# Improving Real Estate Investment Decisions Using Predictive Modeling in Bakersfield, CA

# Chapter 1 - Introduction

The real estate market presents significant opportunities for investors, but maximizing profitability requires precise decision-making regarding pricing. This project explores methods to enhance investment strategies in Bakersfield, California, through the application of predictive modeling. Using a private Multiple Listing Service (MLS) dataset combined with external neighborhood-level data, the project employs machine learning techniques to address the question: What is the ideal price at which to purchase a property in Bakersfield? To forecast property prices, ten machine learning models are trained and evaluated.

# Chapter 2 - Results

## 2.1 Introduction

This chapter presents the results of the predictive modeling analysis, aiming to answer the question: “What is the appropriate price to pay for a property in Bakersfield?”The analysis focuses on the creation and evaluation of multiple machine learning models designed to predict home value (i.e., the final sold price) using various combinations of features, preprocessing steps, and hyperparameter tuning. Models were trained on both the original dataset and a winsorized version. The performance of each model was assessed using Root Mean Squared Error (RMSE) and Adjusted R² on both training and test data. This chapter also examines model interpretability via SHAP analysis, provides visualizations such as actual-versus-predicted plots for the best-performing model, and discusses the implications of these results for real estate investors.

Early sections of this chapter detail the correlation-based feature selection results, which guided the removal of certain independent variables prior to modeling. Next, the RMSE and Adjusted R² performances of each model are discussed, highlighting notable differences between the algorithms as well as the impact of winsorizing the data. The final sections provide an in-depth analysis of the best model’s predictions on the test set, supported by a discussion of outliers, an actual-versus-predicted plot, and a SHAP feature-importance visualization. Concluding remarks summarize the overall performance and offer guidance on how these results can inform short-term real estate investment decisions.

## 2.2 Data Preparation and Feature Selection Results

Initial data exploration involved examining each independent variable’s relationship to the dependent variable—final sold price—and assessing correlations among independent variables. The original dataset contained property-level attributes such as square footage, lot size, number of bedrooms, school ratings, neighborhood safety indicators, commute scores, mortgage rates, and more. During correlation analysis, several variables showed either negligible correlation with the final sold price or significant redundancy with other independent variables. The negligible-correlation set included the age of the property at the time of sale, days on market, high school ratings, bike score, driving distance to the nearest shopping mall, and month of sale. Because these variables demonstrated minimal influence on the model’s predictive strength, they were removed to streamline the feature space. In addition, the variable for bathrooms exhibited high correlation with square footage. Retaining both independent variables risked multicollinearity issues and added model complexity without notable gains in predictive performance. Consequently, bathrooms variable was also dropped, leaving a focused set of features that captured structural attributes, mortgage rates, local safety, commute options, and other community-level information.

After removing these variables, 21 independent variables (including dummy-encoded categorical variables such as zip codes) remained for model-building. The final feature set included Sq Foot, Lot Size, Bedrooms, Elementary School Ratings, Junior High School Ratings, Sex Offender Count, Walk Score, Transit Score, Downtown Distance (Meters), Avg Monthly Mortgage Rate (30Y), Pool\_Inground, Pool\_Community, Pool\_Spa, an interaction term (Sex Offender Count \* Walk Score), and dummy variables for zip codes: Zipcode\_93307, Zipcode\_93308, Zipcode\_93311, Zipcode\_93312, Zipcode\_93313, Zipcode\_93314, and Zipcode\_Other.

Two versions of the filtered dataset were created. One kept all original numeric values. The other winsorized extreme outliers to mitigate the disproportionate influence that exceedingly large or small sale prices and independent variable values can have on model performance. These datasets were subjected to recursive feature elimination (RFE) for certain algorithms—namely linear regression, elastic net, gradient boosting, and random forest—to select the most relevant subset of independent variables. Because RFE in its standard implementation is not compatible with neural networks, the neural network models used all 21 independent variables. Instead of RFE, SHAP analysis was employed to interpret feature importance and assess the model's decision-making process. Both the original and winsorized data were subsequently split into training (85%) and test (15%) sets. The training partition underwent 10-fold cross-validation for model training and hyperparameter tuning, while the final model performance was evaluated on the held-out test subset.

