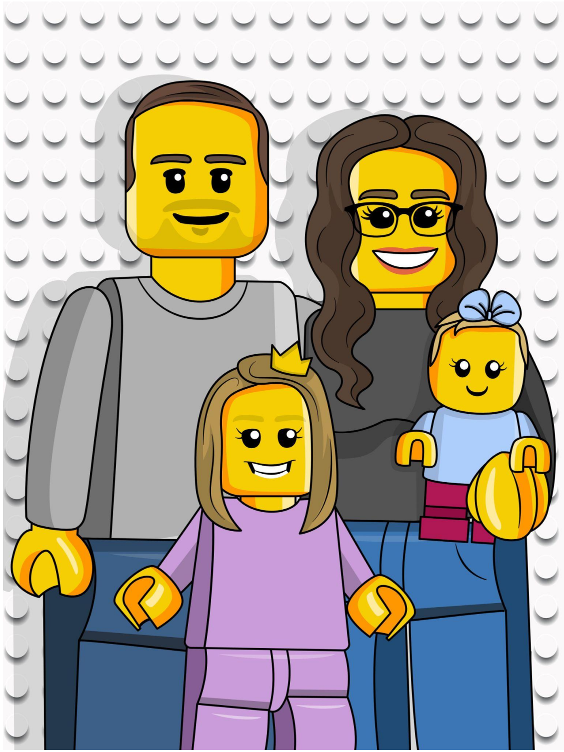
Predictive Model for Utah Housing Values

By Josh Funderburk

Intro to Josh

- Education
  - Bachelors of Science in Information Systems (2019) - University of Utah
  - Masters of Science in Data Analytics (current) - Western Governors University
- Employment
  - Included Health
    - Manager, Member Care Analytics (7/2024 - present)
    - Service Quality Analyst (8/2022 to 7/2024)
  - NexRep
    - Business Analyst Manager (1/2020 - 7/2022)
    - Workforce Management Analyst (12/2018 - 1/2020)
  - Upwell Health
    - Business Analyst (5/2017 - 12/2018)



Problem Statement & Hypothesis

Research Question:

Can property features such as square footage, lot size, and city predict Utah housing unit values?

Hypothesis:

- Null (H0): Property features do not significantly predict housing prices
- Alternative (H1): Property features significantly predict housing price

Why this matters:

- Utah ranks top 5 in population growth
- Need for accurate valuation method
- Traditional methods rely heavily on comparative analysis
- Data-driven approach could improve valuation accuracy and efficiency

Data Analysis: Preparation

1. Initial Data Collection & Cleaning
    - Utah Housing Unit Inventory: 688,270 records
    - Filtered to single family homes
    - Removed nulls
    - Removed invalid values
        i. Example: Total Value = $0
    - Applied outlier handling (IQR method)
2. Feature Engineering
    - Encoded categorical variables (City, County, Subtype)
    - Standardized numerical features
3. Model Setup
    - Split in to Training and Test sets
    - Training set (80%): 533,269 records
    - Test set (20%): 198,017 records

Data Analysis: Model Strategy

1. Algorithm Choice:
   - Random Forest Regression
   - Chosen for ability to handle mixed data types
   - Strong with non-linear relationships
2. Evaluation Framework:
   - R-squared: percentage of total value variation explained
   - Mean Squared Error: average squared distance from predicted values
   - Root Mean Squared Error: average error in dollars
   - Mean Absolute Error: average magnitude of prediction errors
3. Optimization Strategy:
   - Initial baseline model
   - Hyperparameter tuning
   - Final optimized model

Data Analysis: Initial Model

Core Metrics:

- R-squared: 84.2% variance explained
- Mean Absolute Error: ~$56,000

Key Insights:

- Model shows strong predictive power
- Error margin significant but reasonable for housing market
- Good foundation for optimization

```
R-squared Score: 0.842

Mean Squared Error: 10186082188.194
Root Mean Squared Error: 100926.122

Mean Absolute Error: 55948.001
```

