A Machine Learning Approach to the Analysis of Real Estate Prices in the New York Metropolitan Area

Jacky He, Zhiduo Xie, Shou-Kai Cheng, Yibei Li, Yuchuan Zhang
Cornell University
Department of Information Science and Cornell Tech
ph474@cornell.edu, zx324@cornell.edu, sc2745@cornell.edu, yl3692@cornell.edu, yz2947@cornell.edu

Abstract

This project addresses the critical challenge of accurately predicting real estate prices in the New York Metropolitan Area (NYC), a crucial task for market transparency and aiding decisions in the dynamic NYC market. Employing a dataset with detailed property features from Kaggle, we apply and compare three state-of-the-art machine learning models—Linear Regression, Random Forest, and Gradient Boosting—to identify the most effective approach. The system's architecture includes a user-friendly front-end for parameter input and a robust back-end for model processing. Models are evaluated using RMSE, MAE and R^2, with the best-performing model deployed for tailored price estimations. This comparative analysis not only provides insights into the most effective predictive models but also advances machine learning applications in real estate, offering methodologies that could influence future predictive modeling.

1. Introduction

Predicting real estate prices is crucial in dynamic markets like New York City (NYC). Accurate predictions enhance market transparency and aid decision-making. This project develops an application to predict housing prices using advanced machine learning techniques and a comprehensive dataset with detailed property features. We utilize three machine learning algorithms: Linear Regression, Random Forest Regression, and Gradient Boosting Regression, chosen for their effectiveness with large, complex NYC real estate datasets. Previous research often focused on single models without extensive comparisons. Studies combined traditional and non-traditional data for better accuracy. Our work builds on this by comparing multiple models to find the best predictor for NYC real estate prices. The pipeline involves data exploration, preprocessing, model training, and evaluation using RMSE, MAE, and R^2. The final model is deployed in a user-friendly application for customized property price estimations. The machine learning pipeline is shown in Figure 1. This application democratizes access to real estate data and pricing knowledge, promoting a transparent and fair market. Ethical considerations include data privacy and avoiding bias reinforcement in predictive models.

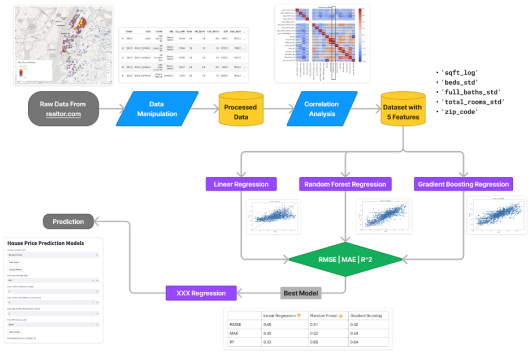


Figure 1: Machine Learning Pipeline

2. Background

Prior research in the realm of housing price prediction has predominantly utilized simpler statistical models such as linear regression to determine the impact of various property-related features on market values. While effective, these models often lack the complexity needed to fully capture the multifaceted nature of real estate markets. Recent advancements have expanded the scope of variables considered, incorporating not only physical attributes of properties but also environmental and socioeconomic factors. More sophisticated algorithms, including Ridge Regression, Support Vector Machines (SVM), and ensemble methods like Gradient Boosting, have been adopted to better account for these complexities.

For instance, Park and Bae's study [3] assesses the efficacy of several machine learning algorithms—including

C4.5, RIPPER, Naïve Bayesian, and AdaBoost—for predicting housing prices with a comprehensive dataset from Fairfax County, Virginia. Their findings underscore the heightened accuracy of RIPPER over the other models tested.

Another significant contribution is the hybrid approach that combines Lasso regression with Gradient Boosting models, enhancing predictive accuracy in real estate price modeling,as discussed by Lu al. (2017) [2].

In our work, we aim to build upon these foundations by employing a comparative analysis of advanced machine learning models, specifically Linear Regression, Random Forest, and Gradient Boosting. This approach is designed to leverage the distinct strengths of each model to more accurately predict housing prices in the highly dynamic and complex New York City real estate market.

Our anticipated approach contrasts with earlier methods by focusing on a direct comparison and evaluation of multiple models to identify which most effectively captures the nuances of NYC's market.

3. End-to-End ML Pipeline

3.1. Data Collection, Exploration & Processing

We have chosen a dataset comprising property sales data from New York City, which was collected from Realtor.com through API calls. This dataset is particularly tailored for developing predictive models for real estate prices and contains 10,000 records spread across 21 features. These features are categorized as follows:

1. **Categorical Data:** 'status', 'style', 'city', 'county'.

