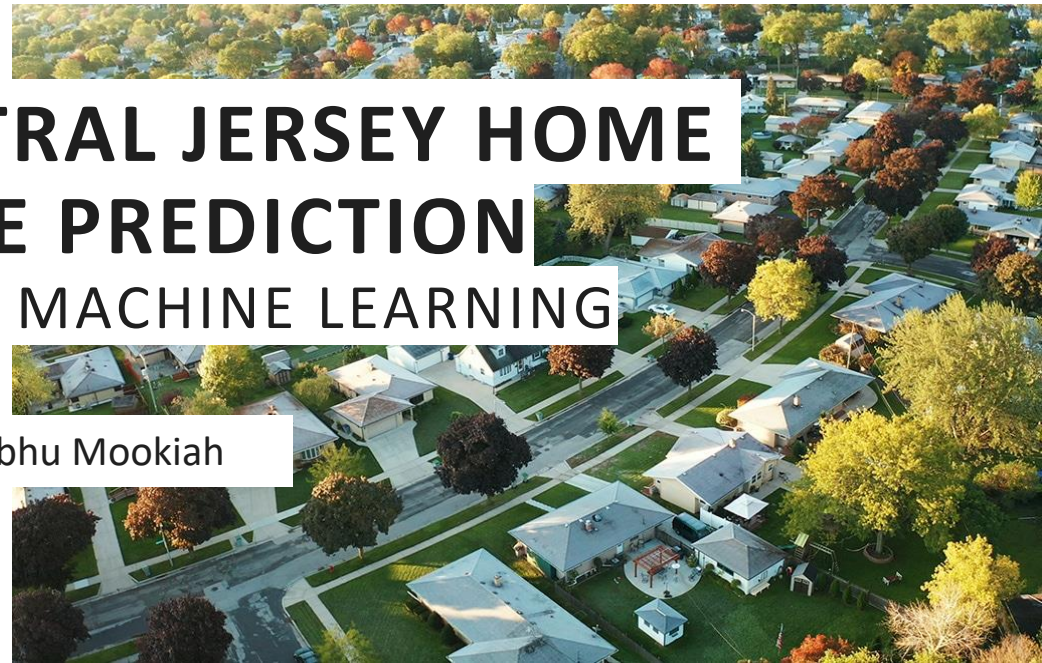
CENTRAL JERSEY HOME PRICE PREDICTION

USING MACHINE LEARNING



Senthil Prabhu Mookiah

AGENDA

Overview

Objectives

Methodology

Data Exploration & Prep

Model Evaluation & Results

Conclusion & Future Work

PROJECT OVERVIEW:

- This project aims to predict home prices in Central Jersey (specifically targets Edison, East Brunswick, South Brunswick, Plainsboro and West Windsor neighborhood) using various machine learning techniques.
- The primary dataset used is the Zillow Central NJ Home Sale Price dataset (Last 3 years data scrapped using Z Real Estate scrapper for Zillow chrome plugin).

PROJECT OBJECTIVES:

- To analyze the factors influencing home prices in Central Jersey.
- To build a predictive model that accurately estimates home prices based on relevant features.
- To provide insights that can help me & my family make informed decisions regarding real estate investments.

METHODOLOGY:

1. **Data Collection:**

   Extract data from the Zillow Central NJ Home Sale Price dataset.

2. **Exploratory Data Analysis (EDA):**

   Analyze the data to identify trends, correlations, and patterns.

