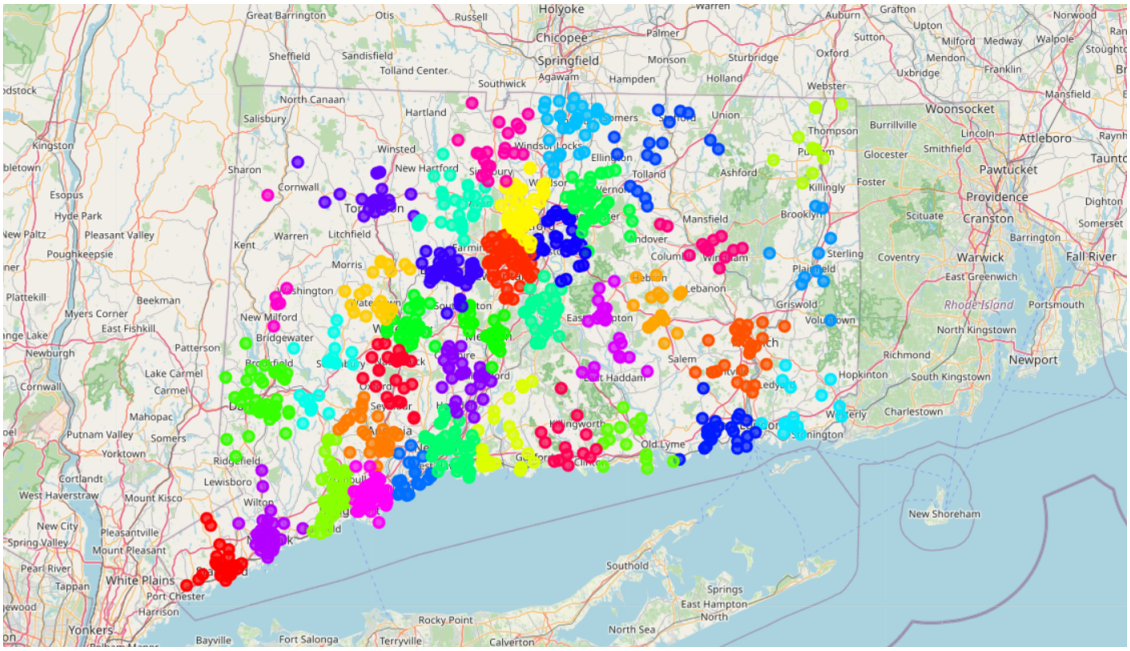

Figure 2: K means Clustering based on Location

4.5 Feature Selection

We generated a heatmap analysis for our property dataset, exploring the relationships between various features. Based on the heatmap results, we strategically selected features that

exhibit significant correlations, informing our subsequent analysis and model development. In our finalized selection of features for machine learning, we have strategically curated a set of variables that encapsulate critical aspects of the real estate data, ensuring a comprehensive representation for predictive modeling. These features include the 'Assessed.Value,' serving as a pivotal component representing the property's assessed monetary value for tax purposes.