## 2.3 Model Performance on Original Dataset

The original dataset yielded five predictive models: linear regression (LR\_Original), elastic net (EN\_Original), gradient boosting (GBR\_Original), random forest (RF\_Original), and a neural network (NN\_Original). Training performance was summarized using cross-validation RMSE, while the final evaluation was based on test RMSE and Adjusted R² on the 15% held-out data. Table 1 shows the performance metrics for different models on the original dataset.

The linear regression and elastic net models displayed similar outcomes. Linear regression (LR\_Original) achieved a training cross-validation RMSE of about 82,968, a training RMSE of approximately 82,453, and a training Adjusted R² close to 79.74%. On the held-out test set, it yielded an RMSE of 81,043 and an Adjusted R² of about 77.22%. Elastic net (EN\_Original) performed almost identically, with a cross-validation RMSE of around 82,953, a training RMSE of 82,414, and a training Adjusted R² of roughly 79.75%. The test RMSE was 80,998 and test Adjusted R² was approximately 77.19%, nearly matching the linear regression’s performance. These results suggest that the inclusion of L1 and L2 regularization in elastic net did not drastically alter predictive capability relative to the simpler linear model.

In contrast, gradient boosting (GBR\_Original) and random forest (RF\_Original) demonstrated significantly stronger training performance. Gradient boosting’s training RMSE (29,751) and Adjusted R² (97.35%) reflected a near-perfect fit on the training data, while its cross-validation RMSE of about 54,492 hinted that it generalized moderately well to unseen samples. On the test set, it delivered an RMSE of 67,174 and an Adjusted R² of 84.25%. Random forest overfit the training set even more, with a training RMSE of 25,445 and an Adjusted R² of about 98.06. Nevertheless, it still generalized effectively, scoring a test RMSE of 63,474 and an Adjusted R² of 85.94%. Both ensemble methods thus captured considerable nonlinearity and feature interactions. The random forest model, in particular, maintained the best overall generalization among these four models on the original dataset.

The neural network (NN\_Original) achieved a strong trade-off between robust training performance and generalization. Its training cross-validation RMSE reached about 67,139, and the final training RMSE was 66,544. These values correspond to an Adjusted R² of 86.79% on the training set. On the test set, the neural network achieved an RMSE of 67,465 and an Adjusted R² of 84.17%. Although its RMSE was slightly higher than the random forest’s 63,474, the neural network’s performance overall placed it within the same range of strong predictive ability. Indeed, random forest’s overfitting tendencies were slightly more pronounced, but both models performed substantially better than linear or elastic net regressions in capturing the complexity of real estate price data.

**Table 1**

*Performance Metrics for Models on the Original Dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **RMSE (Cross Validation)** | **RMSE (Training)** | **Adjusted R² (Training)** | **RMSE (Test)** | **Adjusted R² (Test)** |
| LR\_Original | 82,968 | 82,453 | 79.74% | 81,043 | 77.22% |
| EN\_Original | 82,953 | 82,414 | 79.75% | 80,998 | 77.19% |
| GBR\_Original | 54,492 | 29,751 | 97.35% | 67,174 | 84.25% |
| RF\_Original | 62,311 | 25,445 | 98.06% | 63,474 | 85.94% |
| NN\_Original | 67,139 | 66,544 | 86.79% | 67,465 | 84.17% |

*Note.* In table 1, “Model” indicate the algorithm used: LR (Linear Regression), EN (Elastic Net), GBR (Gradient Boosting Regressor), RF (Random Forest), and NN (Neural Network), all trained on the original dataset.

## 2.4 Model Performance on Winsorized Dataset

A parallel modeling approach was applied to the winsorized dataset, producing five additional models (LR\_Winsorized, EN\_Winsorized, GBR\_Winsorized, RF\_Winsorized, and NN\_Winsorized). Here again, the training sets were subjected to 10-fold cross-validation, with final results reported on the 15% test sample.

