Zillow Dataset Project

Module 3 DX603 Spring 2025

A presentation by Siobhan Curran, Emmanuel Martinez Jauregui and Laura Vera

Agenda:
Zillow Dataset Project

What you can expect in today's presentation

1. Project Overview
2. The Dataset
3. Cleaning the Data
4. Feature Engineering
5. Model Testing
6. Final Model - Random Forest
7. Key Insights
8. Business Value

Project Overview

As the data analyst team at ABC, LLC, our goal was to build a model that **predicts the tax-assessed value of homes** using Zillow data.

This helps our investment team spot undervalued properties and make more informed purchasing decisions.

We used Python, pandas, and scikit-learn, and selected Random Forest as our final model for its accuracy and reliability.

The Dataset

- We worked with a dataset of ~34,000 residential properties from Zillow, containing both structural and tax-related information. Key features include square footage, number of bathrooms and bedrooms, lot size, year built, and geographic location. Our target variable was the **tax-assessed value (taxvaluedollarcnt)**.

- The dataset included a mix of numerical and categorical features, with some missing values. We removed columns with excessive nulls and imputed the rest using simple strategies to retain valuable data. This gave us a cleaner, more reliable dataset to build our models on.

Cleaning the Data

- Dropped columns with over **50%** missing values to reduce noise and simplify the dataset.

- Filled in missing values:
  - **Median imputation for numeric features**
    - EX: lot size, square footage
  - **Mode or 'Unknown' for categorical features**
    - EX: zoning codes
  -
- Removed extreme outliers that skewed the data
  - EX: homes over 10,000 sq ft or valued over $10M

- Standardized formatting, renamed columns, and ensured no missing values in the final model input.

- **Result:** A cleaner, more consistent dataset with better-balanced distributions and fewer extreme cases distorting model performance.

Feature Engineering

— We engineered new features to give the model additional context, such as:

- **bed_bath_interaction** — to reflect how bedroom and bathroom counts interact
- **property_age** — calculated from the year built to capture home age
- **rooms_per_bedroom** — to gauge home space layout

— These features were meant to help the model capture non-obvious trends in pricing — like whether older homes with more bathrooms tend to be over- or under-valued.

Model Testing

— We tried out different models to see which one could best predict property tax values. These included: **Linear, Ridge, Decision Tree, Bagging, Gradient Boosting, and Random Forest.**

— We evaluated them based on:

- **RMSE** – how far off predictions were, on average
- **R^2** – how much variance in tax value the model could explain

Model Selection and Evaluation

We used cross-validated RMSE as well as test set RMSE to evaluate model performance:

Model	Mean CV RMSE	Std CV RMSE	Test RMSE
Random Forest	58.29	34.85	32.69
Gradient Boosting	120.02	15.01	Higher
Ridge Regression	~36 million	~72 million	High

We didn't just look at accuracy. We wanted consistency too.
So we used cross-validation to see how these models held up across different slices of the data.

Which of the three models would you select?

