# Predicting Connecticut Real Estate Prices Using Random Forest Regression: A Distance-Based Analysis

EXECUTIVE SUMMARY

## I. Problem & Hypothesis

- Research Question: Can a Random Forest Regression model predict Connecticut real estate prices with a R^2 score above 0.70 using distance to coast and distance to New York as key predictive features?

- Data analysis through Random Forest Regression and Multiple Linear Regression modeling will provide an insightful understanding of the relationship between property features and house prices.

- A model developed by North Carolina researchers shows coastal property values increasing despite growing risks from sea-level rise (Why Coastal Property Values Grow Despite Climate Risks, 2024). Connecticut properties may be influenced by both coastal distance and their distance from New York.

- Geographic analysis using distance metrics provides unique insights into the Connecticut real estate market, where property values are potentially influenced by both coastal distances and proximity to major states like New York.

- The analysis combines both traditional property characteristics (living area, bedrooms, bathrooms) and economic indicators (CPI, unemployment rate, federal interest rates) to create a comprehensive prediction model that accounts for market conditions.

- **Null hypothesis**: The Random Forest model using distance to coast and distance to New York as key predictive features cannot predict Connecticut real estate prices with a R^2 score greater
  than 0.70.

- **Alternate Hypothesis**: The Random Forest model using distance to coast and distance to New York as key predictive features can predict Connecticut real estate prices with a R^2 score
  greater than 0.70.

## II. Data Analysis Process

- Collection: The real estate dataset was obtained from a publicly available Kaggle source which collected property listings over 8 months. The shapefiles were retrieved from CT DEEP and NY's Civil Boundaries program.

- Extraction: Variables that were related to property characteristics and property price were selected from the Connecticut real estate dataset.
  - Independent variables: loglotArea, livingArea, latitude, longitude, num_full_baths.x, num_half_baths, num_bedrooms.x, days_on_market, yearBuilt.x, fed_rate, lagged_CPI,
    lagged_unemployment, volatility_value, distance_to_coast, distance_to_new_york
  - Dependent variable: price

- Cleaning: Data was cleaned by checking for duplicates, errors, and missing variables. Outliers were checked and removed if outside of the IQR ranges.
- Preparation: Univariate analysis, bivariate analysis, and correlation heat map. The data sparsity was checked with 0%.
- Analysis:
    - Shapiro-Wilk Test:
        - Shapiro-Wilk Statistic: 0.922416
        - Shapiro-Wilk Test P-Value: 0.00
    - Random Forest Regression Analysis: Utilized in this analysis to understand the relationship between property characteristics and property price.
        - Identified the feature importance scores
        - Generated residual plots
        - Preformed cross-validation
        - Normality testing on dependent variable
    - Multiple Linear Regression Analysis: Used as a complimentary baseline comparison with Random Forest Regression to gain additional understanding in how property characteristics directly affect property price through the coefficients.
        - Used backward elimination by checking both VIF values and P-values to reduce an initial model to a refined version with just 8 features.
        - Log transformed dependent variable
        - Identified the coefficients
        - Generated residual plots
        - Preformed cross-validation

## III. Findings

- The following statistically significant property characteristics associated with property prices were identified during Random Forest Regression as: livingArea, longitude, latitude, yearBuilt, and distance_to_new_york.

- The Random Forest Regression analysis indicated the model is reasonably reliable for estimating Connecticut house prices based on the provided features.
    - R^2 score: 0.7274 - indicating about 73% of the variability in the data can be explained by the independent variables. This indicates a strong predictive performance, beating the target R^2 threshold of 0.70.
    - RMSE: 84,120.73 – this is the typical size of prediction errors, showing that most predictions fall within this range of the actual house price.
    - MAE: 59,747.44 – this indicates that, on average, the model's predictions differ from actual house prices by around 60,000. Given that house prices in the dataset range from 100,000 to 800,000 after outliers were handled, these error metrics suggest the model performs well for real estate price prediction.

- The Random Forest model achieved an $R^2$ score of 0.73 on the test set, exceeding the 0.70 threshold. Therefore, the null hypothesis can be rejected, demonstrating that proximity to the coast and New York City, along with other property features, can reliably predict house prices in Connecticut.