Data Analysis: Tuning

- Hyperparameter tuning:
    - Number of trees
    - Maximum tree depth
    - Minimum samples required for node splitting
    - Minimum samples per leaf node
- Tuned on 5% sample of training data due to computational limitations (21,964 samples)
- Scoring metric: R-squared
- Best parameters R-squared = 0.799

Data Analysis: Final Model

Performance Metrics:

1. Explanatory Power:
   - R-squared: 80.5%
   - Slight decrease from initial model
   - More robust and generalizable
2. Error Measures:
   - Mean Absolute Error: $62,548
   - Root Mean Squared Error: $111,984
   - Context: ~12% of median home value ($509,000)
3. Key Takeaway:
   - Model shows strong predictive power
   - Error margin suggests need for traditional methods
   - Ready for production with proper safeguards

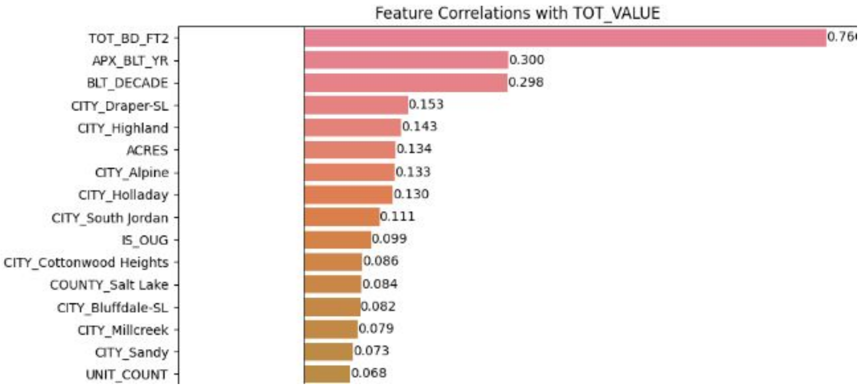
```
R-squared Score: 0.805

Mean Squared Error: 12540528306.686
Root Mean Squared Error: 111984.500

Mean Absolute Error: 62548.030
```

Key Findings

1. Model Performance:
   - Property features explain 80.5% of house price variation
   - Total building square footage strongest predictor (correlation = 0.778)
   - Building age shows meaningful influence (correlation = 0.198)
2. Location Impact:
   - City and county significantly affect values
   - Salt Lake County shows strongest location correlation
   - Higher density areas generally show higher values
3. Practical Implementation:
   - Average prediction deviation: $62,548
   - Context: 12% of median home value ($509,000)
   - Model suitable as supporting tool, not replacement



Limitations

1. Prediction Accuracy Challenges:
   - $62,548 average error on $509,000 median home value (~12%)
   - Current margin limits use as standalone tool
2. Data Coverage Limitations:
   - Only 8 of 29 Utah counties represented
   - Missing critical features (bedrooms, bathrooms, garage)
   - Interior quality data not available
3. Technical Constraints:
   - Complex model reduces interpretability
   - Limited computing power for model tuning
   - Results harder to explain to stakeholders

Proposed Actions

1. Use as baseline alongside traditional methods
    - Support, not replace, current valuation process
    - Provide objective starting points
2. Expand dataset coverage
    - Include remaining 21 Utah counties
    - Collect additional property features
    - Focus on key value indicators
3. Improve model performance
    - Focus on single-family homes initially
    - Enhance computational resources
    - Reduce prediction error margin

Expected Benefits

1. For Appraisers & Assessors:
   - Data-driven starting points for valuations
   - Potential to reduce assessment time by 30-40%
   - Consistent baseline across properties
2. For Property Owners:
   - More objective, transparent valuations
   - Faster assessment processes
   - Better understanding of value drivers
3. For Real Estate Market:
   - Supporting Utah's top 5 population growth
   - Foundation for advanced valuation tools
   - More efficient market operations

Thank you!