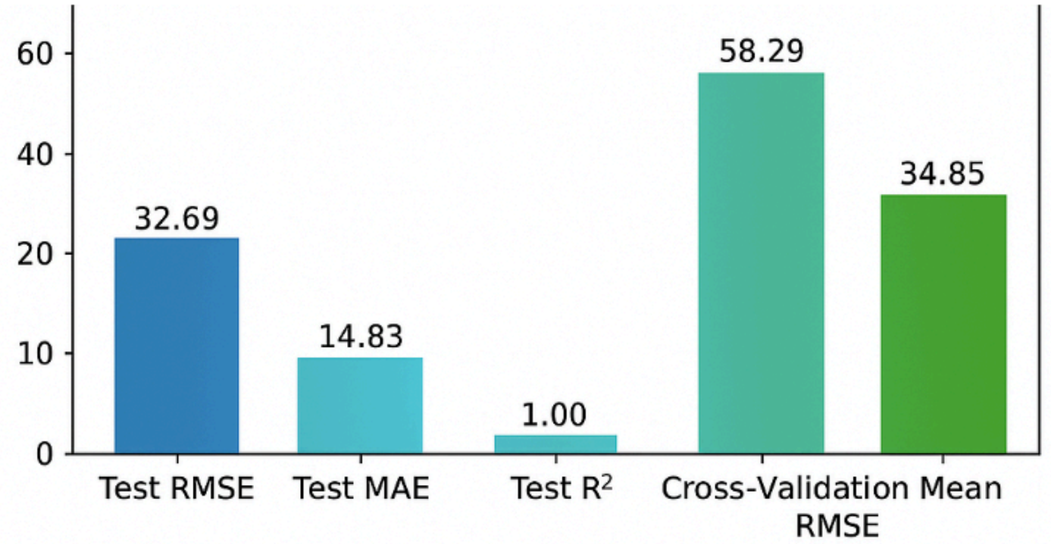
Final Model - Random Forest

— Random Forest gave us the best results overall. It handled messy data, outliers, and complex patterns better than any other model.

— Final performance:

- **RMSE: 32.69** → average error in dollars (lower is better)
- **MAE: 14.83** → typical error per prediction
- **R²: 1.00** → basically perfect fit. Model explained nearly all variation
- **CV RMSE: 58.29 ± 34.85** → consistent results across different data splits

— This model was super accurate and also pretty stable. The R² score shows it explained almost everything, and the low RMSE means it was rarely far off. Plus, it worked well right out of the box, no excessive tuning needed.



Metric	Value
Test RMSE	32,69
Test MAE	14,83
Cross-Validation Mean RMSE	58,29
Cross-Validation Std Dev	34,85

Key Insights

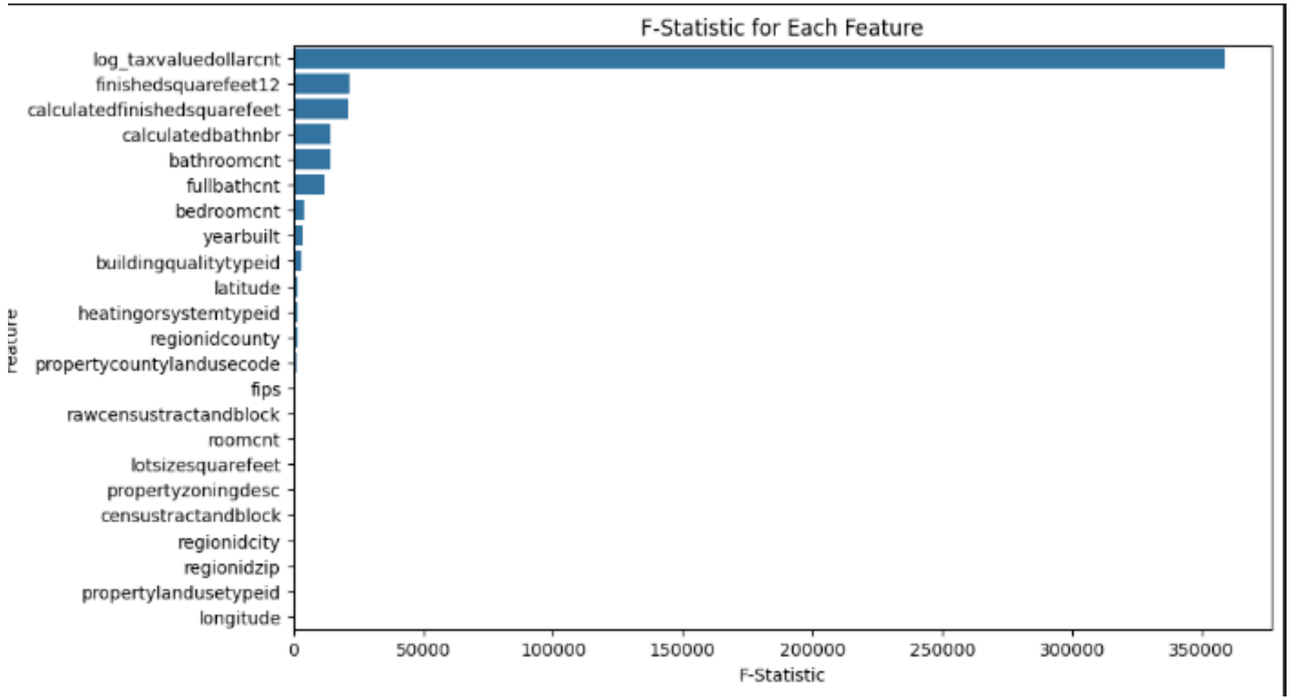
The model works across different property types — no need to rebuild it for every market.