Linear regression (LR\_Winsorized) and elastic net (EN\_Winsorized) converged to nearly identical solutions, with cross-validation RMSE values hovering around 65,503 and 65,500, respectively, and training RMSE near 65,173 for linear regression and 65,178 for elastic net. Both presented training Adjusted R² around 82.45%. Their generalization to the test set was similar as well, with a test RMSE of about 79,152–79,185 and a test Adjusted R² around 78.14%–78.12%. The slight improvement in Adjusted R², compared to the original dataset’s linear or elastic net models, suggests that winsorizing extreme outliers might have alleviated the disproportionately negative effect that very large property prices can have in a conventional linear approach.

The ensemble-based models exhibited some variability. Gradient boosting (GBR\_Winsorized) showed a relatively low training cross-validation RMSE of 61,424 and a strong training RMSE of around 48,770, yielding an Adjusted R² above 90%. Yet, on the test data, its performance deteriorated to an RMSE of about 91,700 and an Adjusted R² of 70.69%. The difference between training and test performance suggests that gradient boosting did not generalize as effectively under winsorization, potentially due to certain high-value properties being capped, thus altering the model’s learned patterns. Random forest (RF\_Winsorized) likewise showed good training results (about 50,997 RMSE and 89.27% Adjusted R²), but its test RMSE of approximately 81,978 and Adjusted R² of around 76.75% indicated stronger generalization than gradient boosting on the winsorized dataset but weaker performance than the best random forest under the original data.

Notably, the neural network (NN\_Winsorized) presented the strongest test performance among the winsorized models. It obtained a training cross-validation RMSE of around 55,411, a training RMSE of 54,715, and an Adjusted R² near 87.63%. The test RMSE was around 71,106, with an Adjusted R² of 82.36%. Although it did not surpass the test-set results of the best neural network or the best random forest on the original data, it outperformed every winsorized competitor by a considerable margin. This indicates that neural networks can remain robust to outliers after capping extreme values, likely due to the architecture’s flexibility in modeling complex functional relationships.

**Table 2**

*Performance Metrics for Models on the Winsorized Dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **RMSE (Cross Validation)** | **RMSE (Training)** | **Adjusted R² (Training)** | **RMSE (Test)** | **Adjusted R² (Test)** |
| LR\_ Winsorized | 65,503 | 65,173 | 82.45% | 79,152 | 78.14% |
| EN\_ Winsorized | 65,500 | 65,178 | 82.44% | 79,185 | 78.12% |
| GBR\_ Winsorized | 61,424 | 48,770 | 90.17% | 91,700 | 70.69% |
| RF\_ Winsorized | 55,993 | 50,997 | 89.27% | 81,978 | 76.75% |
| NN\_ Winsorized | 55,411 | 54,715 | 87.63% | 71,106 | 82.36% |

*Note.* In table 2, “Model” indicate the algorithm used: LR (Linear Regression), EN (Elastic Net), GBR (Gradient Boosting Regressor), RF (Random Forest), and NN (Neural Network), all trained on the winsorized dataset.

## 2.5 Best Model Selection and Visualizations

The final step in evaluating which model best answered “How much should I pay for a property in Bakersfield?” involved comparing performance metrics on the original and winsorized datasets. Although Random Forest (RF\_Original) achieved the lowest test RMSE (about 63,474), it showed a larger gap between its training RMSE (about 25,445) and test RMSE, suggesting a higher likelihood of overfitting. By contrast, the neural network on the original dataset (NN\_Original) displayed a much smaller difference between its cross validation RMSE (about 67,139) training RMSE (about 66,544) and final test RMSE (about 67,465). This consistency in training, cross-validation, and test performance makes NN\_Original the most reliable overall model.

Additionally, NN\_Original maintained an Adjusted R² of 84.17% on the test set, which, while slightly behind Random Forest’s 85.94%, was still a marked improvement over linear or elastic net regressions. Its balanced performance indicates that it is effectively capturing the nonlinearity of Bakersfield’s real estate market without unduly overfitting to training data. This conclusion is reinforced by the small discrepancy between NN\_Original’s cross-validation and test RMSE values, demonstrating stable generalization.

To visualize NN\_Original’s predictions, an actual-versus-predicted scatter plot was generated using the test set. Figure 1 presents this plot, where most data points lie closely along the ideal diagonal, but some outliers are present, indicating instances where the model both underestimates and overestimates home values. This plot highlights the model’s predictive capability while also revealing areas where estimation errors occur.

**Figure 1**

*Actual vs. Predicted Plot for NN\_Original (Best Model)*

A graph with dots and lines

AI-generated content may be incorrect.

## 2.6 SHAP Feature Importance Analysis

To further investigate how specific independent variables influence the best model’s output, a SHAP (Shapley Additive Explanations) analysis was performed on NN\_Original’s test-set predictions. Figure 2 presents a summary bar plot, highlighting the factors that most influenced predicted home values. Among the highest-ranking features was square footage, suggesting that living area plays a dominant role in shaping property prices across Bakersfield. Certain zip codes also displayed high SHAP value magnitudes. For instance, homes in zip codes 93312, 93314, and 93311 contributed strongly to price increases, indicating these neighborhoods’ premium nature in the local market.

Macroeconomic variables also emerged as relevant. The average 30-year mortgage rate had a notable impact, supporting the idea that borrowing costs can decrease or increase property values. In line with neighborhood quality perceptions, the sex offender count in a zip code consistently ranked among the more influential independent variables, as did the presence of an inground pool, which often corresponds with higher market desirability in warm climates like Bakersfield. Lot size, junior high school ratings, walk scores, and transit scores were similarly reflected in the SHAP values, underscoring that access to quality education and commuting options all play a role—though to a lesser extent than living area and zip code location—in shaping property prices. The SHAP results thus confirm the intuitive importance of both structural (size, pool availability), neighborhood-specific (zip code, safety, commute), and economic (mortgage rates) variables in predicting home values.

**Figure 2**

*SHAP Summary Plot: Feature Impact on NN\_Original (Test Set)*

A graph of a number of people

AI-generated content may be incorrect.

## 2.7 Summary

This chapter presented the outcomes of predictive modeling to address the research question, “How much should I pay for a property in Bakersfield?” by estimating home values using a combination of property-level and economic factors. Correlation analyses revealed that certain independent variables (e.g., property age, days on market, bike score, high school ratings) contributed minimal explanatory value and were therefore removed. Bathrooms variable was eliminated due to high correlation with square footage. The final subset of features was used in two versions of the dataset: one preserving all original values and another winsorized to curb outliers. Multiple learning algorithms—linear regression, elastic net, gradient boosting, random forest, and neural networks—were then fit and evaluated using 10-fold cross-validation on the training portion and subsequent validation on a 15% test sample.

Results showed that linear-based models yielded moderate accuracy, whereas random forest and neural networks captured significantly more variance in home values. Random forest attained particularly low training RMSE values, indicative of potential overfitting, although it still performed well in out-of-sample tests. The neural network model (NN\_Original) on the original dataset balanced training accuracy with test-set generalization, achieving a test RMSE near 67,465 and an Adjusted R² of about 84.17%—one of the best performances overall. SHAP analysis emphasized that square footage, certain zip codes, average mortgage rates, and safety indicators, among others, collectively have the strongest influences on predicted home values. Graphical inspection of the neural network’s actual-versus-predicted scatter plot revealed that while most test instances clustered near the ideal diagonal, some outliers indicated instances of both underestimation and overestimation. This highlights the model’s overall predictive strength while acknowledging areas where estimation errors occur.

From an industry standpoint, these results highlight the practicality of applying advanced machine learning to smaller markets like Bakersfield, offering robust price estimates that aid investors in setting bidding strategies and judging potential property values.