Housing Price Prediction Project

Introduction and Background

Literature Review

Housing prices are unpredictable especially due to the many factors impacting the prices of houses. Because of unpredictable housing prices, a machine learning model would be beneficial in determining how housing prices may change.

Adetunji et al. (2022) discuss their RandomForest model that predicts housing prices. They identified significant factors that play a role in housing prices such as crime rate, number of rooms, and accessibility to highways to approximate how high the housing price would be based on these attributes (Adetunji et al., 2022, p. 808-812).

Vineeth et al. (2018) looks at different algorithms like simple linear regression, multiple linear regression, and neural networks for predicting housing prices. The researchers examined factors such as price, number of bedrooms/bathrooms, square feet, and etc. From these algorithms, it was found that the neural networks perform the best (Vineeth et al., 2018).

Gupta et al. (2022) applied a Bayesian dynamic factor model to analyze housing market trends by assessing how macroeconomic uncertainty influences state housing prices. The model found that national factors accounted for 35% of housing price growth. Additionally, a random forest model helped identify key drivers of macroeconomic uncertainty and predict the best-fit model for national house prices.

Dataset Description

NY Housing

The dataset includes prices for 4,800 New York houses with features like house type, size, and location. It helps analyze how these factors influence pricing but lacks historical data for predicting trends over time.

California Housing

This dataset covers 20,640 California houses with features like house age, median income, and location. It enables analysis of how income and house characteristics affect pricing.

Dataset Links

- NY Housing Dataset
- California Housing Dataset

Problem Definition

We aim to tackle the unpredictability of housing prices, which affects buyers, sellers, and developers. Housing data is often noisy and non-linear, so using machine learning models allows for more accurate price predictions. This approach enhances pricing transparency, helps stakeholders make better decisions, and reduces risks in the housing market caused by inaccurate forecasts.

Methods

Data Preprocessing

- scikit-learn StandardScaler to standardize numerical data and improve convergence of the machine learning models. TFIDF Vectorization for categorical and text data.
- pandas Series.string to extract keywords from the type of house and address in the NYC data set.
- pandas DataFrames to merge the two datasets into one DataFrame.

- geopy package Nominatim to convert addresses in the NYC data set into longitude and latitude points to match the California dataset.
- After merging, use sklearn haversine_distances to calculate distances to major city points, schools, and transportation to generate more insights.

Machine Learning Models

- Supervised learning LinearRegression, RandomForest, and MLPRegressor models are applicable since we have clearly defined attributes (median household income, price, location, etc) and a target value (price of the house).

For our implemented solution, we incorporated LinearRegression, MLPRegressor (Neural Network), and RandomForest ML models to predict the housing price based on the features we provided to the models. For choosing the model, we tried many models like random forest and k-means clustering, but we found that LinearRegression and MLPRegressor were the most consistent and statistically valid. Supervised learning LinearRegression and Neural Network models are applicable since we have clearly defined attributes (median household income, price, location, etc) and a target value (price of the house). However, the Random Forest model provided valuable insights during the testing phase, as it helped highlight nonlinear interactions between features and identify important predictors, such as median household income. Supervised learning methods like LinearRegression, MLPRegressor, and Random Forest are applicable since we have clearly defined attributes (median household income, price, location, etc.) and a target value (price of the house). We also chose to test these models because Vineeth et al. (2018) evaluates linear regression and neural networks where the model takes in square foot of living, number of bedrooms, bathrooms, year build, and number of floors, but does not take longitude, latitude, and mean income for consideration which our dataset contains. Additionally, the dataset from the paper contains data from King County, but our model’s data contains data from California and New York City which allows us to see how our model would perform from two distinct locations. Random Forest was a popular model used in Adetunji et al. (2022) paper regarding housing price prediction so we decided to incorporate this into our testing to see how it would impact our analysis and enrich our understanding.

For data preprocessing, we used sklearn StandardScaler to standardize numerical data features for the machine learning models while also utilizing feature selection to keep only relevant features such as median_income and remove unnecessary features such as address. We also used pandas to create DataFrames to merge the two datasets. In addition to this, since we are combining 2 datasets, we needed to handle missing data with some NaN values that were replaced with empty string and 0 values to ensure data stayed consistent and wasn't lost during preprocessing. We also utilized TfidfVectorizer from sklearn to identify significant words from the CSV file, which in our case was used to identify the California and New York dataset when we combine the data with our features.