2. **Numerical Data:** 'zip_code', 'beds', 'full_baths', 'half_baths', 'sqft', 'year_built', 'days_on_mls', 'list_price', 'sold_price', 'assessed_value', 'estimated_value', 'lot_sqft', 'price_per_sqft', 'latitude', 'longitude', 'stories'.

The data, extracted real-time covering the last 365 days as of May 1, 2024, from Realtor.com, ensures that the most recent trends in the real estate market are captured. **Ground truth labels** such as 'sold_price' are provided, which facilitates supervised learning approaches.

3.1.1 Data Cleaning and Preprocessing Techniques

To prepare this dataset for machine learning algorithms, we implemented several preprocessing techniques:

- **Handling Missing Data:**
  - Missing values in 'half_baths' are filled with zeros.
  - Columns with significant missing data or lesser relevance like 'days_on_mls', 'assessed_value', 'estimated_value', 'lot_sqft', and 'stories' are dropped.
  - The median values are used to fill missing data in 'sqft', while the median or mode values are used for 'full_baths', 'beds', 'year_built', 'list_price', 'zip_code', and 'county'.
  - Missing latitude and longitude are imputed geographically based on median values from the same zip code.

- **Outlier Handling:**
  - Outliers in geographical data are removed or imputed, ensuring that 'latitude' and 'longitude' values are within specific bounds relevant to New York City.

- **Data Type Conversions:**
  - 'zip_code' is converted from float to integer after handling missing values.

3.1.2 Feature Engineering and Transformations

To enhance the dataset's predictive capabilities, these feature engineering and transformation steps were executed:

- **One-Hot Encoding:**
  - The 'style' column is transformed into multiple binary columns through one-hot encoding, facilitating its use in numerical modeling.

- **Log and Standardization Transformations:**
  - Logarithmic transformation is applied to 'sqft', 'list_price', and 'sold_price', creating new features: 'sqft_log', 'list_price_log', and 'sold_price_log'.
  - Standardization is applied to newly calculated 'house_age', 'beds', 'full_baths', and 'half_baths', producing standardized features like 'house_age_std', 'beds_std', 'full_baths_std', and 'half_baths_std'.

- **Normalization:**
  - The new 'style_' features created by one-hot encoding are normalized. New features are named like 'style_[feature]_norm', where [feature] is the specific style category.

- **New Feature Creation:**

- 'house_age' is calculated by subtracting 'year_built' from the current year (2024).
- 'total_rooms_std' combines the total standardized counts of bedrooms and bathrooms into a single feature, reflecting the scaled room count impact on house valuation.
- 'price_per_sqft_log' normalizes the price per square foot by dividing 'list_price_log' by 'sqft_log', aiding in the comparison of properties of varying sizes.

- **Outlier Removal:**
  - Outliers in 'sqft_log', 'list_price_log', 'sold_price_log', and 'price_per_sqft_log' are identified and removed using the Interquartile Range (IQR) method.

