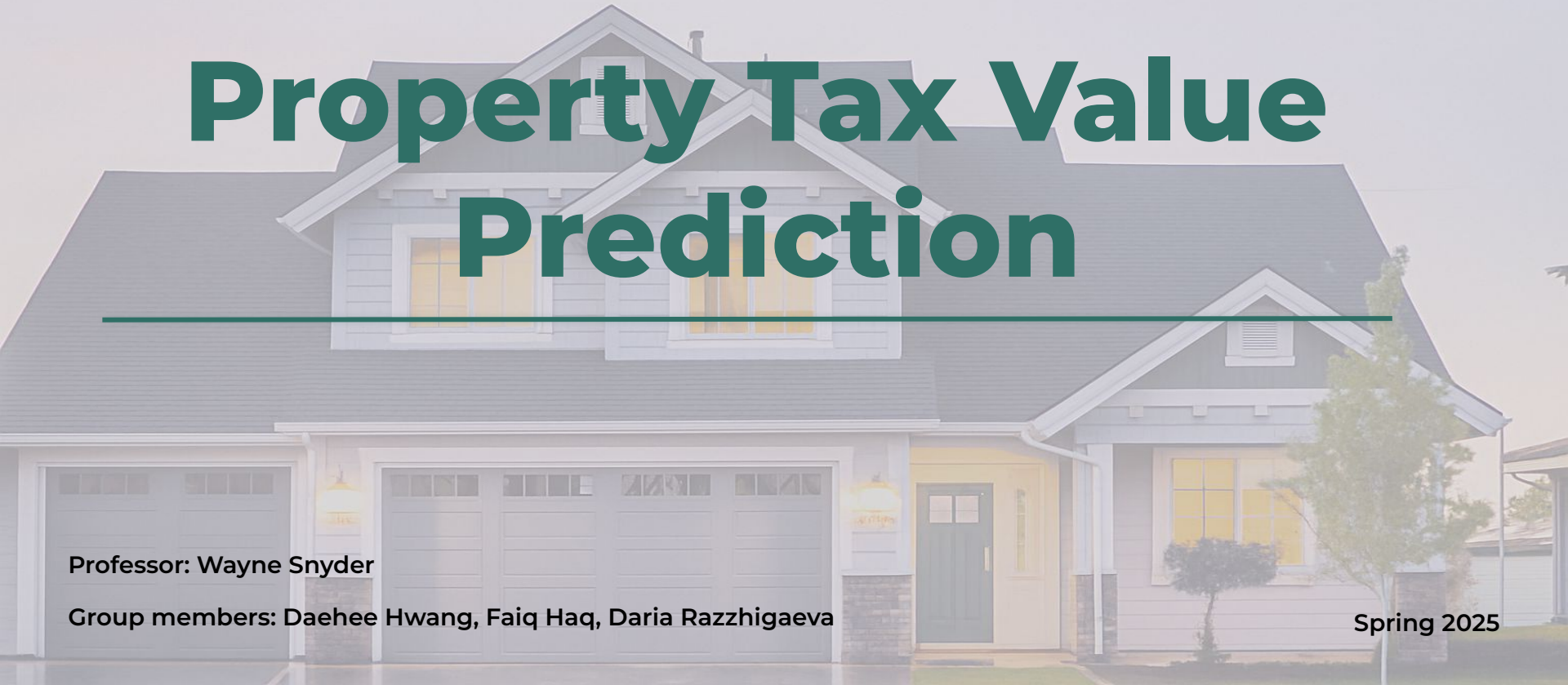
Property Tax Value Prediction



Professor: Wayne Snyder

Group members: Daehee Hwang, Faiq Haq, Daria Razzhigaeva

Spring 2025

Agenda

- Executive Summary
- Introduction
- Data Preparation
- EDA: Top Features
- EDA: Correlation with Target
- EDA: Multicollinearity Insights
- Model Development Process
- Pipeline
- Parameter Sweep Results
- Distribution of Key Features
- Trend - Visualized
- Relationship Between Property Size and Tax Value
- Measuring Success
- Future Directions
- Business or Practical Implications
- Team Contributions

Executive Summary

Developed predictive models to estimate property tax values using a structured dataset of ~77,000 residential properties from Zillow. Applied end-to-end machine learning workflow including data preprocessing, feature engineering, model selection, and tuning.