We split the data for the model using train_test_split() where we split the data into 80% train and 20% test. For scaling for the Neural Network model, we used Sklearn StandardScaler with fit_transform() on our training data and transform() on our testing data. After splitting the data, we trained the model using the MLPRegressor model from sklearn with 100 hidden layers and 500 iterations and the fit method on our x and y train data. For the LinearRegression model and the RandomForest models, there was no scaling and we used sklearn LinearRegression RandomForest models respectively and used the fit method to fit the x and y train data.

Results and Discussion

Quantitative Metrics

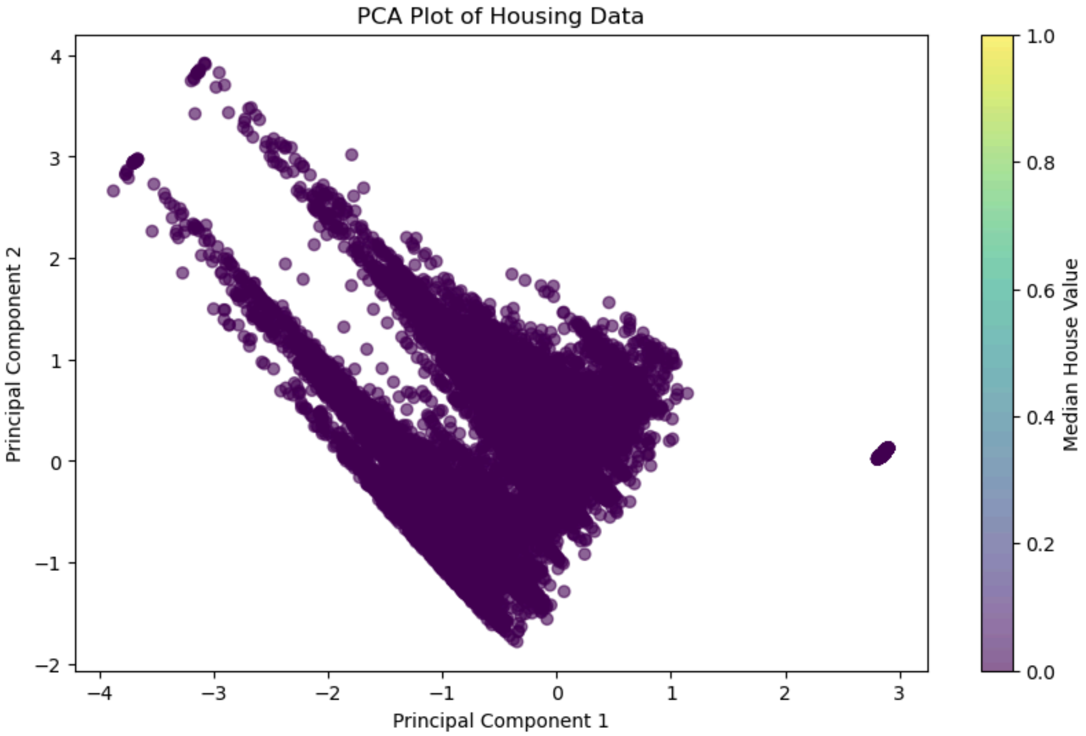
- **Mean Absolute Error (MAE)**: Chosen because Linear Regression is a regression model, and MAE measures the average prediction error in the same unit as the target (dollars).
- **Root Mean Squared Error (RMSE)**: Emphasizes larger errors more than MAE.
- **R-squared (R^2)**: Indicates the proportion of variance in house prices the Linear Regression model explains.