3.1.3 Data Exploration Displays

- **Histogram:**

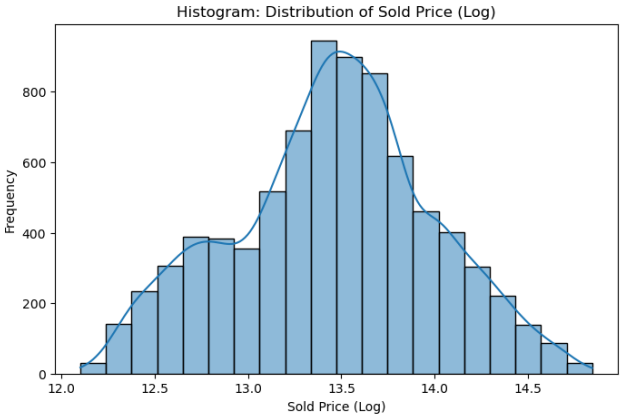


Figure 2: Distribution of Sold Price

- **Bar chart:**

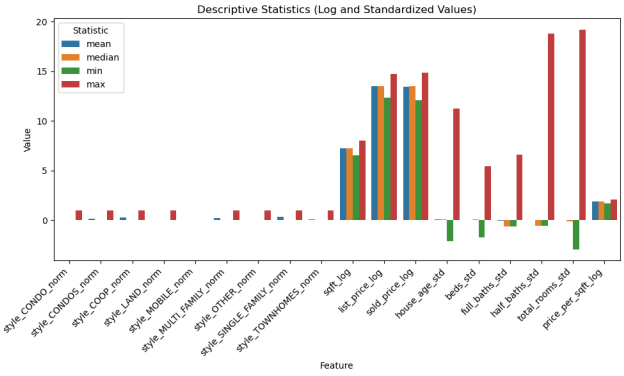


Figure 3: Descriptive Statistics

- **Heat map:**

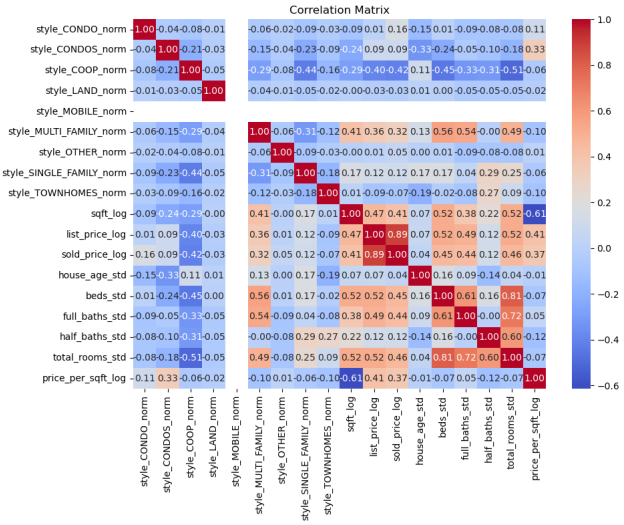


Figure 4: Correlation Analysis

3.2. Methods and Model Training

3.2.1 Machine Learning Techniques and Justification

In this study, we employed three distinct regression techniques to predict real estate prices in the New York Metropolitan area: Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Regression was chosen as the appropriate machine learning approach due to the continuous nature of the output variable (real estate prices), which is well-suited for predicting values based on input features. **Linear Regression** was used to establish a baseline for model performance. Its simplicity and efficiency in terms of computation make it a standard choice for initial modeling. **Random Forest Regression** was selected for its ability to handle non-linear data effectively and its robustness against overfitting, particularly important given the complexity and variability of real estate data. **Gradient Boosting Regression** was utilized for its strengths in reducing bias and variance through ensemble learning, where weak learners are combined sequentially to create a strong predictive model.

3.2.2 Model Inputs and Outputs

The models utilized in this study leverage a set of transformed and standardized input features derived from raw real estate data, each selected for their relevance and predictive power concerning property pricing. The specific features used are as follows:

- **sqft_log:** The logarithm of the square footage, transforming this feature to reduce skewness and improve

model performance.

- **beds_std:** The number of bedrooms, standardized.
- **full_baths_std:** The number of full bathrooms, standardized.
- **total_rooms_std:** The total number of rooms, standardized.
- **zip_code:** The postal code of the property, which is a categorical variable and is treated with appropriate encoding techniques to fit the model.

The output of the model is the **sold_price_log**, which is the logarithm of the sold price of a property. This transformation is applied to normalize the distribution of the prices and to enhance the predictive accuracy of the model. The output being a continuous variable confirms the use of regression techniques in this study. Predicting the logarithm of the sold price, rather than the actual price, helps in reducing the effect of extreme values on the model's performance, thus providing a more stable and reliable prediction.

3.2.3 Training and Validation Procedure

The dataset was split into an 80/20 ratio, with 80% used for training the models and 20% reserved for testing. This split ensures that the models have ample data for learning while still reserving a substantial portion for unbiased evaluation of model performance.

3.2.4 Model Parameter Settings and Justification

For the Linear Regression model building from scratch, we set the learning rate at 0.01 and the iterations at 1000 times. Random Forest Regression parameters such as the number of trees (100 trees), max depth (30), and min samples split (2) were chosen to balance model complexity and training time. Gradient Boosting Regression used 100 boosting stages (n_estimators), a learning rate of 0.1, and a max depth of 3, enabling gradual and thorough learning.

3.2.5 Methods to Avoid Overfitting and Underfitting

For the Linear Regression model, we used regularization techniques (Lasso Regression) to prevent overfitting. However, the performance did not significantly improve in this case. For the Random Forest and Gradient Boosting models, we utilized parameters like max depth and min samples per leaf to control the growth of trees and thus prevent overfitting.