3. **Data Preprocessing:**

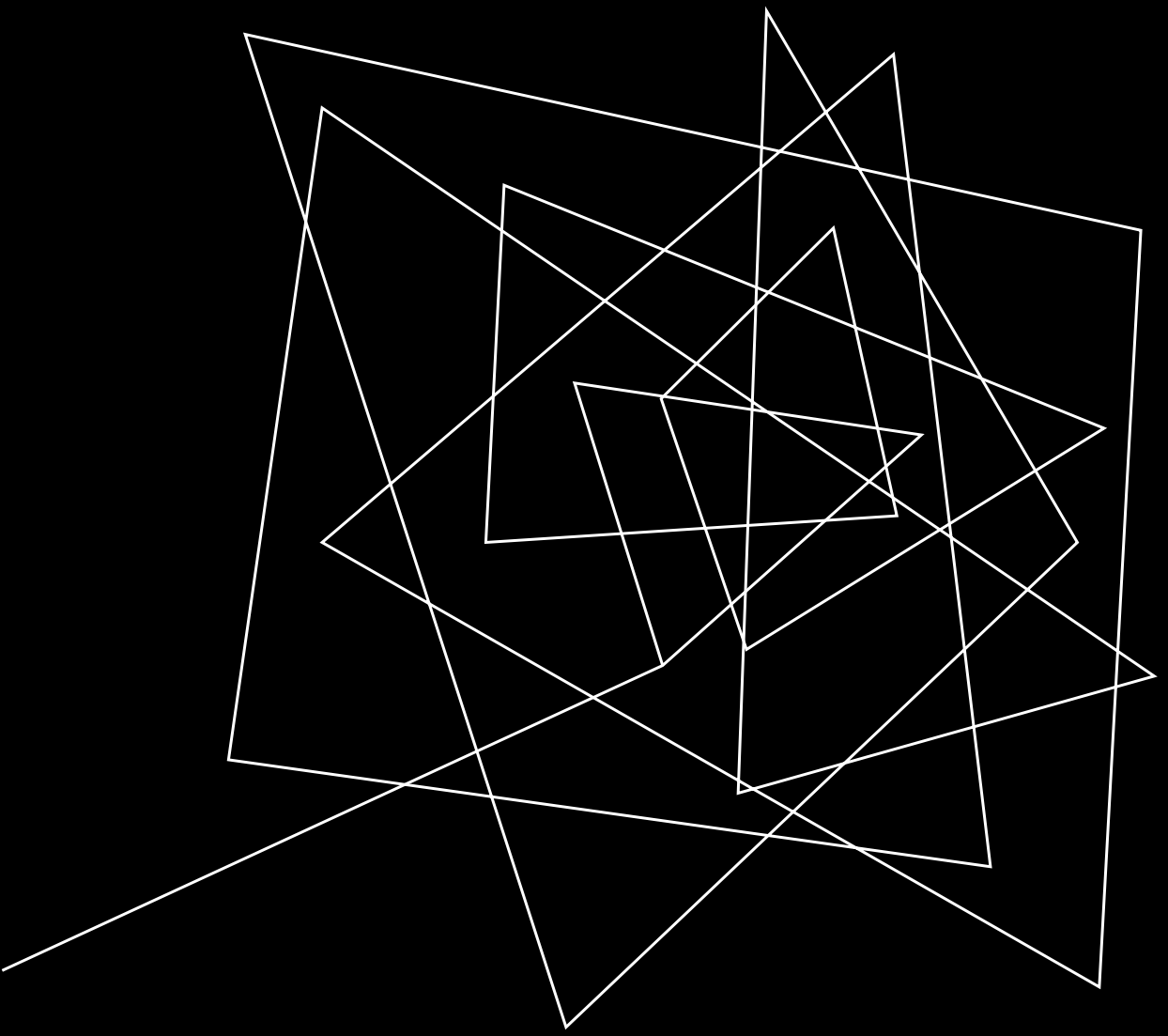
   Clean and prepare the data by handling missing values and encoding categorical variables.

3. **Model Building:**

   Train various machine learning models and select the best performing one.

4. **Model Evaluation:**

   Evaluate the model using appropriate metrics to ensure its accuracy and reliability.