Project Goals

- **MAE**: Less than $10,000
- **RMSE**: Less than $15,000
- **R^2**: Greater than 0.80

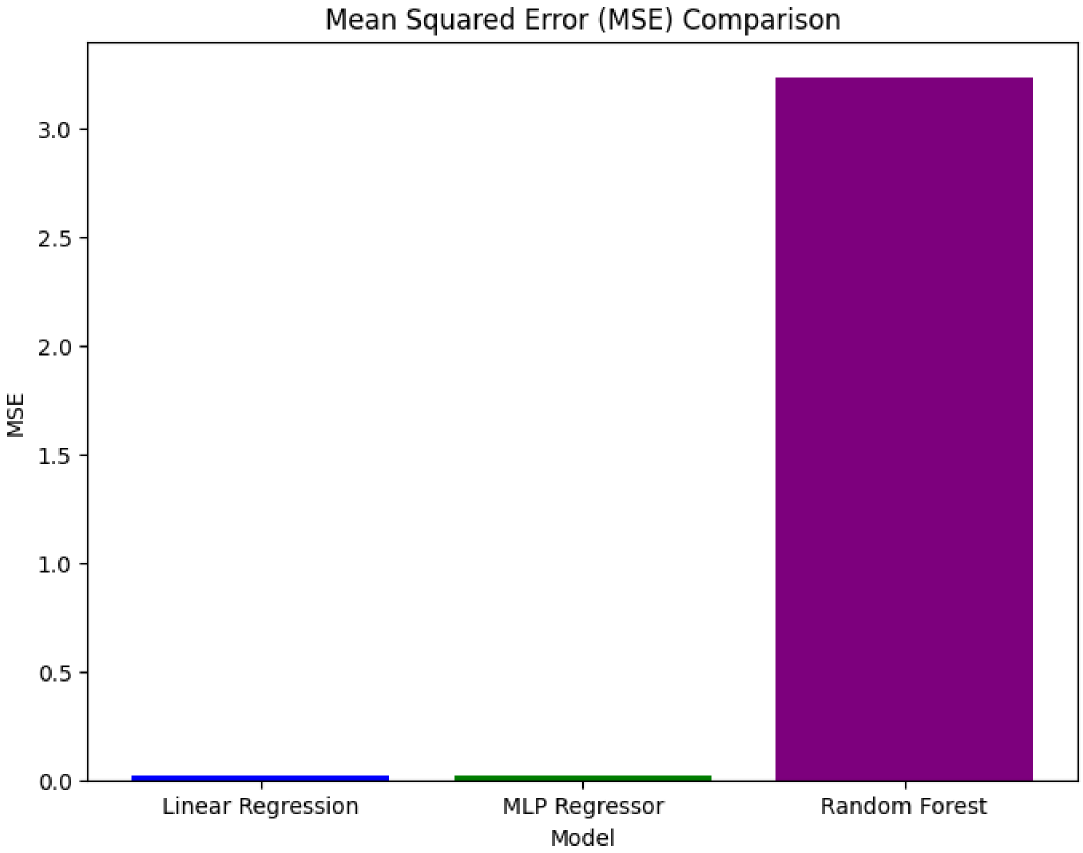
Expected Results

Linear Regression provides reliable price predictions, minimizing error and explaining most price variability, though it may struggle with non-linear relationships. MLP Regressor can capture more complex patterns, while RandomForest highlights the most influential features for predicting housing prices.

