Predictive Analysis of Real Estate Prices in the USA

Samuel R Morales

Western Governors University

Table of Contents

[A. Proposal Overview 4](#_Toc2137083979)

[A.1 Research Question or Organizational Need 4](#_Toc1815560889)

[A.2 Context and Background 4](#_Toc1384268585)

[A.3 and A3A Summary of Published Works and Their Relation to the Project 4](#_Toc2079458267)

[Review of Work 1 5](#_Toc694593955)

[Review of Work 2 5](#_Toc475195288)

[Review of Work 3 5](#_Toc1156669356)

[A.4 Summary of Data Analytics Solution 5](#_Toc374341712)

[A.5 Benefits and Support of Decision-Making Process 6](#_Toc1116633891)

[B. Data Analytics Project Plan 6](#_Toc1561262715)

[B.1 Goals, Objectives, and Deliverables 6](#_Toc1217623361)

[B.2 Scope of Project 7](#_Toc427690838)

[B.2.A Included in Project Scope 7](#_Toc1243851001)

[B.2.B Not included in Project Scope 7](#_Toc2036717847)

[B.3 Standard Methodology 7](#_Toc204762054)

[B.4 Timeline and Milestones 8](#_Toc1049840195)

[B.5 Resources and Costs 8](#_Toc313275103)

[B.6 Criteria for Success 9](#_Toc1337049307)

[C. Design of Data Analytics Solution 9](#_Toc764513894)

[C.1 Hypothesis 9](#_Toc1169566000)

[C.2 and C.2.A Analytical Method 9](#_Toc1622461191)

[C.3 Tools and Environments 10](#_Toc53928856)

[C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance 10](#_Toc1386867513)

[C.5 Practical Significance 11](#_Toc157862960)

[C.6 Visual Communication 11](#_Toc1263386613)

[D. Description of Dataset 12](#_Toc34323025)

[D.1 Source of Data 12](#_Toc1867631323)

[D.2 Appropriateness of Dataset 12](#_Toc418252330)

[D.3 Data Collection Methods 12](#_Toc1665164903)

[D.4 Observations on Quality and Completeness of Data 12](#_Toc57902398)

[D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances 13](#_Toc1649406101)

[References 14](#_Toc1322159537)

[Appendix A 14](#_Toc1046046101)

[Title of Appendix 15](#_Toc804972029)

[Appendix B 15](#_Toc128761729)

[Title of Appendix 16](#_Toc1200922991)

[Appendix C 16](#_Toc1897962072)

[Title of Appendix 17](#_Toc1862588384)

[Appendix D 17](#_Toc1292381938)

[Title of Appendix 18](#_Toc390418831)

# A. Proposal Overview

## A.1 Research Question or Organizational Need

What are the key factors influencing real estate prices in the USA and how can they be used to correctly predict prices?

## A.2 Context and Background

Real estate prices in the USA are influenced by a complex dynamic of factors such as economic, demographic, and geographical factors. Understanding the driving features of property prices is vital for stakeholders. The ability to accurately predict real estate prices for individuals can influence decisions in investing, selling, or buying property. This project aims to dissect property features and harness them in a predictive model.

## A.3 and A3A Summary of Published Works and Their Relation to the Project

### Review of Work 1

This article from Charles Schwab discusses factors that influence real estate prices, with a focus on interest rates, government policies and unexpected events. This works provides valuable insight for the research question as understanding these factors is crucial for predicting real estate prices and market trends. (Factors that influence real estate prices, 2023)

### Review of Work 2

McKinsey’s article explores the potential of real estate analytics and emphasizes the use of unconventional data along with advanced analytics like machine learning to predict real estate prices. This source demonstrates the importance of integrating diverse data sources and advanced analytical methods to develop accurate and sophisticated predictive models. (Asaftei, Doshi, Means, & Sanghvi, 2018)

### Review of Work 3

This study published on Hindawi explores housing price prediction using multiple linear regression. It identifies key factors and categorizes them into transportation, environmental, and house factors. This paper directly addresses the research question and applies a statistical method to predict real estate prices. It provides a methodological approach to quantify the impact of various factors on housing prices. (Zhang, 2021)

## A.4 Summary of Data Analytics Solution

## The analytical solution for real estate price prediction will involve utilizing Python and specific libraries (Pandas, Matplotlib, Seaborn, Scikit-learn). The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology will be used during analysis. The data will be assessed, prepared, and cleaned which will then be used to create a regression model. The model’s statistical significance will be determined by the coefficient of determination (R2) and the Mean Squared Error (MSE). This will provide a comprehensive assessment of the model’s accuracy and predictive power.

## A.5 Benefits and Support of Decision-Making Process

The analysis will benefit decision-making in real estate by providing an accurate model to predict real estate prices and insight into some of the driving factors behind their cost. By analyzing market data, we can identify trends and patterns which will allow stakeholders to make informed decisions about buying, selling, or investing in properties. The use of R2 and MSE as metrics ensures the reliability of predictions. This data-driven approach reduces the reliance on intuition and supports a more strategic and evidence-based process in the real estate sector.