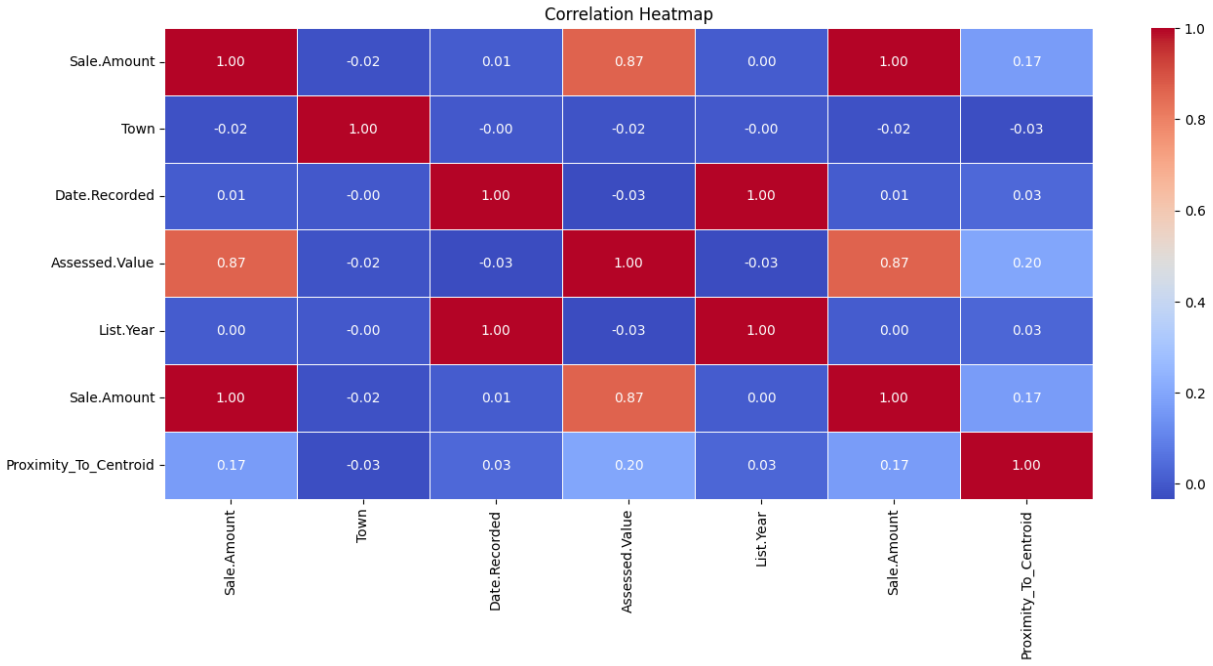

Figure 3: Heatmap for various features

The 'Date.Recorded' feature captures the temporal dimension of property transactions, and 'List.Year' offers insights into the year a property was listed. We retain 'Property.Type' to consider the property's categorical type and introduce the 'Proximity_To_Centroid' feature, derived from our k-means clustering analysis, indicating the spatial proximity of each property to its respective cluster centroid. Additionally, the 'geo' feature, composed of 'Longitude' and 'Latitude,' preserves the geographical coordinates crucial for spatial analysis. This refined feature set aims to capture multifaceted aspects of property dynamics, providing a robust foundation for our machine learning model to discern patterns and relationships within the Connecticut real estate market.

4.6 Preparing Machine Learning Model

The machine learning model is meticulously designed to predict 'Sale.Amount' for property prices based on a comprehensive analysis of historical data and various price indicators. The model construction is encapsulated in the BuildModel function, which comprises a pipeline incorporating a ColumnTransformer to handle key features such as 'Assessed.Value,'

'Date.Recorded,' 'Property.Type,' 'Proximity_To_Centroid,' 'List.Year,' and geographical coordinates ('Longitude' and 'Latitude'). Numerical features undergo standardization using StandardScaler, and the pipeline integrates a specified regressor for prediction. In the TrainModel function, data preprocessing includes handling missing values, encoding categorical features, and splitting the dataset into training and testing sets. The model is then trained on the training set. the training and testing split is executed within the TrainModel function using the train_test_split function from the scikit-learn library. In this snippet, the dataset is split into features (X) and the target variable (y). The test_size parameter specifies the proportion of the dataset to include in the test split (in this case, 40% for testing). The random_state parameter is set to ensure reproducibility by fixing the random seed. After this split, the training set (X_train and y_train) is used to train the machine learning model, while the testing set (X_test and y_test) is reserved for evaluating the model's performance . The TestModel function assesses the model's performance on the test set, providing critical metrics such as Mean Absolute Error (MAE) and R-squared (R2) score. This robust pipeline aligns with our overarching goal of predicting property 'Sale.Amount' based on thorough historical data analysis and diverse price indicators, ensuring that our model is well-equipped to capture and interpret the intricate dynamics of the Connecticut real estate market.

5 Methods

5.1 Random Forest Regression

Random Forest Regression is employed in our machine learning model for predicting Sale Amounts from the property dataset due to its robust and versatile nature. This ensemble learning algorithm builds multiple decision trees and combines their predictions to achieve a more accurate and stable result. Random Forest is effective in capturing complex relationships within the data, handling non-linear patterns, and minimizing overfitting. Its ability to provide feature importance rankings also aids in understanding which features play a crucial role in predicting Sale Amounts. This makes Random Forest a valuable choice for our model, allowing it to handle diverse factors influencing property prices in the Connecticut real estate market.

5.2 Linear Regression

Linear Regression is chosen for its simplicity and interpretability, making it a suitable algorithm to gain insights into the linear relationships between features and the target variable, Sale Amount. While Linear Regression assumes a linear connection between the features and the target, it still performs reasonably well when the relationship is approximately linear. Linear Regression's coefficients provide a clear understanding of how each feature contributes to the predicted Sale Amount. This transparency can be valuable for stakeholders seeking a straightforward interpretation of the factors influencing property prices.