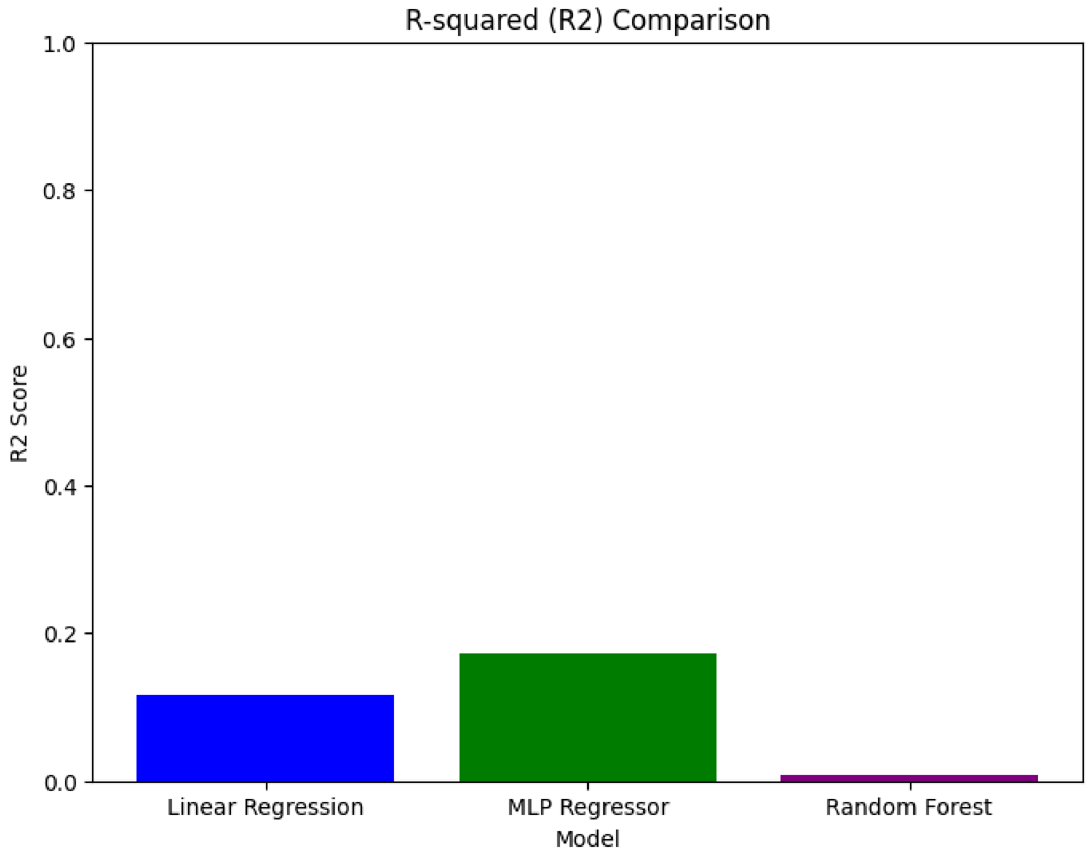


This displays the PCA visualization for our combined housing dataset reduced to two principal components. The color bar represents the range of normalized median house values based on the geographical influence.

We used Mean Squared Error (MSE) and R-squared correlation as the quantitative scoring metrics for both models.



Mean Squared Error (MSE) measures the average squared differences between the actual values and predicted values. MSE is chosen as a metric because it penalizes larger errors more than smaller ones, making it sensitive to outliers. It is a valuable metric to indicate how well a model is performing, with lower values implying better fit – ideally close to 0. We train two models, a Linear Regression model and MLP Regressor, to predict normalized housing values. The visualization shows that the Linear Regression Model has a MSE of 0.0186 and the MLP Regressor Model achieved a slightly lower MSE of 0.0175, reflecting that the performance of both models is highly accurate and fits the housing data well. This could likely result from the longitude and latitude features of the dataset being good identifiers that the models can use for more accurate prediction of the price of a home. Additionally, the California Dataset contains median incomes, which is an important indicator of housing prices, contributing to the accuracy of both models. The Random Forest Model has a dramatically much higher MSE of 3.2336, reflecting a significantly poorer performance compared to both the Linear Regression and MLP Regressor models. This could be attributed to overfitting to noisy training data or poor hyperparameter tuning. While features like median income and geographic coordinates are key identifiers to predict housing prices, the Random Forest Model may overfit to noisy or less relevant patterns in the data, leading to a reduced generalization. Another reason why the model performs poorly is because the hyperparameters such as the number of trees or maximum depth are not optimized. Overall, the performance of the MLP Regressor model is the most accurate with a lower MSE because of its neural network structure, which uses hidden layers to train and optimize weights, allowing it to capture complex relationships that the Linear Regression and Random Forest models cannot.