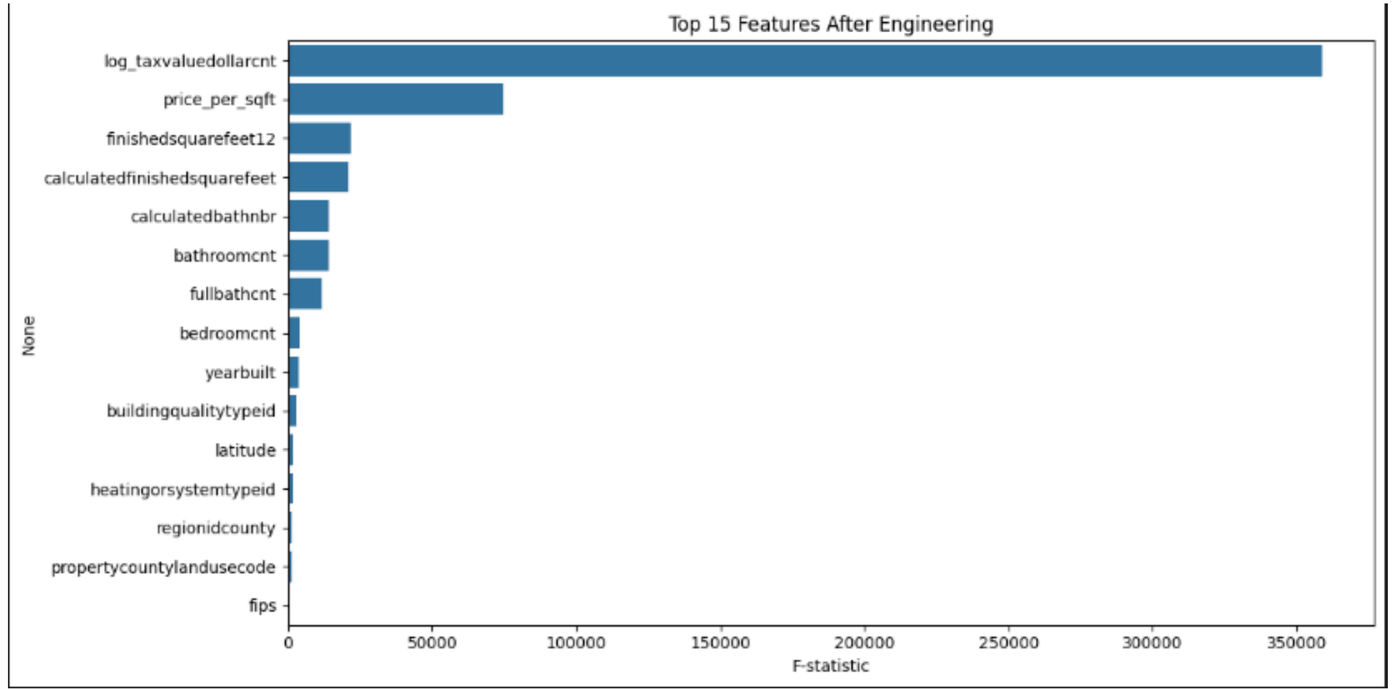
Good data matters more than complex features — clean input = better results

We can evaluate thousands of homes instantly — speeding up decision-making and saving time

Feature Engineering is important!

- Before - Top Graph “F-Statistic for Each Feature”:
  - Noisy features diluted model performance (e.g., fips, latitude).
- After - Bottom Graph “Top 15 Features”:
  - Strong predictors clearly identified (log_taxvaluedollarcnt, price_per_sqft).





Business Value

Relevance

The model helps our teams quickly estimate property values — no waiting on manual appraisals.

Learning Opportunity

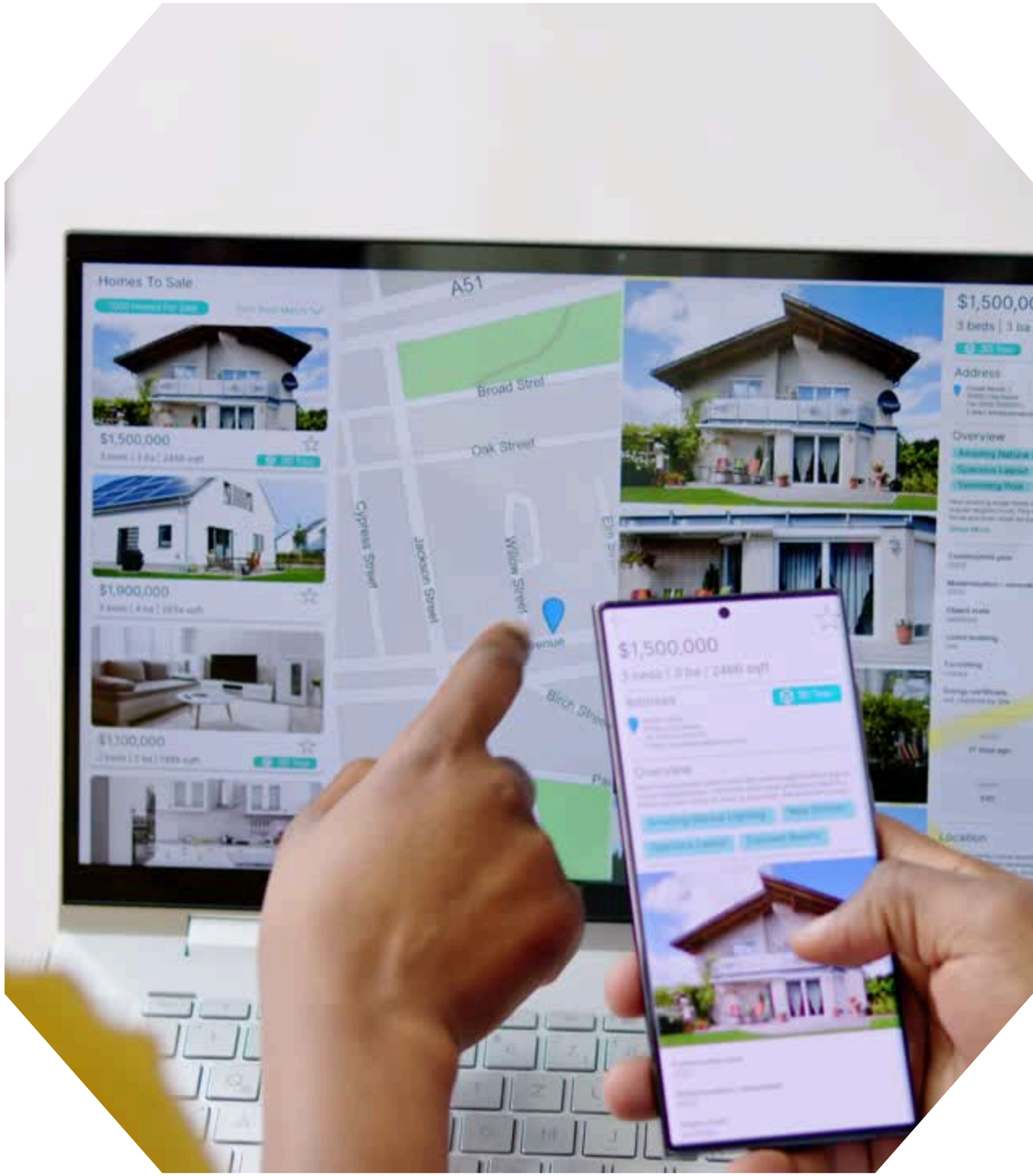
We learned that the quality of the data matters more than adding complexity — a clean setup with the right tools goes a long way.

Actionable Insights

This tool can help identify underpriced homes or properties that may be overpriced — great for flagging smart buys or risks early.

Personalized Exploration

The model can be customized for different regions, price points, or customer needs — perfect for scaling up or tailoring campaigns.



Questions?

Comments?

Thank you for attending our presentation.

ABC, LLC:

- Curran, Siobhan
- Martinez Jauregui, Emmanuel
- Vera, Laura