EXPLORATORY DATA ANALYSIS

DATA ANALYSIS

Zillow Attributes

Last Date Sold, Latitude, Longitude, Zip, Price, Price Per Sq. Ft., Lot Area, Lot Area Units, Beds, Baths, Footage, Address, Zestimate, Days on the Market (For Sale), Year Built, Stories, Heating, Cooling, Fireplaces, Flooring, Foundation, Garage Capacity, Sewer, Subdivision, Zoning Desc., Property Condition, Roof Type

Key Attributes

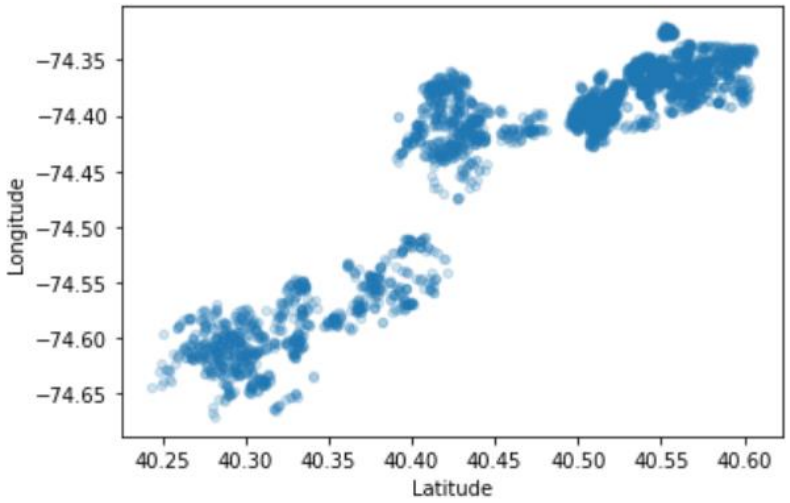
- Last Date Sold
- Latitude
- Longitude
- Zip Code
- Price
- Price Per Sq. Ft.
- Lot Area
- Beds
- Baths
- Footage
- Zestimate

Correlation

- Price 1.000000
- Zestimate 0.873873
- Baths 0.735028
- Footage 0.734201
- Beds 0.346140
- Price Per Sq. Ft. -0.006659
- Lot Area -0.006781
- Latitude -0.190306
- Longitude -0.251584
- Zip -0.255347

