PERSONALIZED REAL ESTATE VALUATION SYSTEM WITH PRICE PREDICTION

Team members

Jackson Jose &
Jijo John
Date: 01-Mar-2025 – 10-May-2025

Summary

This project focuses on developing a robust machine learning system for accurately predicting real estate property values. Utilizing the Zillow Housing Dataset, the system employs an ensemble regression approach to capture complex relationships between property features, location, market trends, and seasonal effects. The project encompasses comprehensive data cleaning and preprocessing, exploratory data analysis (EDA) to uncover insights, training of multiple regression models, and the development of a Stacking Regressor for final predictions. Key deliverables include the ensemble model, feature importance analysis, and the groundwork for a future interactive Flask web application for property assessment. Evaluation metrics such as RMSE, MAE, and R^2 are used to demonstrate the model's predictive power and accuracy.

1 Introduction

Accurate real estate valuation is crucial for buyers, sellers, investors, and financial institutions, but traditional methods can be subjective and slow. Primary goal to develop a data-driven machine learning system for accurately predicting property sale prices.

Significance of the project is to demonstrates the efficacy of machine learning in providing more objective, efficient, and potentially more accurate property valuations compared to traditional approaches.

Core Methodology: Employs ensemble regression techniques, culminating in a Stacking Regressor, to capture complex market dynamics.

Key Project Phases: Includes meticulous data cleaning, insightful exploratory data analysis (EDA), training diverse models, and rigorous evaluation.

Target Outcome: Aims to deliver a robust predictive model, analyze key value drivers (feature importance), and lay groundwork for a practical valuation tool (e.g., Flask app).

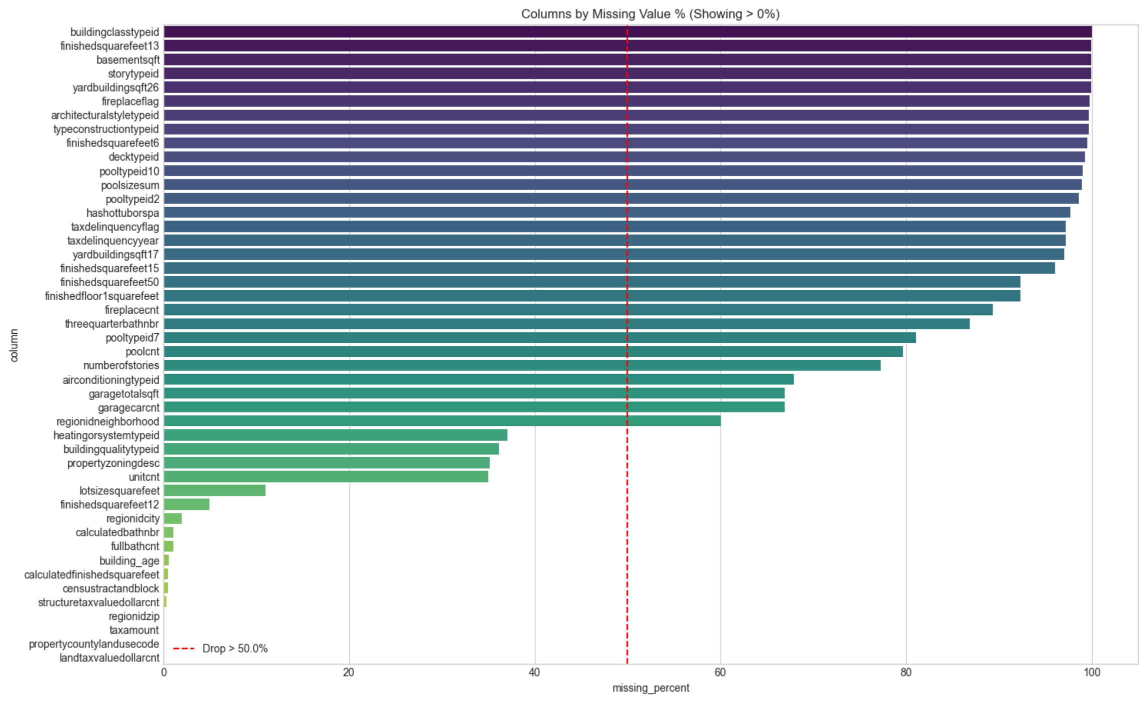
2. Dataset Overview

Dataset Used: Leverages the comprehensive Zillow Housing Dataset, rich in property features and transaction data.

3. Data Cleaning and Preprocessing

A rigorous data cleaning and preprocessing pipeline was implemented to prepare the data for modeling:

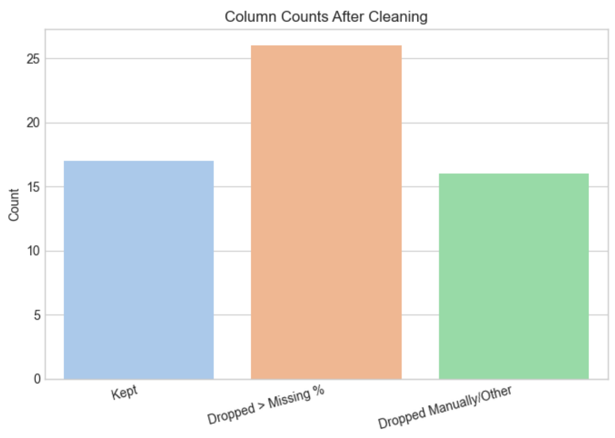
- Initial Data Loading & Merging:
- Dictionary Mapping & Target Variable Engineering:
- Handling Missing Values:
    1. An initial analysis of missing values across all features was conducted. The distribution of missing percentages is visualized below, highlighting features with substantial missing data.



- Columns with a missing value percentage greater than 50% were dropped to reduce noise and dimensionality.

Column Dropping (Manual/Redundancy/Leakage):

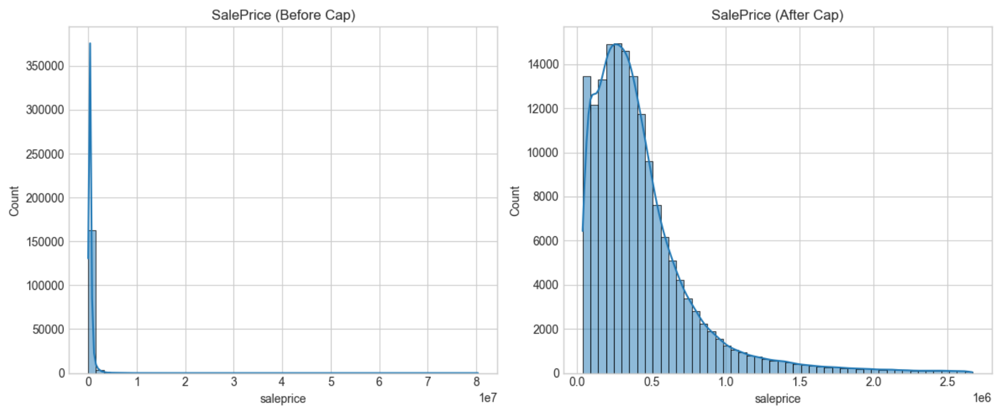
- Identifier columns (`parcelid`), redundant location identifiers (while keeping `regionidzip` and `fips`), raw census data, and tax-related columns that could lead to data leakage were manually removed.
- The following chart summarizes the column reduction process:



(Caption: Figure 2: Summary of columns kept versus those dropped due to high missing percentages or manual selection.)

Outlier Handling:

- The distribution of the target variable, sale price, was examined before and after outlier capping (e.g., at 1st and 99th percentiles) to ensure a more robust modeling process.



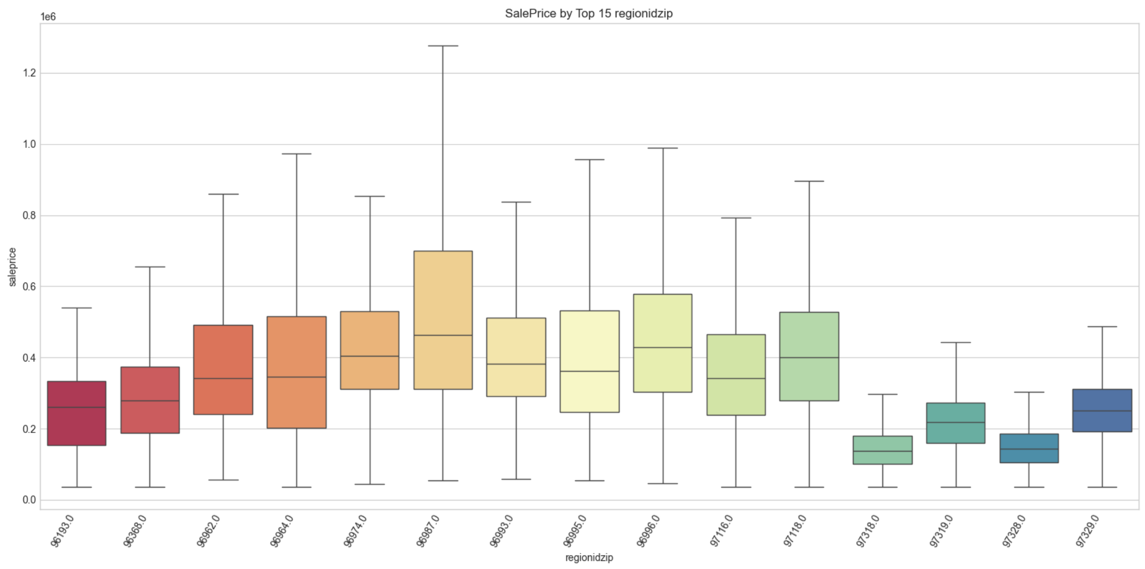
- Similar outlier treatment was applied to key numerical features like calculatedfinishedsquarefeet
- The final cleaned dataset was then used for EDA and model building.

4. Exploratory Data Analysis (EDA)

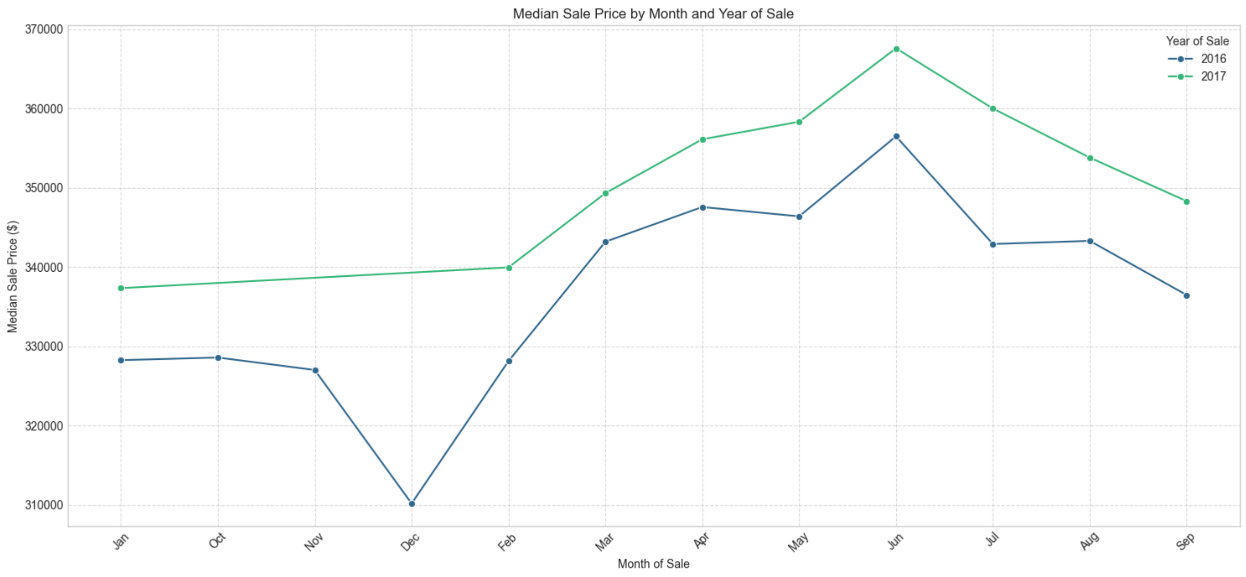
EDA was performed on the cleaned dataset to uncover patterns...

Sale Price Distribution Across Neighborhoods:

- Boxplots show the Sale Price distribution across the top 15 (or 20, adjust based on your chart) most frequent ZIP codes. This reveals significant price variations by locality.

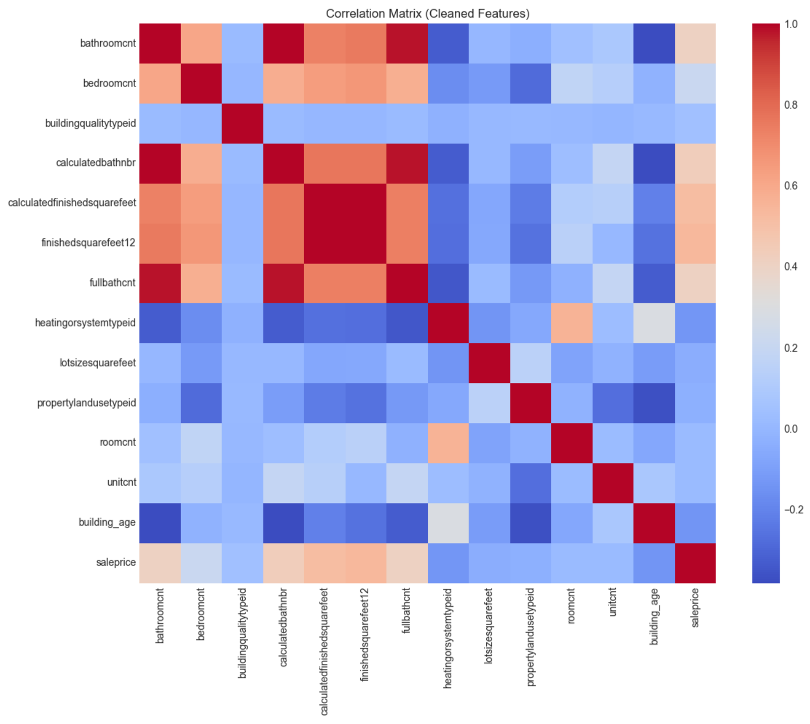


- **Seasonal Trends in Property Prices:**
    - *As described previously, assuming seasonal_price_median_lineplot.png and seasonal_price_multiyear_lineplot.png were generated if data allowed. If not, this section might state that seasonal features were explored but showed minimal clear patterns or data was insufficient.*

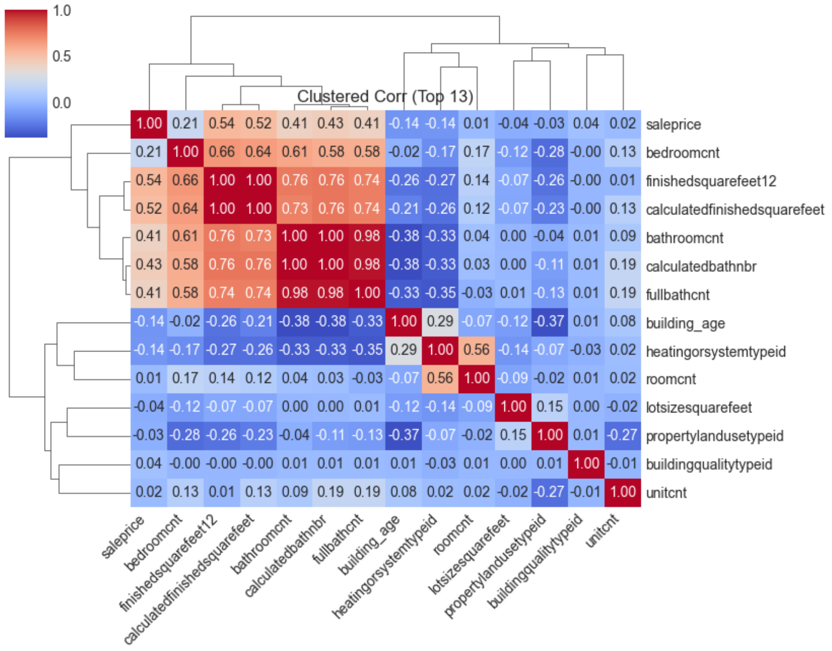


Correlation Analysis:

- A correlation heatmap was generated for key numerical features and sale price from the cleaned dataset to identify linear relationships.



- A clustered heatmap was also created for the top 13 (or as per your chart) most correlated features with saleprice (or a relevant subset), reordering features based on similarity to highlight groups of correlated variables.



	saleprice	bedroomcnt	finishedsquarefeet12	calculatedfinishedsquarefeet	bathroomcnt	calculatedbathnbr	fullbathcnt	building_age	heatingorsystemtypeid	roomcnt	lotsizesquarefeet	propertylandusetypeid	buildingqualitytypeid	unitcnt
saleprice	1.00	0.21	0.54	0.52	0.41	0.43	0.41	-0.14	-0.14	0.01	-0.04	-0.03	0.04	0.02
bedroomcnt	0.21	1.00	0.66	0.64	0.61	0.58	0.58	-0.02	-0.17	0.17	-0.12	-0.28	-0.00	0.13
finishedsquarefeet12	0.54	0.66	1.00	1.00	0.76	0.76	0.74	-0.26	-0.27	0.14	-0.07	-0.26	-0.00	0.01
calculatedfinishedsquarefeet	0.52	0.64	1.00	1.00	0.73	0.76	0.74	-0.21	-0.26	0.12	-0.07	-0.23	-0.00	0.13
bathroomcnt	0.41	0.61	0.76	0.73	1.00	1.00	0.98	-0.38	-0.33	0.04	0.00	-0.04	0.01	0.09
calculatedbathnbr	0.43	0.58	0.76	0.76	1.00	1.00	0.98	-0.38	-0.33	0.03	0.00	-0.11	0.01	0.19
fullbathcnt	0.41	0.58	0.74	0.74	0.98	0.98	1.00	-0.33	-0.35	-0.03	0.01	-0.13	0.01	0.19
building_age	-0.14	-0.02	-0.26	-0.21	-0.38	-0.38	-0.33	1.00	0.29	-0.07	-0.12	-0.37	0.01	0.08
heatingorsystemtypeid	-0.14	-0.17	-0.27	-0.26	-0.33	-0.33	-0.35	0.29	1.00	0.56	-0.14	-0.07	-0.03	0.02
roomcnt	0.01	0.17	0.14	0.12	0.04	0.03	-0.03	-0.07	0.56	1.00	-0.09	-0.02	0.01	0.02
lotsizesquarefeet	-0.04	-0.12	-0.07	-0.07	0.00	0.00	0.01	-0.12	-0.14	-0.09	1.00	0.15	0.00	-0.02
propertylandusetypeid	-0.03	-0.28	-0.26	-0.23	-0.04	-0.11	-0.13	-0.37	-0.07	-0.02	0.15	1.00	0.01	-0.27
buildingqualitytypeid	0.04	-0.00	-0.00	-0.00	0.01	0.01	0.01	0.01	-0.03	0.01	0.00	0.01	1.00	-0.01
unitcnt	0.02	0.13	0.01	0.13	0.09	0.19	0.19	0.08	0.02	0.02	-0.02	-0.27	-0.01	1.00

5. Modeling Approach

- Split data into train/test sets (80/20 split)
- Trained multiple models: Ridge, Lasso, Random Forest, Gradient Boosting, LightGBM
- Combined models using Stacking Regressor for improved performance
- Evaluated using RMSE, MAE, and R^2 Score

6. Model Evaluation Summary

- **Performance Metrics:** Models were evaluated using RMSE, MAE, and R^2 Score. The table below summarizes the performance of the key models tested on the held-out test set.

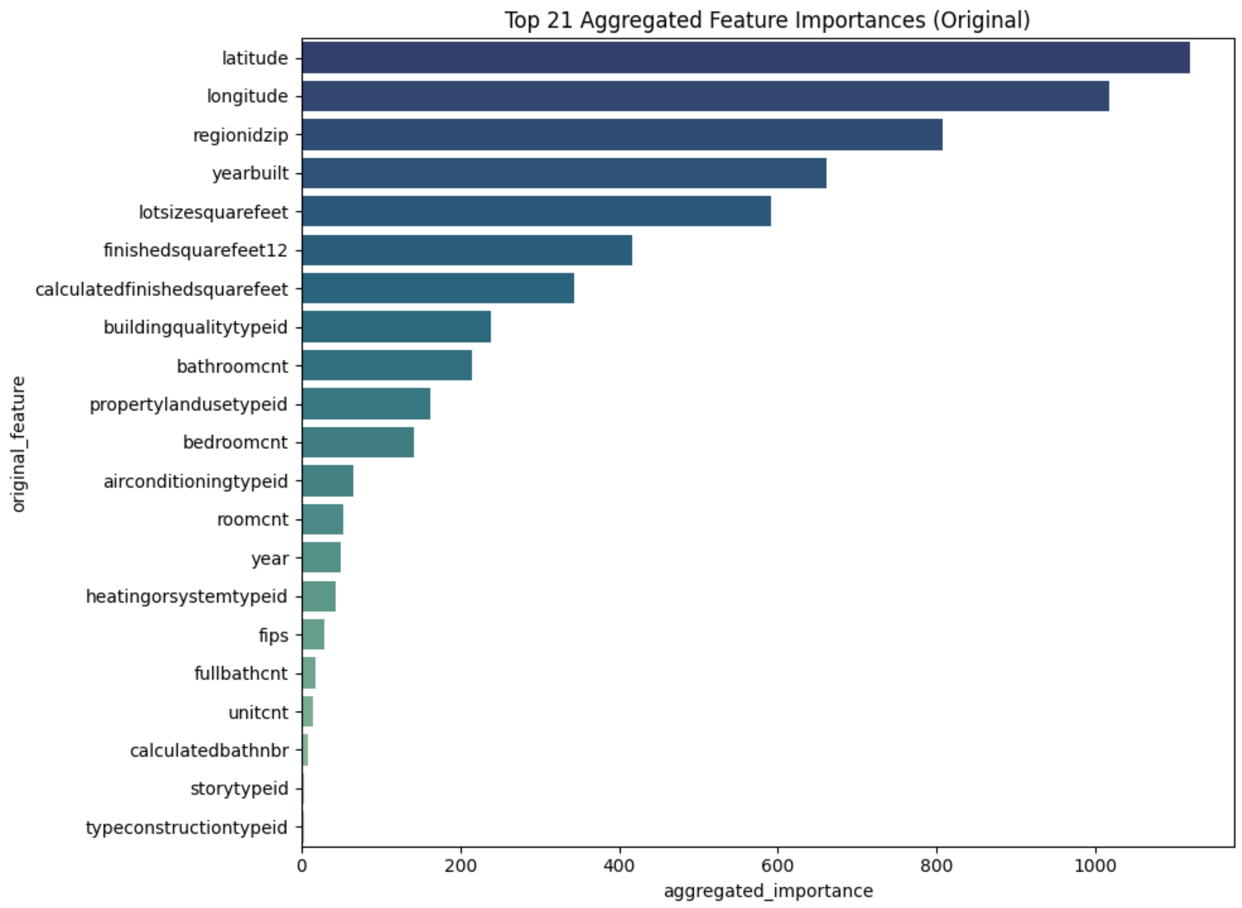
- **Table 1: Model Performance Comparison**

Model	RMSE	MAE	R2	Training Time (s)
Ridge	276846.6443	188488.9682	0.358018	0.10938
Lasso	276842.0568	188486.7022	0.358039	71.763522
Linear Regression	276839.628	188484.7048	0.358051	0.652342
Random Forest	238581.7946	157308.2173	0.523219	38.579632
LightGBM	238404.455	160702.7364	0.523928	1.514389
Stacked Ensemble	236202.6629	157769.7973	0.532681	180.655839

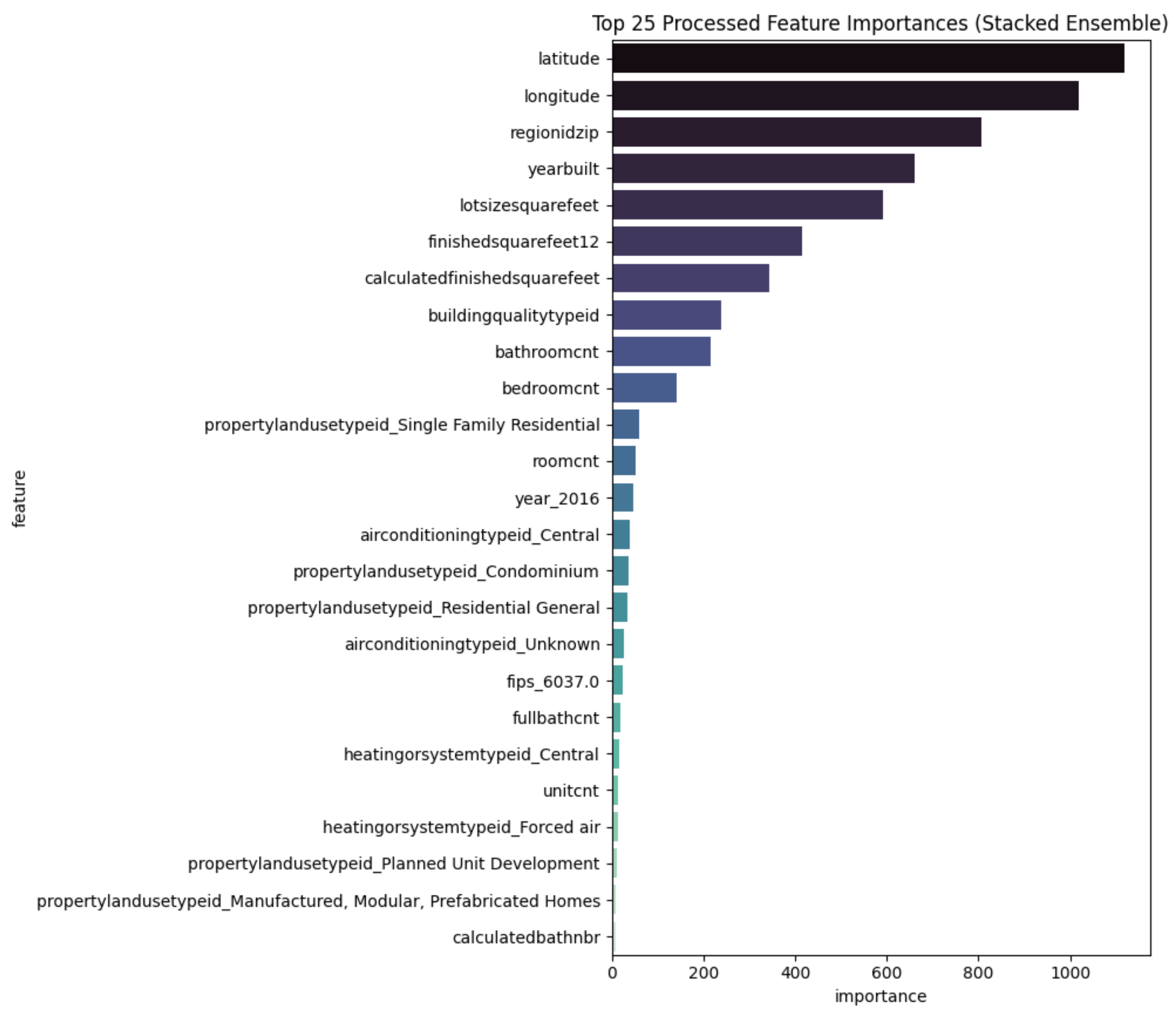
Best Performing Model: Stacked Ensemble

Feature Importance:

- Feature importances were extracted to understand the key drivers of property value. The "Top 16 Aggregated Importances" chart shows the influence of the original input features (after aggregating importances from OHE components where applicable). Location identifiers (regionidzip, fips) and property size (finishedsquarefeet12, calculatedfinishedsquarefeet) often rank highly.



- The "Top 25 Processed Importances" chart details the importances of features as seen by the model after preprocessing (including individual One-Hot Encoded categories). This often highlights specific ZIP codes or FIPS areas if the model was, for example, Ridge, where OHE'd features get individual coefficients.



7. Artifact Saving

To ensure reproducibility, enable deployment, and facilitate future analysis, key components of the developed system were serialized and saved. This included:

- **Final Model Pipeline (final_model_pipeline.joblib):** The complete, fitted pipeline containing the optimal preprocessor and the best-performing trained ensemble model (e.g., Stacked Regressor). This is the primary artifact for making new predictions.
- **Fitted Preprocessor (preprocessor.joblib):** The preprocessor (ColumnTransformer) instance, fitted on the training data, saved separately for potential independent use or inspection.
- **Feature Definitions & UI Data:**

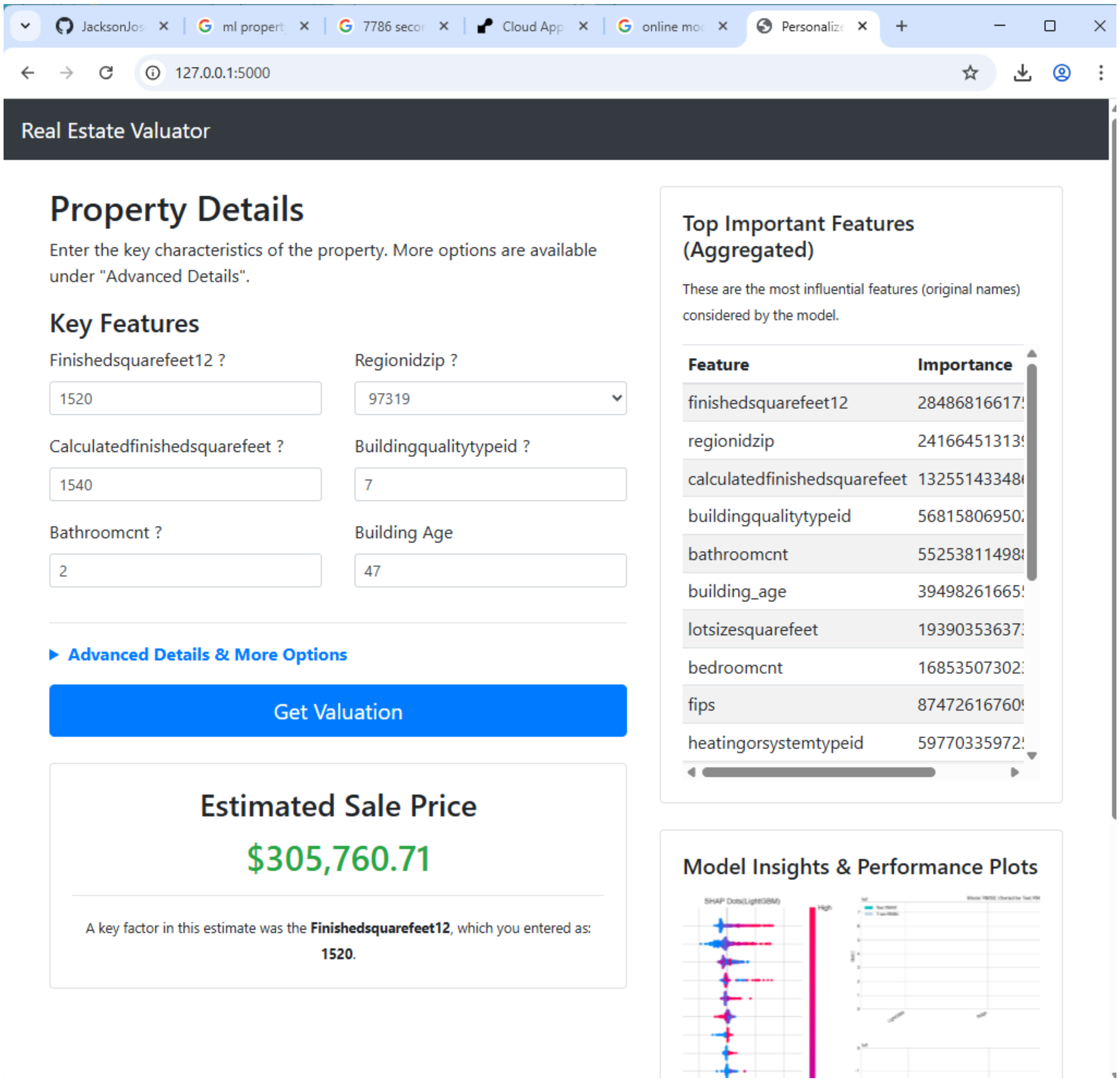
    - expected_columns_before_preprocessing.joblib: List of raw input feature names the model expects.
    - final_feature_names_post_ohe.joblib: List of feature names after One-Hot Encoding, used for interpreting processed feature importances.
    - feature_importance_table_data.joblib: Data (top N features and scores) for the UI's detailed feature importance table.
    - field_order_preference.joblib: List of original feature names, sorted by aggregated importance, to guide form field order in the UI.
    - Supporting dictionaries (field_descriptions.joblib, id_value_mappings.joblib) and original feature type lists from data processing were also saved for the Flask application.
- **Visualizations:** All relevant plots generated during EDA and feature importance analysis were saved as image files (e.g., in static/plots/) for inclusion in reports and the web application.

This systematic saving of artifacts ensures that the model can be easily loaded and utilized in other environments, such as the planned Flask web application.

8. Conclusion & Future Work

Conclusion: This project successfully developed an ensemble regression model for predicting real estate prices. The model captures influences from location, size, and other property attributes. The EDA provided valuable insights into data characteristics and price drivers.

Flask Web Application: The generated artifacts pave the way for deploying an interactive Flask application.


127.0.0.1:5000

Real Estate Valuator

Property Details

Enter the key characteristics of the property. More options are available under "Advanced Details".

Key Features

Finishedsquarefeet12 ?	Regionidzip ?
1520	97319
Calculatedfinishedsquarefeet ?	Buildingqualitytypeid ?
1540	7
Bathroomcnt ?	Building Age
2	47

▸ Advanced Details & More Options

Get Valuation

Estimated Sale Price

$305,760.71

A key factor in this estimate was the **Finishedsquarefeet12**, which you entered as: **1520**.

Top Important Features (Aggregated)

These are the most influential features (original names) considered by the model.

Feature	Importance
finishedsquarefeet12	2848681661
regionidzip	2416645131
calculatedfinishedsquarefeet	1325514334
buildingqualitytypeid	5681580695
bathroomcnt	5525381149
building_age	3949826166
lotsizesquarefeet	1939035363
bedroomcnt	1685350730
fips	8747261676
heatingorsystemtypeid	5977033597

Model Insights & Performance Plots

Docker: Deploy Docker Image and Configure Service

Git : https://github.com/JacksonJosekwt/SUBMISSION2

Future Work:

Include more granular location data, advanced feature engineering, hyperparameter tuning, alternative encoding, time series analysis, SHAP, regular retraining, confidence intervals.