- The Multiple Linear Regression analysis showed several significant coefficients that indicate how different features affect house prices in Connecticut:

    - The model identified positive coefficients for: livingArea, num_full_baths, loglotArea, yearBuilt, and num_half_baths, indicating these features increase property values when they increase.
    - The negative coefficients found: distance_to_new_york, longitude, and distance_to_coast, suggesting that property values decrease as these distances increase or as properties are located further east in the state.

- The Multiple Linear Regression analysis provided a complementary approach to predicting Connecticut house prices, though with lower accuracy than the Random Forest model.
    - R^2 score: 0.6034 - indicating about 60% of the variability in house prices can be explained by the selected features. While this shows moderate predictive power, it falls below the target threshold of 0.70.
    - RMSE: 101,463.73 – this represents the typical prediction error range, showing slightly higher uncertainty compared to the Random Forest model.
    - MAE: 70,491.12 – this indicates that, on average, the model's predictions differ from actual house prices by around 70,000. While less accurate than the Random Forest model, the MLR provides valuable insights into individual feature impacts on price.

- Overall, the combination of both models provides a great way for understanding Connecticut's real estate market, with Random Forest offering superior predictive power and MLR providing understandable insights into feature relationships. Using multiple models enables both accurate price predictions and clear explanations of market dynamics, making it valuable for various stakeholders in the real estate industry.

## IV. Limitations

- The dataset only included real estate listings from June 2022 to February 2023, which made it difficult to see how seasons affected house prices throughout a full year.

- Prior to conducting the analysis, the distance calculations for each property from the coast and New York were calculated using map coordinates because these important measurements were not in the original dataset.

- Some information in the dataset, like crime rates, showed the same values for all Connecticut properties, making these features useless for predicting house prices.

- The Random Forest model could only predict prices similar to what it had seen before, which meant it struggled with very expensive houses or unusual market conditions.

- The Multiple Linear Regression model assumed straight-line relationships between features and prices, which was too simple for the complex real estate market.

- Many location-related features were too closely related to each other, requiring the removal of certain features which potentially contain useful information.

- The analysis required removing many properties from the dataset, such as those outside of Connecticut and non-single-family homes, which might have limited how well the results apply to other situations.

- While the study showed connections between house features and prices, it couldn't prove these features directly caused the price changes.

## V. Recommended Actions

- The Random Forest model performed well at predicting house prices ($R^2 = 0.73$), indicating it can be used as the primary prediction tool, while the Multiple Linear Regression model can be used to explain price factors to potential homebuyers.

- Future studies could collect historical data to understand seasonal changes in house prices and examine how economic conditions influence the importance of location features. Using a larger dataset over a longer time-period may improve these models.

- Another suggestion for future studies is dividing Connecticut into regions for more detailed local market analysis, helping understand how coastal and New York City proximity effects vary across the state.

- Additional features should be added to improve the models, including demographic information, economic indicators, and neighborhood characteristics like proximity to other important geographic points. Distances from schools, hospitals, cities, and public transportation are all potential candidates to gain a better understanding of the localized market.

- The models should be tested with data from different time periods and similar housing markets in other states to validate their performance under various market conditions.

- Further analysis should focus on how property values change with distance from key locations and regional differences in feature importance across Connecticut, helping create more accurate local market predictions.

## VI. Benefits of Study

- The study successfully demonstrated that a Random Forest model could predict Connecticut house prices with high accuracy, helping real estate professionals make better pricing decisions.

- Through Multiple Linear Regression analysis, the research identified specific features that influenced house prices, with living area showing the strongest positive effect and proximity to New York City showing a significant negative effect, providing clear insights for market participants.

- The research rejected the null hypothesis by showing that distance-based features, along with other property characteristics, could predict house prices above the target accuracy threshold.

- The combined use of two different models offered both accurate predictions and

easy-to-understand explanations of price factors, making the findings useful for real estate agents, homebuyers, and sellers. Tools for potential buyers could even be created by using this model to predict prices across the state.

- The study revealed important relationships between property location and value in Connecticut, showing how proximity to the coast and New York City significantly affected house prices.

- The analysis contributed to the understanding of Connecticut's real estate market by quantifying the impact of various features on property values, helping improve market transparency and decision-making.

Panopto Video Presentation Link:
https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=85b2b8b5-dd13-4b4e-853a-b036011ff1f3