The R squared value is a representation of how well the model fits the data based on how well the model's inputs account for changes in the predicted data. The R squared value is an integer ranging from 0 to 1 and the higher the value, the better the model fits the data. This metric is appropriate for determining the quality of the linear regression model and MLP regressor model because it indicates how well the model explains the variability in the predicted house prices. After running the two models on the housing data, the visualization for this metric shows an R squared value of around 0.12 for linear regression and an R squared value of 0.17 for MLP regressor. Given that the scores are low, this indicates that both models explain only a small portion of the variation of the housing prices, which means they are not capturing the key factors that impact the data. These low R squared values could likely be due to insufficient features in the housing data as it doesn’t include variables such as crime rates, proximity to business districts, school ratings, or age of the home which are reflective of housing markets. Additionally, the California dataset includes the median income while the New York dataset doesn’t, and the

New York dataset includes the type of home which the California dataset doesn’t, and this could have also led to a variation in the housing predictions. The R-squared value for the Random Forest model is 0.00894, which is significantly lower than the R-squared values of the linear regression model (0.12) and MLP regressor model (0.17) This means that the Random Forest model accounts for very little variability in the housing prices. This indicates that the model isn't capturing key patterns in the housing data, and just like the linear regression and MLP models, this could likely be due to the lack of important variables. Additionally, Random Forest models create decision trees, and since there are few features in this dataset, the model might not be able to efficiently learn from them. This might have caused overfitting or underfitting to the data due to improper hyperparameter tuning. Overfitting happens when the model becomes too complex, learning noise in the training data, which affects its ability to generalize new data. Underfitting occurs when the model is too simple and fails to capture the important patterns, leading to poor performance even on the training data. Both issues can occur from not properly tuning hyperparameters such as the number of trees, the depth of trees, or the number of features considered for splits. Overall, while the MLP regressor has a higher R-squared and is better suited for non-linear relationships, all 3 of the models don’t provide accurate predictions, which suggests that adjustments are needed to the features included in the data.

For the next steps in improving the performance of our models on the housing dataset, we plan to focus on hyperparameter tuning and model expansion. Specifically, we will experiment with various hidden layer sizes, learning rates, maximum iterations, and activation functions to optimize the MLP regressor. For the Random Forest model, we can also explore tuning parameters such as the number of trees, tree depth, and the number of features considered at each split to improve its predictive power. Additionally, we will apply cross-validation techniques to improve the accuracy of our model metrics. To improve the performance of our initial models, we also plan to expand the dataset by adding additional features indicative of the housing market, such as public transportation and employment rates. Through these steps, we hope to develop the models and obtain more accurate housing predictions.