Addressed substantial missing data across many features (some with over 98% missingness) by applying context-specific imputation strategies, removing uninformative features, and carefully preserving valuable records to maintain dataset integrity.

Implemented multiple regression models, evaluating both linear (Linear, Ridge, Lasso) and ensemble methods (Decision Tree, Random Forest, Gradient Boosting), with repeated cross-validation using RMSE as the main metric to ensure consistent, interpretable evaluation.

Selected the best-performing model through randomized hyperparameter tuning and feature selection (including backward elimination and tree-based importance). Ensemble models, especially Gradient Boosting, showed the strongest performance and generalization.

Introduction

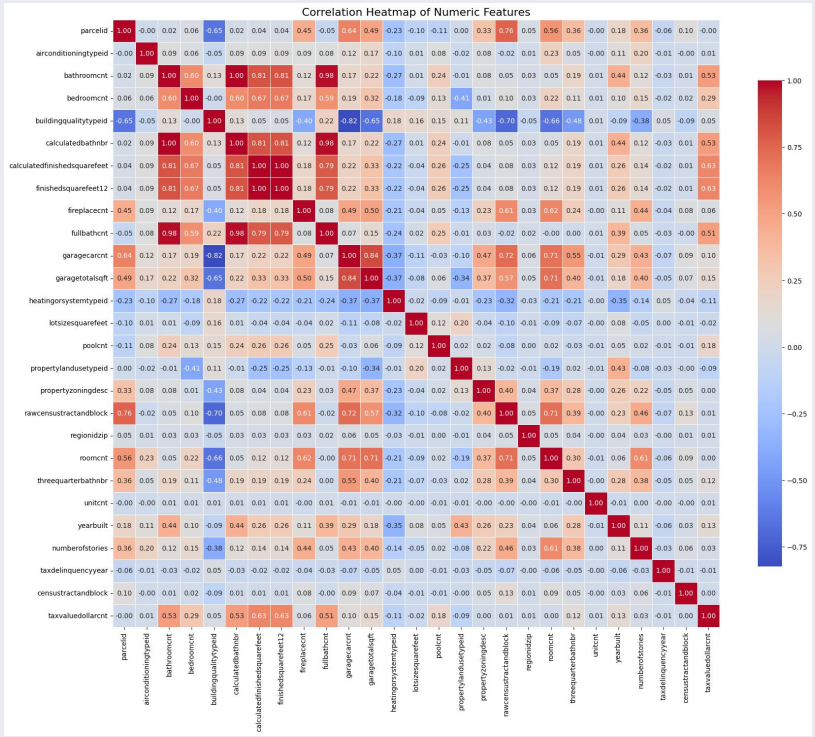
In the context of a real estate analytics initiative, our data science team was tasked with developing a model to estimate property tax values for residential properties using structured data. The primary goal was to build a predictive framework capable of producing accurate and interpretable tax value estimates for a wide range of properties, supporting business decisions related to valuation, investment, and forecasting.

The dataset provided was a large extract from Zillow's internal property database, containing both numerical and categorical features that describe physical characteristics, location details, and usage types. Our objective was to apply the complete machine learning workflow from data exploration and preprocessing to model development and evaluation while identifying key features that most influence assessed property values.

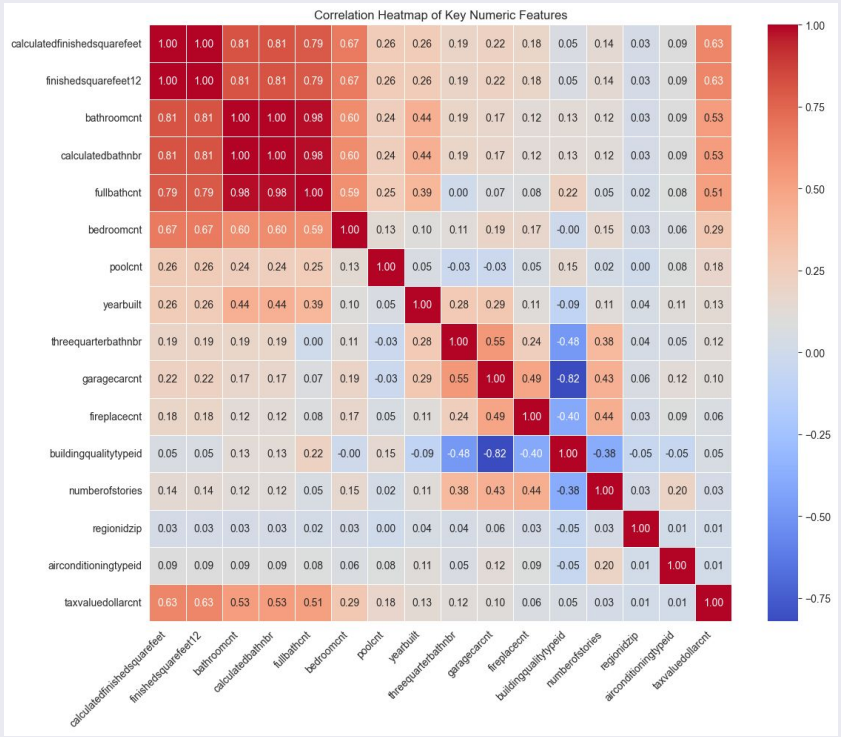
Data Preparation

- Feature Removal: Redundant Geographical Identifiers, High missing/null value features, normalization through removing outliers that misrepresent the data.
- Feature Engineering: Hot-encode categorical features, fill-missing values with appropriate correlated data (garage was reported but garage size is not reported), Default value imputation (unit count and stories defaults to 1 if not reported)
- Hot-encoding, normalization, logarithmic transformation, ratio-transformation.

Data Preparation



Before Feature engineering



After Feature engineering

EDA: Top Features

Top correlations with Tax Valuation:

Finished Square Footage	0.625351
Bathroom Count	0.525055
Full-bath Count	0.512469
Bedroom Count	0.290866
Pool Count	0.178949
Garage size	0.151741
Year of construction	0.134256

EDA: Correlation with Target

- Property Size features such as calculatedfinishedsquarefeet, finishedsquarefeet12 show **strong positive** correlations with the target taxvaluedollarcnt (0.63).
  - Larger homes tend to have higher property tax values.
- Bathroom and bedroom counts features like bathroomcnt, fullbathcnt, and calculatedbathnbr are highly correlated with each other and **moderately** with the target (0.51-0.53).
  - More bathrooms typically suggest higher property valuation.
- Bedroom count (bedroomcnt) is correlated weaker (0.29) than bathrooms.
- Garage-related features (garagecnt, garagetotalsqft) have **low** impact individually (<0.15)
- No strong correlation was observed with other features such as location or pool presence.

EDA: Multicollinearity Insights

- Strong multicollinearity detected between:
  - calculatedbathnbr, bathroomcnt and fullbathcnt (0.98-1.0)
  - calculatedfinishedsquarefeet and finishedsquarefeet12 (1.00)
- These features carry redundant information, risking instability in linear models
- To mitigate:
  - Keep only one feature from highly correlated pairs when necessary.
  - Tree-based models (like Random Forest, Gradient Boosting) should be prioritized, as they handle multicollinearity robustly without performance loss.

Model Development Process

- While we may now know how to identify non-linearity by observing the preliminary bivariate exploratory data analysis - For the purpose of this project we used trial and error. We ran the data through both linear and non-linear regressors.LinearRegression. Ridge, lasso, DecisionTreeRegressor, BaggingRegressor, RandomForestRegressor, GradientBoostingRegressor. Which were also cross-validated using RepeatedKFold tactic.
- We then used the mean Root Square Error (RMSE) as a marker to determine our best regressor for the dataset.

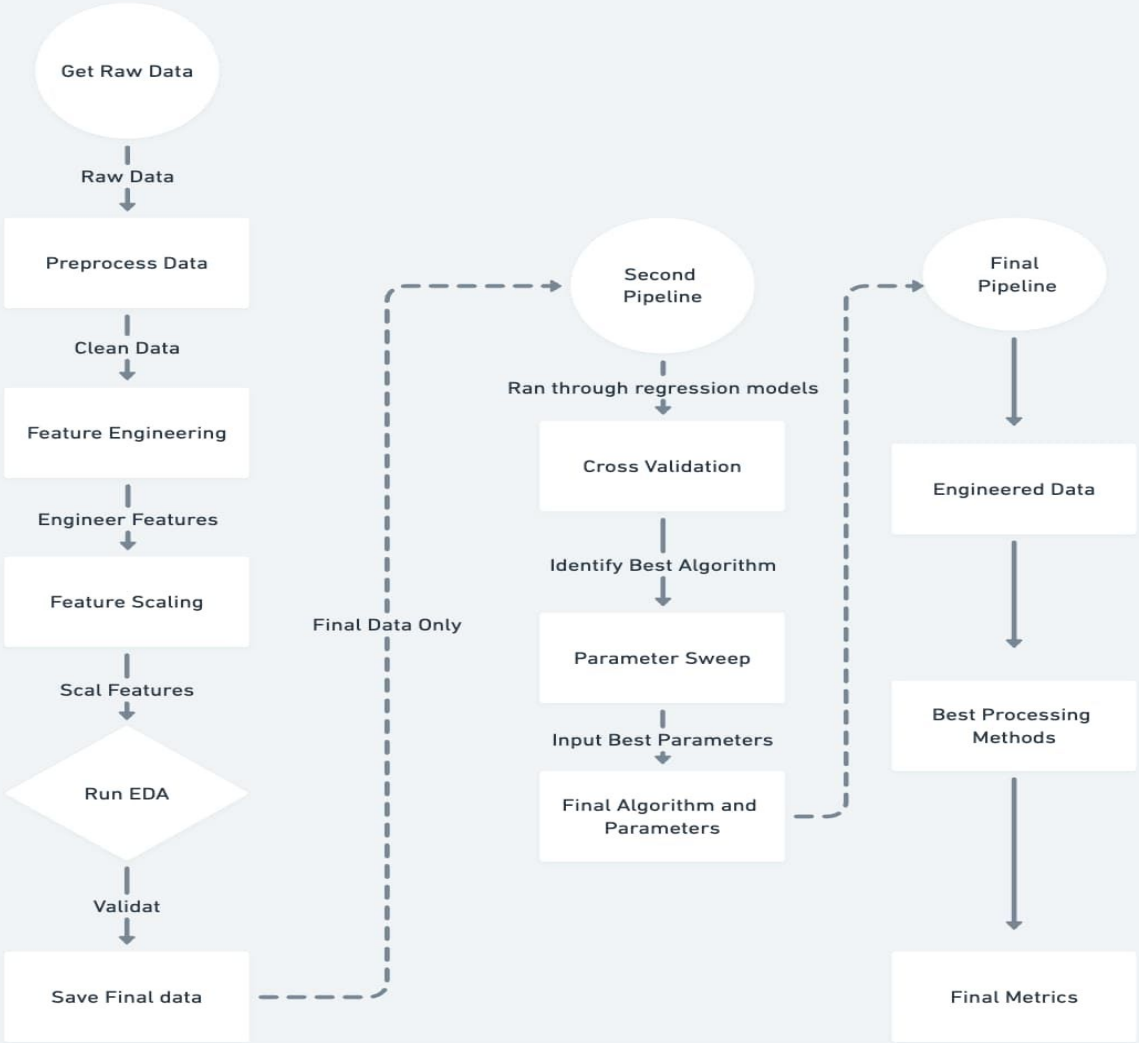
Model Development Process (Cont.)

- We chose RMSE as the guiding metric as we know that the dataset has complicated dimensionality and many of the features do not follow a normal distribution. Which R^2 score performs best on.
- We have also removed outliers that do not represent the overall data well during our preprocessing we are capturing many more opportunities with RMSE.

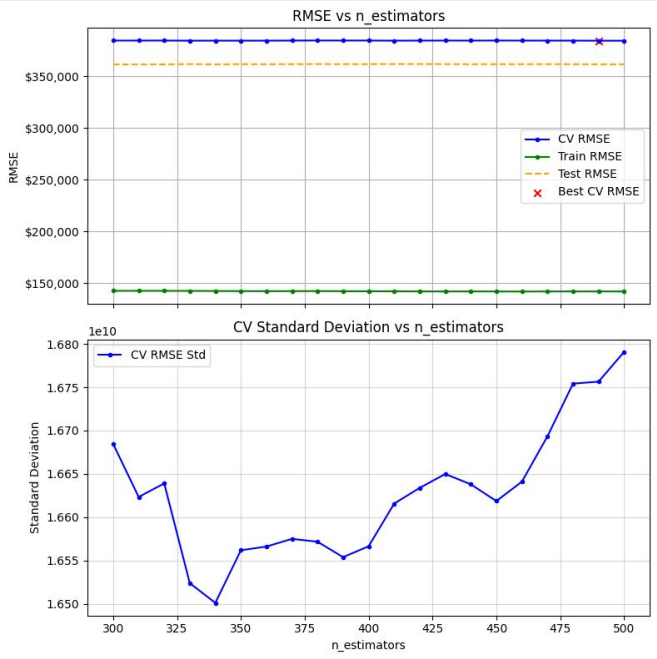
Model Development Process (Cont.)

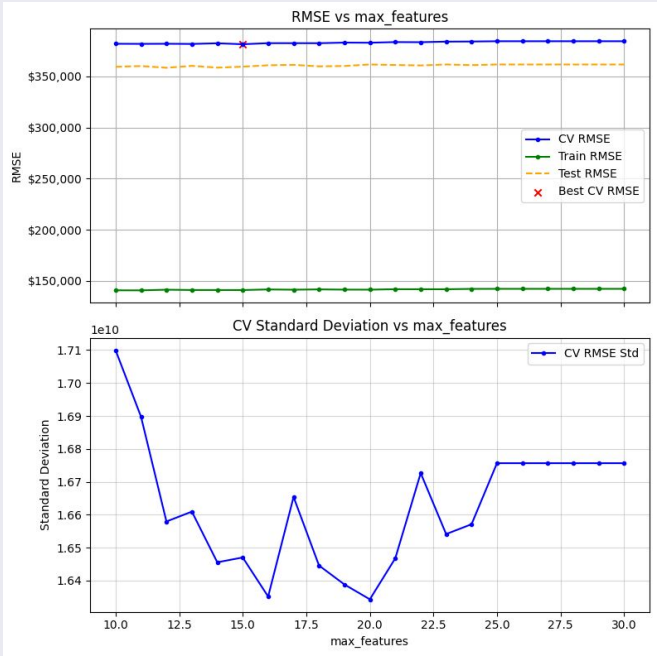
- The data was then ran through backward and foward feature selection process to verify each features importance and weight on the model. Which helps with avoiding over-generalization or over-engineering.
- Finally, we ran our dataset through the processes selected as best fitted for our dataset in isolation to confirm the findings.

Pipeline


Get Raw Data
Raw Data
Preprocess Data
Clean Data
Feature Engineering
Engineer Features
Feature Scaling
Scal Features
Run EDA
Validat
Save Final data
Final Data Only
Second Pipeline
Ran through regression models
Cross Validation
Identify Best Algorithm
Parameter Sweep
Input Best Parameters
Final Algorithm and Parameters
Final Pipeline
Engineered Data
Best Processing Methods
Final Metrics

Parameter Sweep Results

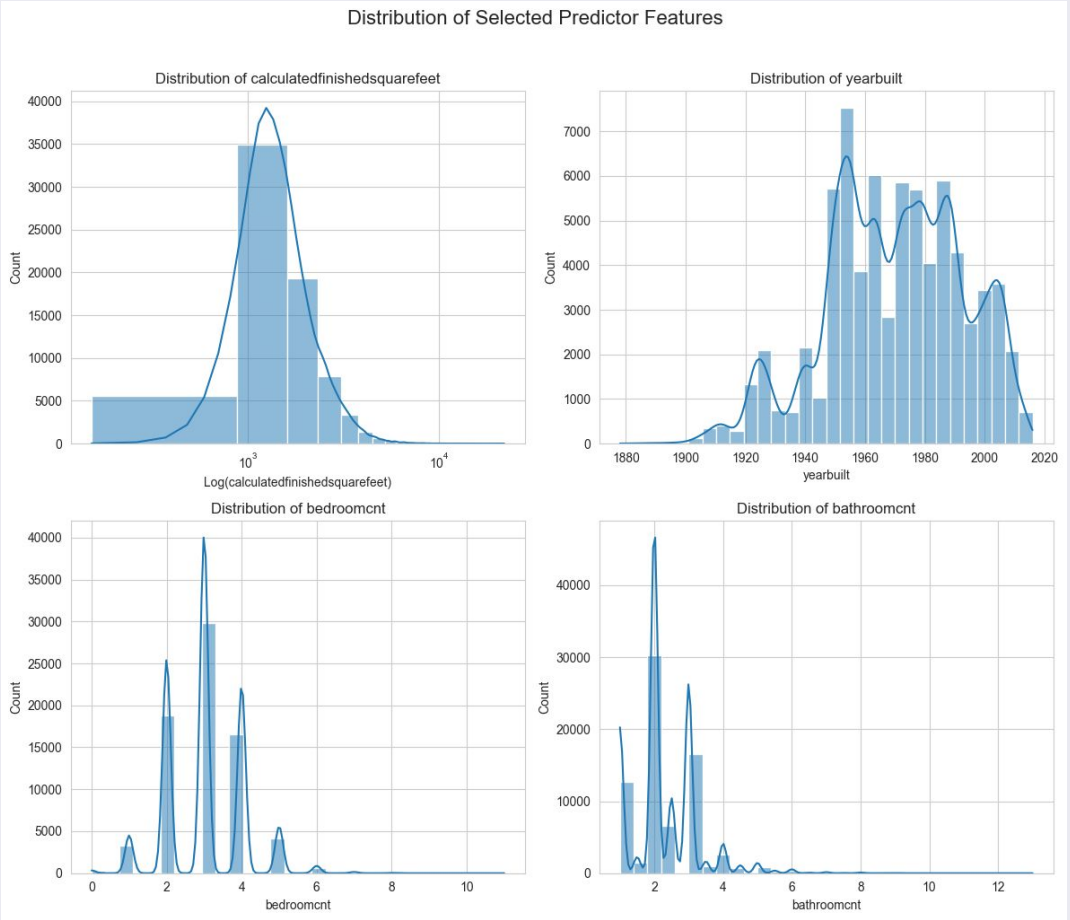




Parameter n_estimators = 490 RMSE = $384,444.12
{'n_estimators': 490, 'max_features': None, 'max_depth': None, 'bootstrap': True, 'random_state': 42, 'MSE_found': 147797280919.67545}

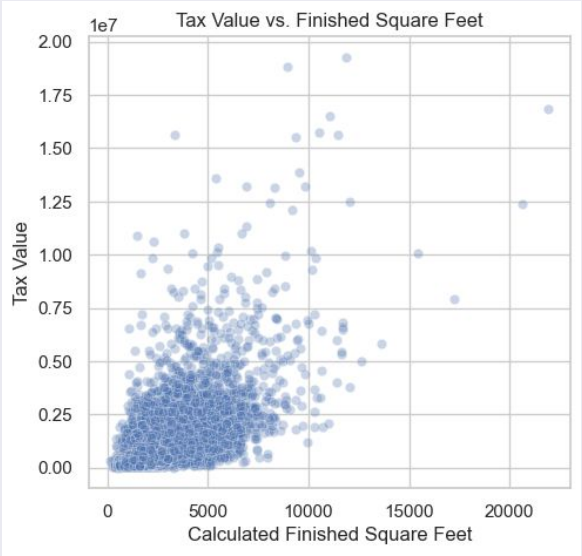
Parameter max_features = 15 RMSE = $381,511.42
{'n_estimators': 490, 'max_features': 15, 'max_depth': None, 'bootstrap': True, 'random_state': 42, 'MSE_found': 145550963251.203}

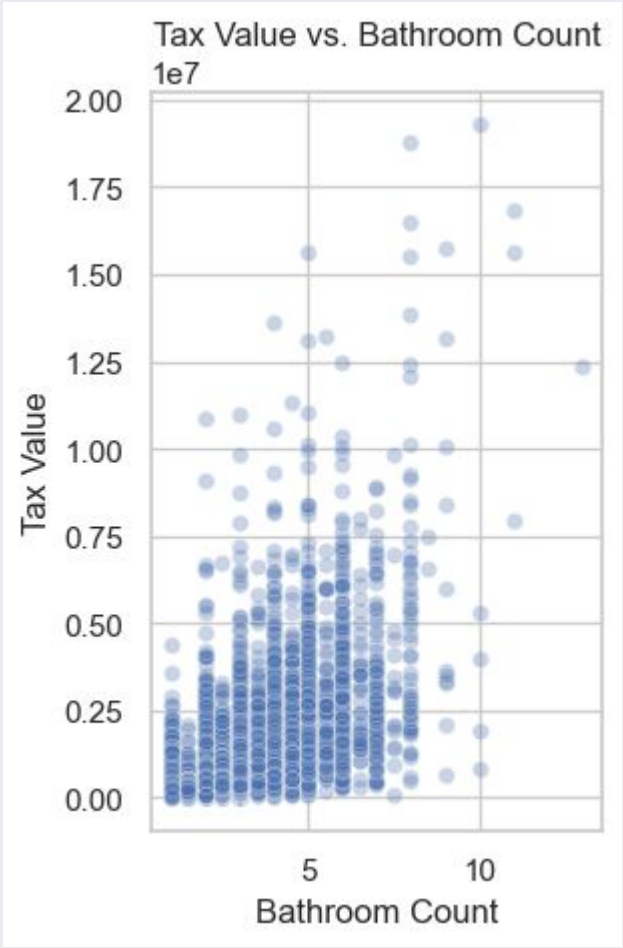
Distribution of Key Features



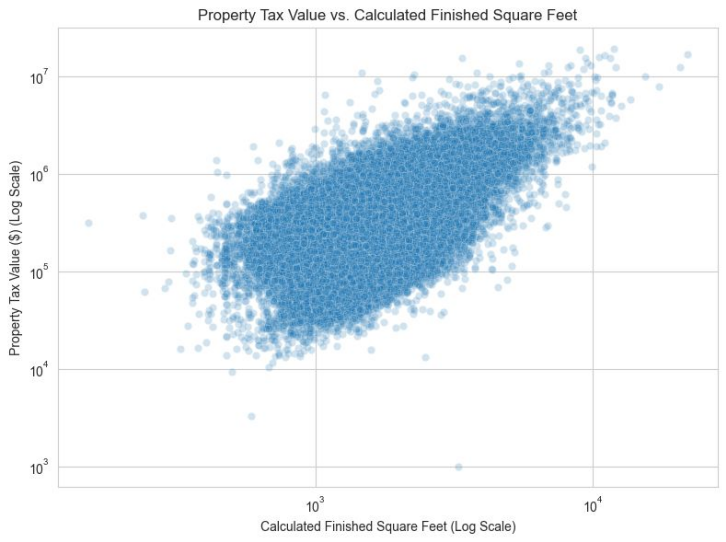
Trend - Visualized

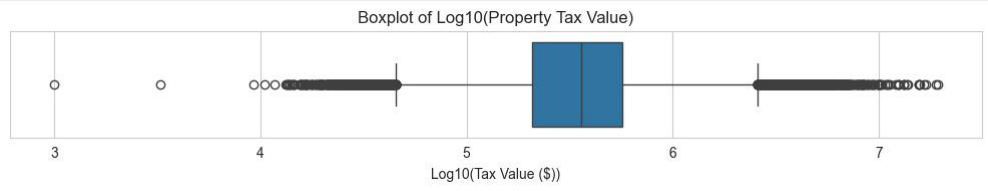
- A positive trend is visible
- Generally Larger homes have a higher tax value
- More bathrooms also correlates with a higher tax value





Relationship between property size and tax value





Measuring Success

- Normalized margin of error for RMSE was at 1.97% (RMSE/range) ($383,521.59/ ($19,280,700.00-$1,000.00))
- 2 percent is normally considered an excellent margin of error.
- Ultimately we would want to measure our success with real-world housing sales execution prices to evaluate applied success of our model
- For example, 2018 Zestimate had a margin of error of 13% (executed sales price vs Zestimate price) While the winning team had 4% error.

Future directions

- Advanced Feature Engineering: Incorporate external data (neighborhood demographics, school ratings, economic trends) and develop more complex geospatial features.
- Sophisticated Imputation: Employ methods like K-Nearest Neighbors (KNN) for missing values.
- Outlier Handling & Error Analysis: Conduct a thorough analysis of prediction errors, identify outlier impacts, and potentially develop specific sub models or use robust regression techniques less sensitive to extreme values. Consider alternative metrics like Mean Absolute Error (MAE).
- Target Refinement: If market value data becomes available, retrain the model to predict market prices directly or model the relationship between tax value and market value.

Business or practical implications

For a real estate company and investors, this model offers a data driven tool to estimate assessed property values.

- It can serve as a baseline for internal appraisals
- help identify potentially under or over assessed properties
- provide insights into the key drivers influencing property valuations

For individuals, this model will remove much of doubt in determining the correct price for buying/selling a property.

Team Contributions

Every individual in the team contributed equally to each aspect of the project.

There were certain times an individual had less to contribute due to nature of the project. But everyone was a team player and made up the work in the following days.