5.3 Gradient Boosting Regression

Gradient Boosting Regression is selected for its powerful predictive capabilities and flexibility in capturing complex relationships. This algorithm builds an ensemble of weak learners sequentially, iteratively correcting errors made by previous models. Gradient Boosting excels in handling non-linear patterns, making it adept at capturing intricate nuances within the dataset. Its ability to optimize performance through boosting and the fine-tuning of hyperparameters contributes to its superior predictive accuracy. For our model, Gradient Boosting Regression proves beneficial in capturing the subtle dynamics that influence Sale Amounts, providing a nuanced and accurate prediction.

6 Result Analysis

In our analysis of three machine learning algorithms for property price prediction, Random Forest Regression emerges as a robust performer, showcasing a Mean Absolute Error (MAE) of 44308.42 and an impressive R-squared (R2) Score of 0.7979. The algorithm's strengths lie in its ability to capture intricate relationships and minimize overfitting, attributed to the ensemble of decision trees that collectively contribute to accurate predictions. Its adaptability to diverse data patterns makes it particularly well-suited for the complexities inherent in property price prediction. Moving to Linear Regression, with a MAE of 46770.26 and an R2 score of 0.7909, it demonstrates reasonable predictive performance, leveraging its simplicity and interpretability to provide valuable insights. However, Linear Regression's assumption of a linear relationship may limit its effectiveness in capturing more complex, non-linear patterns present in the data. Finally, Gradient Boosting Regression outshines the others with a low MAE of 42868.96 and a high R2 score of 0.8026. Its iterative boosting approach allows for fine-tuned adjustments, resulting in enhanced predictive accuracy, while its adaptability to non-linear patterns contributes to its superior results. In summary, each algorithm offers unique strengths, providing stakeholders with versatile tools for property price prediction in the dynamic real estate landscape.

Algorithm	**Mean Absolute Error (MAE)**	**R-squared (R2) Score**
Random Forest Regression	44308.42	0.7979
Linear Regression	46770.26	0.7909
Gradient Boosting Regression	42868.96	0.8026

Table 1: Performance Metrics of Machine Learning Algorithms for Sale Amount Prediction

All three models demonstrate a solid performance in predicting Sale Amounts from the property dataset. Gradient Boosting Regression stands out with the lowest MAE and the highest R-squared score, indicating superior predictive accuracy. The choice of the most suitable model depends on specific priorities, with Random Forest offering robustness, Linear Regression providing interpretability, and Gradient Boosting excelling in accuracy. Further fine-tuning of hyperparameters and model evaluation could enhance the performance of each algorithm.

7 Conclusion

In conclusion, our comprehensive exploration of the real estate market in Connecticut, USA, has been driven by the primary goal of empowering potential property buyers with the necessary tools for well-informed investment decisions. The critical challenge of accurately predicting sale amounts of house properties has been addressed through a meticulous data-driven approach, leveraging big data, cutting-edge analytics techniques, and advanced machine learning methodologies.

Our dataset, comprising a substantial number of housing sales records across several columns, serves as a rich foundation for analysis. The task of predicting sale amounts involves not only historical data but also innovative features derived from existing ones, including geolocation data and spatial clusters. The machine learning model, incorporating Random Forest Regression, Linear Regression, and Gradient Boosting Regression, is designed to discern intricate patterns and relationships that influence property sale prices.

Key techniques such as data cleaning, outlier removal, geolocation addition, clustering, and feature selection contribute to a refined dataset and a robust machine learning model. The selected features, guided by heatmap analysis, aim to capture various aspects of the real estate landscape, ensuring a nuanced understanding of factors influencing property values.

The analysis of our machine learning model results indicates distinct strengths for each algorithm. Random Forest Regression demonstrates robust predictive performance, excelling in capturing complex relationships and minimizing overfitting. Linear Regression provides reasonable predictive capabilities with simplicity and interpretability, while Gradient Boosting Regression outperforms others with superior accuracy and flexibility in handling complex relationships.

In summary, our approach combines traditional and cutting-edge methods, providing stakeholders with versatile tools for property price prediction. This report lays the groundwork for further exploration and collaboration in understanding the intricacies of the Connecticut real estate market.