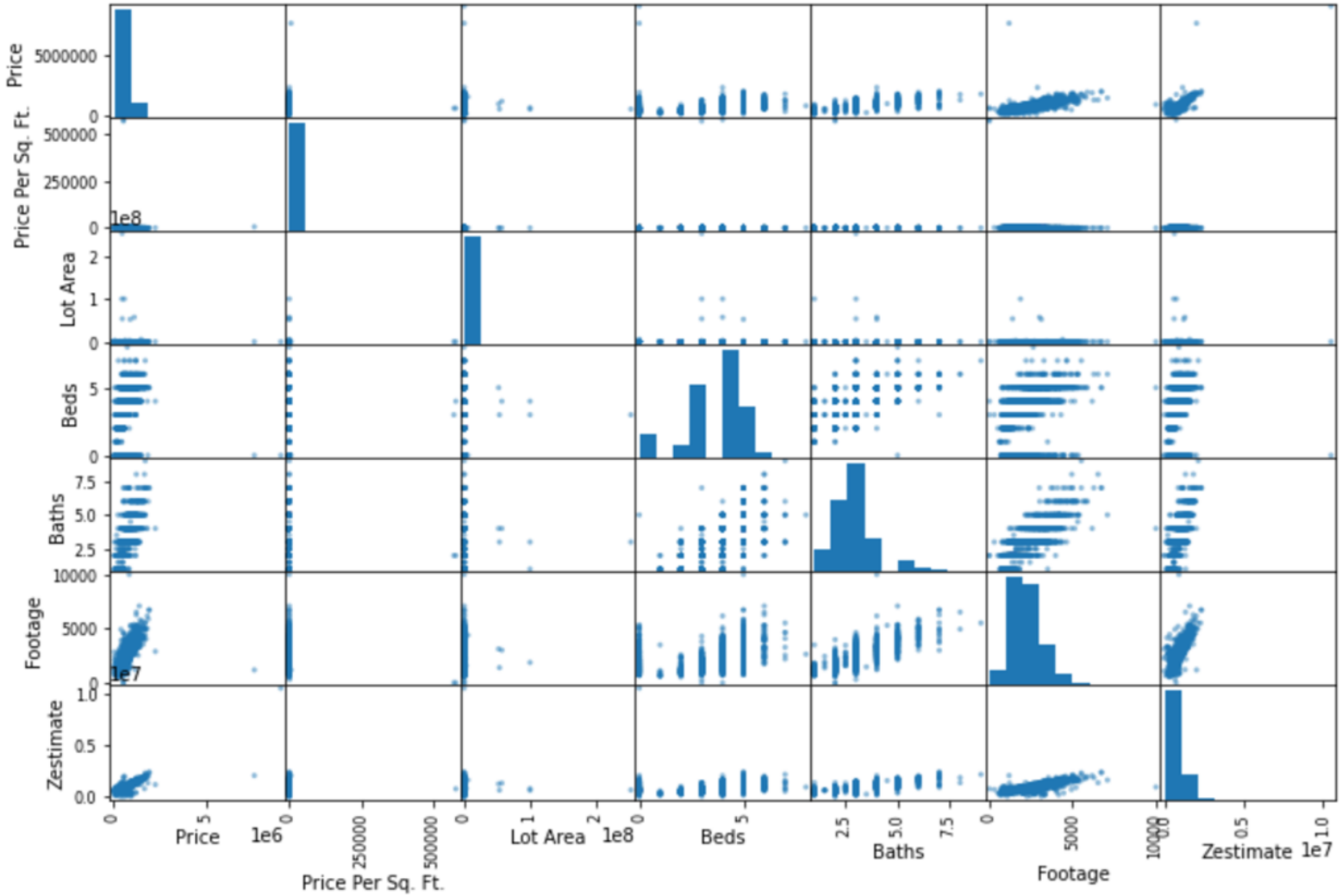
Significant correlations found between Price and Zestimate, Footage, and Baths.

SCATTER MATRIX



Prepare Data for ML

- Scikit-Learn Imputer class to fill missing values with Median
- Remove the text attributes because Imputer works on numerical attributes
- Transform Zip code category using One-Hot dense encoding
- Apply Feature scaling on all numerical attributes



CHOOSING THE RIGHT ALGORITHM

- Linear Regression
- Decision Trees
- Random Forests

HOW WELL DOES OUR MODEL PERFORM?

Features	Importance Score
Zestimate	0.767
Price Per Sq. Ft.	0.160
Footage	0.029
Lot Area	0.162
Baths	0.011
Beds	0.007
Zip code	0.001

Feature Importance:

The Zestimate, price per square foot, footage, and lot area are the most influential factors in predicting home prices.

Training and Test Set:

- Length of training set: 3297
- Length of test set: 825

METRIC	Linear Regression	Decision Tree	Random Forests
Root Mean Squared Error (RMSE)	1,65,458.60	3,373.48	61,284.09
Cross Validation Mean RMSE	1,51,631.23	1,52,574.44	1,37,188.61
Standard deviation	72,402.00	1,12,566.69	1,18,150.41

Model Comparison: The Decision Tree model initially appears to have the lowest RMSE, suggesting it is the best-performing model. However, the cross-validation results show a high standard deviation for the Decision Tree, indicating its performance may vary significantly depending on the specific data split.

Hyper Parameter Tuning:

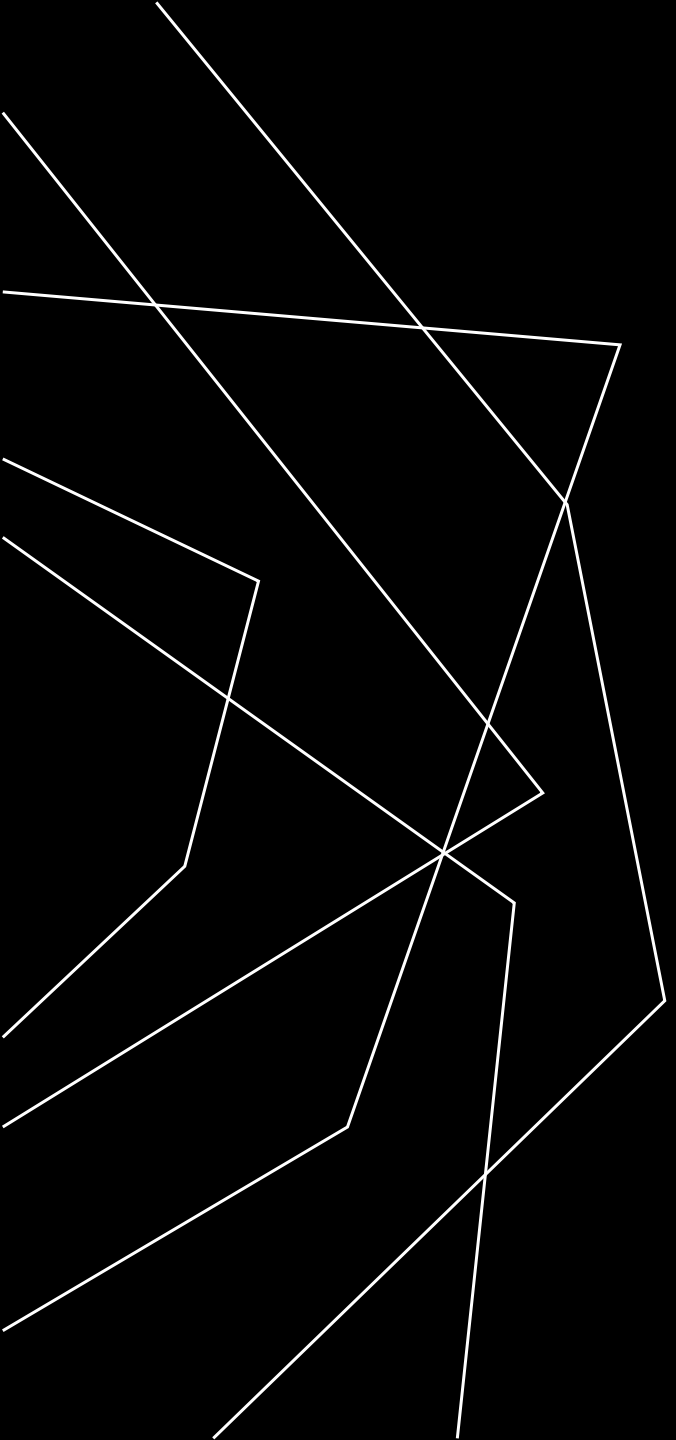
Final Root Mean Squared Error (RMSE) : 5,42,429.63

CONCLUSION

- The Decision Tree model initially showed promise in predicting home prices with the lowest RMSE on the test set.
- However, after hyperparameter tuning using GridSearchCV, the final RMSE significantly increased to 542,429.63, suggesting potential overfitting.
- This highlights the importance of careful hyperparameter tuning to avoid overfitting and ensure robust model performance.

NEXT STEPS

- **Re-evaluate Hyperparameter Tuning**: Overfitting may have occurred with GridSearchCV for the Decision Tree model. Re-examine the hyperparameter grid, considering a wider range of values, especially those affecting overfitting like tree depth and minimum samples per split.
- **Consider Alternative Tuning Methods**: Explore other techniques such as Random Search or Bayesian Optimization for more efficient search space exploration and better generalization.
- **Regularization Techniques**: Try regularization methods like pruning for Decision Trees or ensemble methods like Random Forests to mitigate overfitting and improve generalization.
- **Feature Selection**: Revisit feature selection and engineering to ensure the most relevant features are used, reducing overfitting and enhancing performance.



THANK YOU

Questions and Discussion

Senthil Prabhu Mookiah