References

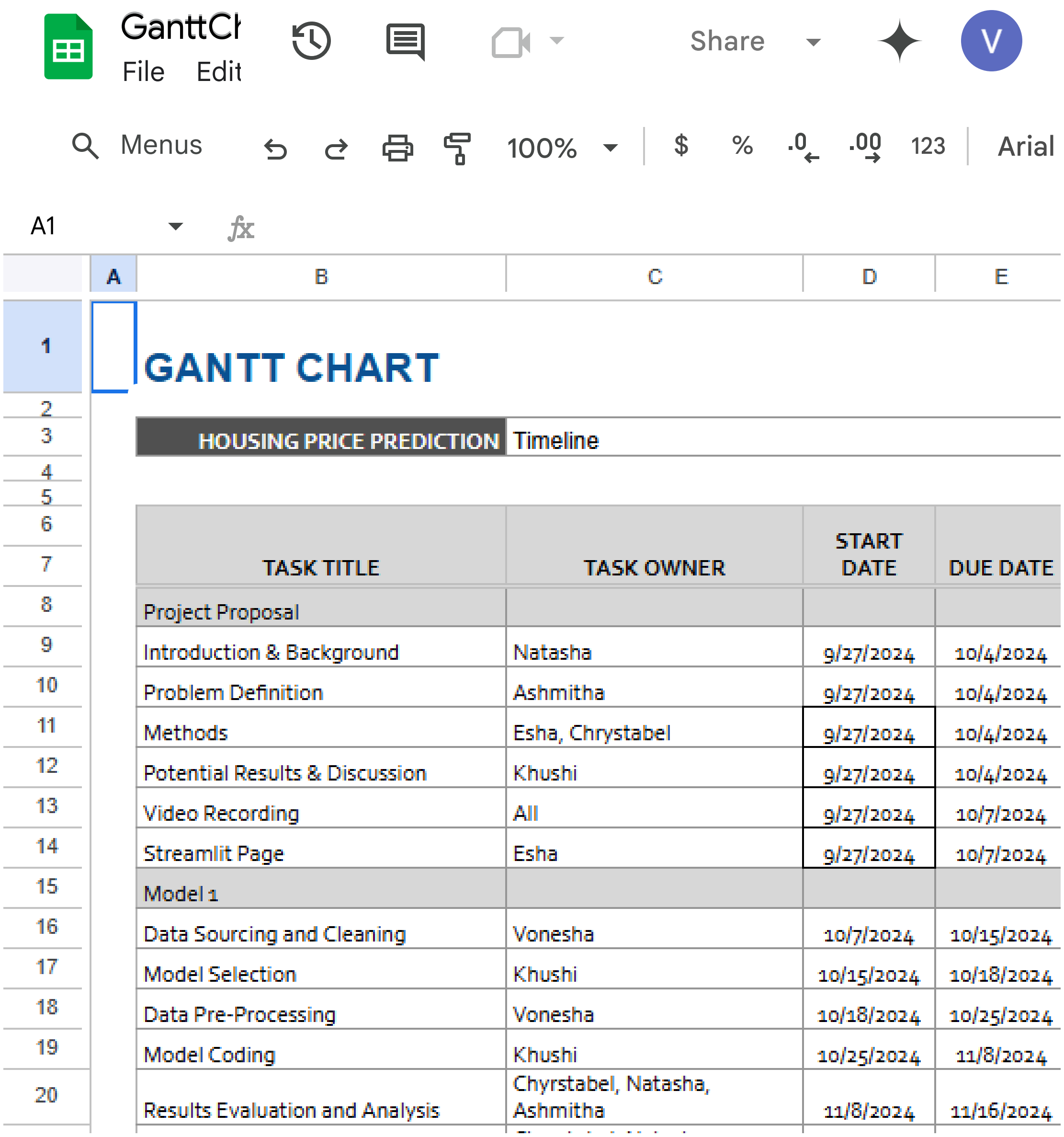
Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique.

Procedia Computer Science, 199, 806-813.

Gupta, R., Marfatia, H. A., Pierdzioch, C., & Salisu, A. A. (2022). Machine learning predictions of housing market synchronization across US states: the role of uncertainty. *The Journal of Real Estate Finance and Economics,* 1-23.

Vineeth, N., Ayyappa, M., & Bharathi, B. (2018). House price prediction using machine learning algorithms. *In Soft Computing Systems: Second International Conference, ICSCS 2018, Kollam, India, April 19–20, 2018, Revised Selected Papers 2* (pp. 425-433). Springer Singapore.

Gantt Chart



GANTT CHART

HOUSING PRICE PREDICTION Timeline

TASK TITLE	TASK OWNER	START DATE	DUE DATE
Project Proposal			
Introduction & Background	Natasha	9/27/2024	10/4/2024
Problem Definition	Ashmitha	9/27/2024	10/4/2024
Methods	Esha, Chrystabel	9/27/2024	10/4/2024
Potential Results & Discussion	Khushi	9/27/2024	10/4/2024
Video Recording	All	9/27/2024	10/7/2024
Streamlit Page	Esha	9/27/2024	10/7/2024
Model 1			
Data Sourcing and Cleaning	Vonesha	10/7/2024	10/15/2024
Model Selection	Khushi	10/15/2024	10/18/2024
Data Pre-Processing	Vonesha	10/18/2024	10/25/2024
Model Coding	Khushi	10/25/2024	11/8/2024
Results Evaluation and Analysis	Chyrstabel, Natasha, Ashmitha	11/8/2024	11/16/2024

Contribution Table

	Team Member	Contribution
0	Ashmitha Aravind	Worked on PCA plot, fixed bugs with dataset setup, worked on a
1	Khushi Gupta	Worked on regression model as well as metric evaluation and vi
2	Natasha Setidadi	Worked on writing methods section, fixed bugs with the linear v
3	Vonesha Shaik	Worked on pulling datasets, cleaning data, preprocessing data,
4	Chrystabel Sunata	Worked on PCA plot, fixed bugs with dataset setup, worked on a

Video

Project overview: