Predictive Analysis of Real Estate Prices in the USA

Samuel R Morales

Western Governors University

Table of Contents

[A. Project Highlights 4](#_Toc1868160802)

[B. Project Execution 5](#_Toc2011082291)

[C. Data Collection Process 5](#_Toc1191971028)

[C.1 Advantages and Limitations of Data Set 6](#_Toc10980602)

[D. Data Extraction and Preparation 6](#_Toc512824111)

[E. Data Analysis Process 6](#_Toc1571068809)

[E.1 Data Analysis Methods 6](#_Toc1409685802)

[E.2 Advantages and Limitations of Tools and Techniques 7](#_Toc737998485)

[E.3 Application of Analytical Methods 7](#_Toc825242289)

[F Data Analysis Results 7](#_Toc1939531306)

[F.1 Statistical Significance 7](#_Toc109772385)

[F.2 Practical Significance 8](#_Toc1419440648)

[F.3 Overall Success 8](#_Toc1210557849)

[G. Conclusion 9](#_Toc1495356842)

[G.1 Summary of Conclusions 9](#_Toc613065263)

[G.2 Effective Storytelling 9](#_Toc16322062)

[G.3 Recommended Courses of Action 9](#_Toc660897025)

[H Panopto Presentation 9](#_Toc2048387211)

[References 12](#_Toc1410715514)

[Appendix A 12](#_Toc1695981861)

[Title of Appendix 13](#_Toc1655579738)

[Appendix B 13](#_Toc1686431257)

[Title of Appendix 14](#_Toc334098601)

[Appendix C 14](#_Toc1371871733)

[Title of Appendix 15](#_Toc38061623)

[Appendix D 15](#_Toc724305577)

[Title of Appendix 16](#_Toc1812292947)

# A. Project Highlights

The research question addressed by this project was identifying the key factors influencing real estate prices in the USA and how they can be used to accurately predict those prices.

The scope of the project included collecting and preprocessing a real estate dataset, developing and tuning a predictive machine learning model, and creating visualizations to interpret the model’s predictions.

The solution utilized Python with libraries such as Pandas, NumPy, and LightGBM within a Visual Studio Code environment with the Jupyter extension. The CRISP-DM methodology guided the execution of the project and ensured a structured approach to data analysis and model development.

# B. Project Execution

The original project plan outlined in Task 2 was followed, and the goal and all deliverables and objectives were met without any variance. The CRISP-DM methodology was followed but took less time than expected. The model used can accept categorical features directly, eliminating the need for one-hot encoding. Additionally, the modeling and evaluation phase overlapped with deliverables 2 and 3, saving a significant amount of time. The revised deliverable timeline is below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone or deliverable | Duration  (days) | Projected start date | Anticipated end date | Actual End Date |
| Deliverable 1: Cleaned Dataset | 14 | 12/18/23 | 1/1/24 | 12/20/23 |
| Deliverable 2: ML Model | 20 | 1/2/24 | 1/22/24 | 12/25/23 |
| Deliverable 3: Visualizations | 8 | 1/23/24 | 1/31/24 | 12/26/23 |

# C. Data Collection Process

The original plan of data collection for real estate data was followed, and the dataset was downloaded from Kaggle as a CSV file. No obstacles were encountered during data collection, and no unplanned data governance issues arose.

## C.1 Advantages and Limitations of Data Set

* Advantage: The extensive number of entries in the dataset allowed a large sample size, which enhanced the reliability of predictive modeling. For example, the large number of entries allowed incomplete records and outliers to be removed without worrying about over-shrinking the dataset.
* Disadvantage: Categorical features such as city, state, and zip code held many unique values that impacted model choice. This is why LightGBM was chosen, as it can handle categorical variables naturally and reduces project complexity and development time. Without this feature, one-hot encoding and Principal Component Analysis would have been necessary but still may have led to an unreasonable dimensionality.

# D. Data Extraction and Preparation

Data extraction was simple as the dataset was downloaded as a CSV file from Kaggle with a web browser. The data preparation used the following tools:

* Python: Chosen language for preparation, analysis, and model development.
* Pandas: For exploration and cleaning the dataset efficiently.
* Matplotlib & Seaborn: For creating visualizations of the dataset to aid preparation.
* Visual Studio Code with Jupyter: IDE to run the code and allow interactive analysis.

These tools were used as they provided an efficient and effective path for data preparation. The steps for preparation were:

* Removal of the records with missing values. Imputation of critical features of a house might introduce bias or inaccuracies. Data integrity was better preserved through removal, given the size of the dataset.
* Removal of irrelevant columns. The previously sold date column was dropped as it is not a feature of the house. Additionally, the status was dropped as all entries were the same value.
* Boxplots of noncategorical features such as acre lot, bath, bed, house size, and price were created to understand the distribution of values and identify outliers.
* Outliers were removed from the dataset using the interquartile range method.

These steps were critical to prepare the dataset for model development and avoid outliers that would negatively impact its predictive power.

# E. Data Analysis Process

## E.1 Data Analysis Methods

LightGBM Model: The LightGBM model is an advanced gradient boosting framework that is efficient for large datasets and capable of handling many features. It was used for its ability to manage the dataset’s complexity, delivering fast and accurate predictions for real estate prices. It was suitable for the project’s goal due to its feature importance capability, which directly informs the significance of hypothesized factors affecting real estate prices.