3.3. Model Evaluation

3.3.1 Experiments and Error Analysis

The study conducted a comprehensive evaluation of three regression models: Linear Regression, Random Forest, and Gradient Boosting. These models were assessed using a subset of the data, specifically designated as the test set, which accounted for 20% of the total data. The error analysis focused on comparing the predicted real estate prices against actual prices in the dataset, thereby identifying the precision and robustness of each model.

The models were evaluated using the following metrics:

- **Root Mean Squared Error (RMSE)**: Measures the average magnitude of the errors, providing a clear indication of model accuracy.
- **Mean Absolute Error (MAE)**: Offers an average of the absolute errors, advantageous for its robustness to outliers.
- **R-squared (R^2)**: Represents the proportion of variance in the dependent variable predictable from the independent variables, useful for assessing the explanatory power of the model.

MAE, RMSE, and R^2 are widely recognized and utilized in the field of real estate price prediction and other regression-related tasks. These metrics are standard for evaluating the performance of predictive models and are noted for their ability to provide insight into model accuracy and variance explanation. Previous works such as Hyndman and Koehler (2006)[1] discuss these metrics extensively, emphasizing their relevance and utility in statistical model evaluation.

3.4. Results

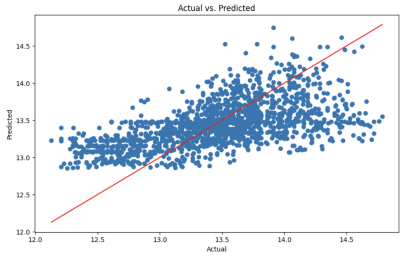
The ML results for the three regression models—Linear Regression, Random Forest, and Gradient Boosting—are quantitatively summarized using two main performance metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), along with R-squared (R^2) to evaluate the models' explanatory power. Table 1 shows the table of summarized metrics for each model.

	Linear Regression 👎	Random Forest 👍	Gradient Boosting
RMSE	0.45	0.31	0.32
MAE	0.35	0.22	0.24
R^2	0.33	0.68	0.64

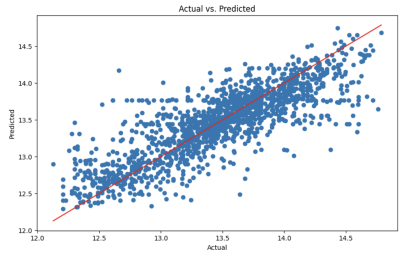
Table 1: Evaluation Metrics for Regression Models

Figure 5 shows the visual comparison through plots of predicted versus actual values further supports the quantitative findings. Among the models, Random Forest Regression demonstrates the best overall performance with the

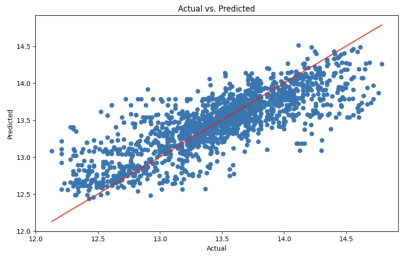
lowest RMSE and MAE, indicating the highest predictive accuracy and lowest errors in predictions. Additionally, it achieves the highest R^2 value, suggesting it is most effective at explaining the variability in real estate prices among the models tested.



(a) Linear Regression



(b) Random Forest



(c) Gradient Boosting

Figure 5: Comparison of model predictions: Actual vs. Predicted values

3.5. Model Deployment

The application of our project significantly impacts real estate transactions by providing a tool that can accurately predict property prices in the New York Metropolitan Area. This capability is vital for enhancing market transparency, aiding potential buyers and sellers in making more informed decisions, and possibly stabilizing property prices in a volatile market.

- **Price Estimation:** Uses advanced algorithms (Linear Regression, Random Forest, Gradient Boosting) to predict real estate prices.
- **Market Transparency:** Improves transparency and supports informed decisions in NYC's real estate market.
- **Market Stability:** Aids both buyers and sellers with accurate forecasts, stabilizing the market and widening access to real estate investments.

The use of this application raises important ethical and societal questions, primarily centered around the fairness and transparency of machine learning algorithms. Ensuring that the models do not perpetuate or exacerbate existing biases in real estate pricing is crucial. This concern is addressed by employing techniques such as cross-validation and regularization to avoid overfitting and to ensure that the models generalize well across different segments of the market. Furthermore, there is an emphasis on privacy and the ethical handling of user data, crucial for maintaining user trust and compliance with data protection laws.

