Executive Summary: Predictive Model for Utah Housing Values

Josh Funderburk

Western Governors University

Data Analytics Graduate Capstone (D214)

Task 3: Presentation of Findings

January 30, 2025

Executive Summary: Predictive Model for Utah Housing Values

Problem Statement and Hypothesis

Utah's real estate market is experiencing significant growth, ranking as a top 5 population growth state (Williams, 2024). According to Woodman (2024), while traditional valuation methods rely heavily on comparative market analysis, a data-driven approach using property features could provide more objective valuations. This analysis examines whether property features like square footage, lot size, and city location can predict Utah housing unit values. The study hypothesized that these property features would significantly predict housing prices, with the null hypothesis stating no significant predictive relationship exists.

Data Analysis Process

The analysis utilized the Utah Housing Unit Inventory dataset from the Utah AGRC's Open Data Portal, containing 18 variables and 688,270 records. The dataset includes continuous and categorical variables related to property characteristics such as square footage, acreage, location details, and build year. Of the 18 variables, Total Value is used as the dependent variable while 9 variables were selected for the model. Initial data exploration revealed significant variation in housing unit values, ranging from modest single-family homes to luxury properties, necessitating careful statistical treatment.

The data preparation process involved multiple stages:

1. Filtering the data set to only include the housing unit type of single family.
2. Removing null values from the dataset and records with invalid values (zeros) for any of the Total Value, Acres, and Total Building Square Feet variables were removed.
3. Applying Interquartile Range method to handle statistical outliers in the Total Value variable. A factor of 3.0 is used for the outlier detection in order to employ

a more conservative approach to outlier removal. Factors less than 3.0 remove too many outliers and cause the model to not capture some normal extremes in housing values. This and previous steps brought the data to a clean state with 549,101 records.

4. Encoding categorical variables of City, County, and Subtype using one-hot encoding.
5. Standardizing all numerical features using StandardScaler to ensure consistent scale.
6. Splitting data into an 80% training set (533,269 records) and a 20% test set (198,017 records).

Random Forest Regression is the algorithm employed in this analysis and R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are all used for model evaluation. The initial model performs considerably well with a R-squared indicating that 84.2% of the variance is explained by the model.

```
R-squared Score: 0.842

Mean Squared Error: 10186082188.194
Root Mean Squared Error: 100926.122

Mean Absolute Error: 55948.001
```

Following the initial model development, hyperparameter tuning is conducted on the model. GridSearchCV explores various combinations of common parameters to tune (Nolan, 2023), including the number of trees, maximum tree depth, minimum samples required for node splitting, and minimum samples per leaf node. Due to limitations with computation, only 5% (21,964 samples) of the training data set were used for hyper parameter tuning. However, several tests show consistency in the best parameters, so this analysis can be confident in trusting the results of the tuning. GridSearchCV determined the best parameters to be max

depth of 30, min samples leaf of 1, min samples split of 10, and n estimators of 300. The best parameters have an R-squared score of 0.799 on the subset of the training data.

The final model implemented with the best parameters achieved slightly different performance metrics than the initial model. The R-squared score adjusted to 0.805 and a Mean Absolute Error of $62,548.030. While these metrics show a slight decrease in performance compared to the initial model, they represent a more robust and generalizable solution, less likely to overfit the training data (Wikipedia, 2025).

```
R-squared Score: 0.805

Mean Squared Error: 12540528306.686
Root Mean Squared Error: 111984.500

Mean Absolute Error: 62548.030
```

Findings

The analysis results support the alternative hypothesis, with the final Random Forest model achieving an R-squared value of 0.805, indicating that approximately 80.5% of the variance in housing values can be explained by the selected property features. However, with a Mean Absolute Error of $62,548.03 on homes with a median value of approximately $509,000, the model's practical application is limited, as the error margin is roughly 12% of a typical property's value. This suggests that while the model demonstrates the predictive power of property features, there is significant room for improvement in prediction accuracy.

To better understand these results, analysis of individual feature correlations reveals total building square footage as the strongest predictor of housing value (correlation = 0.778), while building age demonstrates meaningful influence (correlation = 0.198). Location features like city and county reveal notable correlations, with areas like Salt Lake County exhibiting

robust associations with housing values. Property characteristics including acreage and density (DUA) demonstrate moderate predictive power in the model.

Technique and Tool Limitations

The primary limitations involve the dataset. Only 8 of Utah's 29 counties are represented in the dataset, potentially affecting the model's generalizability to other counties. Additionally, critical features such as the number of bedrooms, bathrooms, garage capacity, and interior finish quality are not available in the UGRC data set. These features are traditionally strong indicators of home value in real estate appraisals. The inclusion of these variables would likely improve the model's prediction accuracy and reduce the current error margin, making it more practical for real-world applications.

Random Forest Regression has several limitations as an analytical technique. The Random Forest algorithm's inherent complexity creates interpretability challenges compared to simpler models, particularly when explaining specific predictions to non-technical audiences. Several technical constraints impact the model's development and performance. Computational limitations necessitate using only 5% of training data (21,964 samples) for hyperparameter tuning.

Proposed Actions

Based on these results, the recommended course of action is to use the Random Forest model as a baseline for property valuations, while working to acquire more comprehensive property data. The model's current accuracy level suggests it should be used cautiously and only alongside traditional appraisal methods. However, several improvements could help it function independently.

The first step is to address computational limitations. Future modeling work should utilize more robust computing resources to enable hyperparameter tuning with the full training dataset

rather than the 5% sample currently used. Beyond computational resources, the primary focus should be on expanding the dataset to include the remaining 21 Utah counties not currently represented. Data collection efforts should prioritize gathering additional property features known to influence housing unit values, such as the number of bedrooms, bathrooms, garage capacity, and interior finish quality.

With or without additional features, another action would be to refine the analysis by focusing exclusively on single-family, single-unit properties to reduce variability in the dataset.

Expected Benefits

The Random Forest model offers several meaningful benefits as a supplementary valuation tool, providing a data-driven foundation that explains 80.5% of variance in housing values. This quantifiable foundation has the potential to streamline the assessment process by offering an objective starting point for valuations.

The model's predictive power is expected to improve with the inclusion of additional counties and property features. This improvement would reduce the current Mean Absolute Error of $62,548.03. By focusing exclusively on single-family, single-unit properties, the refined analysis provides more accurate valuations for this specific market segment, establishing a strong foundation before expanding into more complex property types.

As computational resources are enhanced and the dataset grows to include critical features like bedroom count and interior finish quality, the model's practical applications will strengthen. These improvements will ultimately lead to more efficient and objective property valuations across Utah's rapidly growing real estate market.

References

Nolan, R. (2023, October 24). Random Forest Regressor in Python: A Step-by-Step Guide.

YouTube. https://www.youtube.com/watch?v=YUsx5ZNlYWc

Wikimedia Foundation. (2025, January 27). Bias–variance tradeoff. Wikipedia. https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Williams, C. (2024, December 19). Utah returns to the top 5 in growth as US population soared in 2024. KSL.com. https://www.ksl.com/article/51214522/utah-returns-to-the-top-5-in-growth-as-us-population-soared-in-2024

Woodman, C. (2024, December 2). AVM in real estate: Automated Valuation Models explained. New Silver. Retrieved January 26, 2025, from https://newsilver.com/the-lender/avm-in-real-estate