Linear Regression: In the final stages of analysis, Linear Regression was employed as a benchmark to verify the performance of the LightGBM model. Its simplicity and interpretability provided a baseline to ensure that the more complex LightGBM model was performing correctly and significantly better than the baseline model, thus supporting the hypothesis with a comparative perspective.

## E.2 Advantages and Limitations of Tools and Techniques

LightGBM Model:

* Advantage: Efficient with large datasets and handles categorical features intrinsically, boosting predictive accuracy.
* Limitation: Its complexity can lead to overfitting if not properly tuned, and results can be less interpretable compared to simpler models

Linear Regression:

* Advantage: Highly interpretable, making it easier to understand the relationship between variables and the outcome.
* Limitation: Assumes a linear relationship between variables, which can be too simplistic for capturing complex patterns.

## E.3 Application of Analytical Methods

LightGBM Model:

* Steps
  1. Encoded categorical variables to be compatible with LightGBM.
  2. Separated features from the target variable.
  3. Split data into training and testing sets.
  4. Trained model on the training set.
  5. Evaluated model performance on the testing set using R2, MSE, RMSE, and MAE.
  6. Ensured model was not overfitting the dataset by using cross-validation.
  7. Found optimal hyperparameters with Hyperopt.
  8. Evaluated model performance again with new hyperparameters. Early stopping rounds, bagging fractioning, feature fractioning, and cross-validation were used to ensure the model was not overfit.
* Requirements and Verification:
  1. Preprocessing ensured no missing values were present, meeting the model’s input requirements.
  2. Categorical variables were set to the appropriate type required by the model.
  3. Data was split to ensure a distinct separation of training and testing sets.
  4. Hyperparameter tuning and model performance were confirmed by implementing a robust validation strategy.

Linear Regression:

* Steps:
  1. Performed cross-validation with five splits to ensure robust model validation.
  2. Applied frequency encoding for categorical variables within each fold to handle non-numeric data and avoid data leakage.
  3. Trained Linear Regression model on the training set.
  4. Predicted and evaluated the model on the validation set.
* Requirements and Verification:
  1. The model’s results were verified by calculating the same metrics as the LightGBM model, allowing for a direct comparison of performance.
  2. Categorical variables were encoded using frequency encoding as categorical variables are not natively supported.

# F Data Analysis Results

## F.1 Statistical Significance

Model Analysis:

* Model: Supervised Regression
* Algorithm and Process: LightGBM model with tuned hyperparameters and cross-validation.
* Metrics Used: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²).
* Benchmark: R2 ≥ 0.6 and MSE < 2,500,000,000
* Performance:
  + MSE: 1,095,151,933
  + MAE: 15033
  + RMSE: 33088
  + R2: 0.9865
* Conclusion: The model achieved an R2 of 0.9865, substantially exceeding the benchmark, demonstrating a high level of predictive accuracy and an MSE that is less than half of the benchmark. These provide strong support for the hypothesis that key factors can be used to accurately predict real estate prices.

## F.2 Practical Significance

The practical significance of the LightGBM model's results lies in its high predictive accuracy, as indicated by the R2 of 0.9865. This level of accuracy means the model can reliably predict real estate prices based on the identified key factors.

For example, a real estate company could use this model to assess the market value of properties, enhancing their pricing strategy and investment decisions. The model’s accuracy ensures that the pricing is closely aligned with market expectations, reducing the risk of overvaluing or undervaluing properties. This could lead to more lucrative transactions and higher profitability in the real estate market.

## F.3 Overall Success

The project was successful in creating a highly accurate predictive model for real estate prices in the USA. The LightGBM model’s performance with an R2 of 0.9865 demonstrates its effectiveness in predicting prices based on key factors. This high level of accuracy ensures practical significance, as the model can reliably inform real estate pricing strategies and investment decisions. The project met its goals by providing a data-driven solution to understand and predict real estate market dynamics, supporting more informed and efficient decision-making in the industry.

# G. Conclusion

## G.1 Summary of Conclusions

The project concluded that key factors such as property size, location, and number of bedrooms and bathrooms can be effectively used to predict real estate prices in the USA. The LightGBM model, with its high accuracy, effectively captured the relationships between these factors and property prices. The results demonstrate the feasibility and effectiveness of using advanced machine learning techniques in real estate market analysis. These findings can be applied to various practical scenarios to aid stakeholders in making data-driven decisions in the real estate domain.

## G.2 Effective Storytelling

The visualizations created for this project significantly enhanced the storytelling aspect of the data analysis. The scatterplot of residuals from the LightGBM model illustrated the variance between predicted and actual prices, highlighting the model’s accuracy. The barplot of feature importance from the model visually conveyed which factors were the most influential in predicting real estate prices, offering intuitive insights into market dynamics. These visuals, developed using Python libraries like Matplotlib and Seaborn, effectively communicated complex data in an accessible manner, making the analytical findings more understandable and actionable for stakeholders.

## G.3 Recommended Courses of Action

1. Implement Model in Real Estate Valuation: Real estate companies should integrate the predictive model into their valuation processes. Given its high accuracy, the model can significantly improve pricing strategies, aligning them more closely with market dynamics.
2. Adapt Model for Market Trend Analysis: The model’s ability to identify key price influencers can be leveraged for market trend analysis. This would aid investors and real estate professionals in understanding emerging trends and making informed investment decisions.

# H Panopto Presentation

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=16d1a956-5201-43e6-ad09-b0f0002e1250>

# References

No sources were cited.