- **Democratization of Access:** Improves market transparency and equity by making complicated real estate data available.
- **Data Privacy:** Secures user data to protect personal information.
- **Bias Mitigation:** Updates and refines models to prevent perpetuating market inequities.
- **Techniques for Fairness:** Employs cross-validation and regularization to enhance model reliability and compliance with data protection laws.

3.6. Front-End (Streamlit)

3.6.1 Home Page

Quick Start Guide: Instructions on how to use the application to get housing price estimates.
User Inputs: Features a file uploader where users can upload their dataset. Once a file is uploaded, a preview of the data is displayed.

3.6.2 Data Exploration Page

Interactive Charts: Users can see the preview of the housing data as shown in Figure 6
Visualization: Users can select different visualizations of the processed housing dataset, as shown in figure 7
Map Views: Users can see the map views of the housing price in NYC. Demo in Figure 8

3.6.3 Prediction Page

Input Form: Users input property details through dropdown menus.
Train Model Button: Triggers the prediction model.
Show Result Button: Shows the predicted housing price.

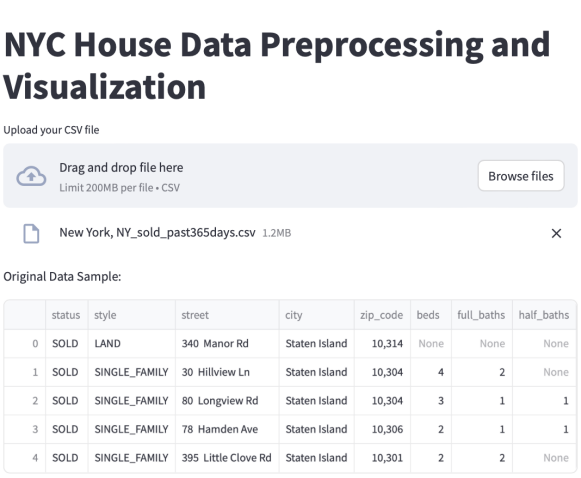


Figure 6: Dataset uploader and preview

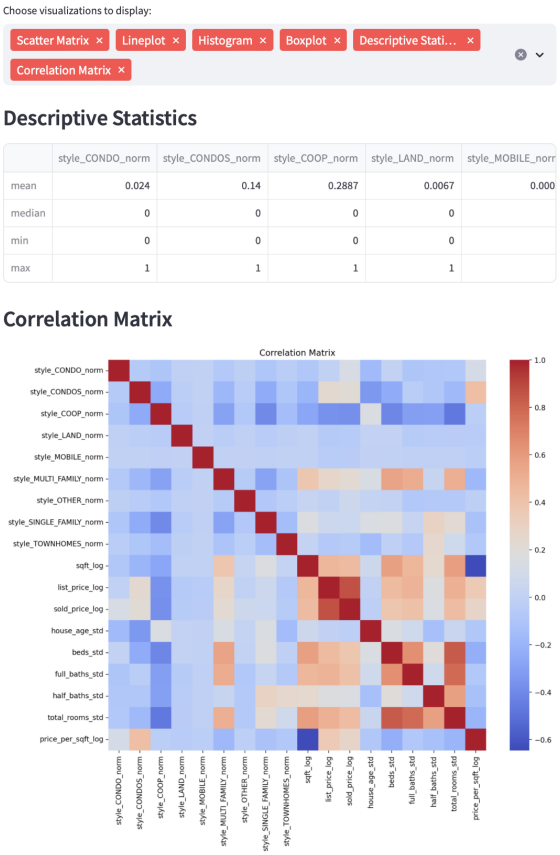


Figure 7: Users can select different visualizations of the processed housing dataset

Menus and Input Widgets include a Navigation Bar at the top of page,and various input widgets such as dropdowns. Demo in Figure 9

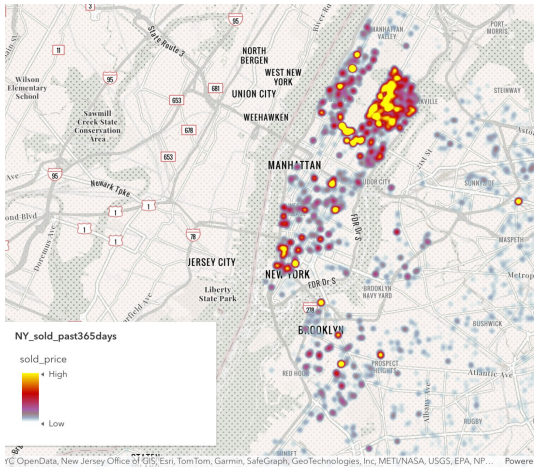


Figure 8: Map view of the housing price distribution across New York City.

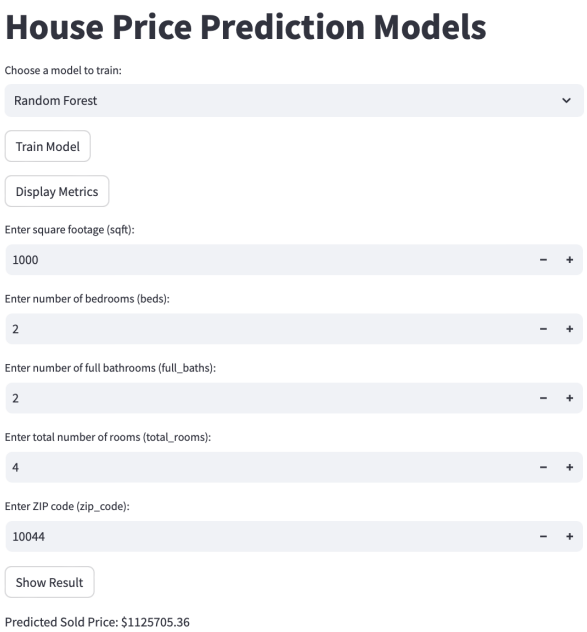


Figure 9: Housing price prediction

3.6.4 Integration of Front- and Back-End

Model training is triggered by user actions, such as clicking train model button and submitting the input form. The results will be shown based on user inputs and selections, providing an engaging user experience.

3.6.5 Design Considerations

The UI is responsive and accessible, with high contrast and easy navigation options to accommodate users with visual

impairments.

4. Conclusion

The project aimed to develop a robust application for accurately predicting real estate prices in the New York Metropolitan Area. Using advanced machine learning techniques, we compared three models—Linear Regression, Random Forest, and Gradient Boosting—to determine the most effective approach for price prediction. Our methodology involved a comprehensive data processing pipeline, rigorous model evaluation, and deployment in a user-friendly interface.

- **Goals:** The primary objective was to enhance market transparency and decision-making by providing accurate housing price predictions.
- **Approach:** We employed three state-of-the-art machine learning models and evaluated them using RMSE, MAE, and R^2 metrics. The models were trained and tested on a comprehensive dataset from Kaggle, which included detailed property features.
- **Results:** Among the models tested, the Random Forest model demonstrated the best performance with the lowest RMSE and MAE and the highest R^2 value, indicating superior predictive accuracy and explanatory power. By implementing and fine-tuning these models, we achieved well-performing models that provide reliable predictions.

By using machine learning, we anticipate more accurate housing price predictions. The models not only offer higher accuracy but also provide deeper insights into market dynamics, improving the understanding of price influences. This project not only provides immediate practical benefits but also lays the groundwork for future enhancements in predictive modeling and data analysis, ultimately contributing to a more efficient and transparent market environment.

- **Well-performed Model:** By implementing and fine-tuning the three models, we can achieve well-performing models.
- **Higher Accuracy:** Using advanced machine learning, we anticipate more accurate housing price predictions.
- **Market Insights:** The model will provide deeper insights into market dynamics, improving understanding of price influences.

5. Team Member Contribution

5.1. Technical Components

Front-end: Yuchuan Responsible for data visualization. This involves creating interactive graphics and charts to display the data, allowing users to understand trends and insights visually.
Back-end: **Yibei** Handles data processing and Arcgis. This includes cleaning, organizing, and preparing data for analysis, ensuring that the data input into the machine learning models is accurate and structured. **Zhiduo, Jacky and Shou-Kai** Collaboratively work on ML model training. They are tasked with developing, training, and refining machine learning algorithms that the project relies on for data analysis and predictions.

5.2. Writing Components

Introduction and Background: Jacky
Data Collection, Exploration & Processing: Yibei
Integration of Front- and Back-End: Yuchuan
Machine Learning Models: Zhiduo
Results and Conclusion: Shou-kai

References

[1] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. 4

[2] Sifei Lu, Zengxiang Li, Zheng Qin, XuLei Yang, and Rick Goh. A hybrid regression technique for house prices prediction. pages 319–323, 12 2017. 2

[3] Bobae Park and Jae Kook Bae. Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data. *Expert Systems with Applications*, 42(6):2928–2934, 2015. 1