# B. Data Analytics Project Plan

## B.1 Goals, Objectives, and Deliverables

* Goal: Develop an accurate predictive model for real estate prices in the USA
  + Objective 1: Obtain and preprocess a dataset of real estate listings in the USA
    - Deliverable 1: A cleaned and structured real estate dataset in a suitable format for analysis such as CSV.
  + Objective 2: Create and tune a predictive model using machine learning.
    - Deliverable 2: Fully trained and developed machine learning model capable of predicting real estate prices with high accuracy.
  + Objective 3: Create visual representations to interpret the model’s predictions and key factors used for determining real estate price.
    - Deliverable 3: A series of charts and graphs that display the importance of each feature and provide further data insight.

## B.2 Scope of Project

### B.2.A Included in Project Scope

Prepare the USA Real Estate Dataset from Kaggle for analysis and machine learning use. Select an appropriate model that is suitable for the problem and available data. Create visualizations that provide insight into the model’s performance and feature importance.

### B.2.B Not included in Project Scope

The data utilized in the analysis will not be directly collected by me or in real time. Recommendations for investment strategies or market interventions based on the model’s predictions will not be made.

## B.3 Standard Methodology

**CRISP-DM**

1. Business Understanding: What are the key factors influencing real estate prices in the USA and how can they be used to correctly predict prices?
2. Data Understanding: Review and explore the USA Real Estate Dataset to gain familiarity and identify data quality problems.
3. Data Preparation: Clean and prepare the data for analysis and model training.
4. Modeling: Select a suitable model for accurate prediction, feature insights, and the dataset in use. Calibrate the model’s parameters for optimal performance.
5. Evaluation: Evaluate the model’s performance using R2 and MSE to determine its capability of satisfying the project’s goal. Validate the model’s ability to generalize to new unseen data using cross-validation techniques.
6. Deployment: Present the model’s price prediction performance and identified key features.

## B.4 Timeline and Milestones

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

## B.5 Resources and Costs

1. Computer: N/A (Already owned)
2. Python: Free
3. Libraries (pandas, numpy, scikit-learn): Free
4. 80 work hours: 0$/hour

## B.6 Criteria for Success

1. Data Preparation: Perform necessary cleaning steps to prepare the data for model development. Including removal or imputation of missing values, removing irrelevant data, normalizing or removing outliers, and encoding categorical features as necessary. Successful completion of this step should have the data fully prepared for model training.
2. Model Development: Develop a machine learning model that predicts real estate prices based on the information available in the dataset. A model capable of predicting a price based on the features of properties will constitute success, regardless of the model’s performance measured by R2 and MSE.
3. Visualizations: Produce and utilize visualizations to aid data analysis, identify model’s performance, and understand key factors. Visualizations created throughout analysis to achieve the previously mentioned component will determine successful completion of this criteria.

# C. Design of Data Analytics Solution

## C.1 Hypothesis

## House features like property size, number of bedrooms and bathrooms, and location have a significant impact on real estate prices.

## C.2 and C.2.A Analytical Method

## The LightGBM algorithm, a gradient boosting framework that uses tree-based learning models will be used to support the hypothesis and research question. LightGBM is efficient in processing large datasets with many features. It is capable of capturing complex relationships between predictors and the target variable. The model will be trained on a subset of the data and evaluated against the remaining unseen data. The model’s predictive performance will be measured using R2 and MSE. Furthermore, LightGBM can provide feature importance scores which directly informs the significance of the hypothesized factors to influence real estate prices.

## C.3 Tools and Environments

* Python: Programming language used for data manipulation, analysis, and model development.
* Pandas: A Python library used for data manipulation, analysis, and ideal for handling and cleaning the dataset.
* NumPy: A Python library for numerical computations, which is incredibly efficient on large arrays.
* LightGBM: An open-source gradient boosting framework for Python provided by Microsoft, used for our predictive model.
* Scikit-learn: A Python library for machine learning that provides tools for data analysis.
* Matplotlib and Seaborn: Python libraries for creating visualizations.
* Visual Studio Code with Jupyter Extension: IDE that supports Python and interactive data analysis through Jupyter Notebooks.

## C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance

* Type of model: Supervised regression.
* Algorithm: LightGBM (Light Gradient Boosting Machine)
* Metrics Used for Assessment: R-squared (R2) and Mean Squared Error (MSE)
* Benchmark for Success: The model will be considered successful if the R2 value is ≥ 0.6, indicating that the model explains at least 60% of the variance in real estate prices. For MSE, the model should be less than 2,500,000,000.
* Justification: Supervised regression is appropriate due to the continuous nature of the response variable (real estate prices). LightGBM is chosen for its efficiency with large datasets and ability to handle a variety of features. R2 is an appropriate metric as it provides a measure of how well new unseen data will be predicted by the model. MSE is relevant as it measures the average squared difference between the estimated values and the actual value, this is vital to ensure the accuracy of the predictions in a real estate context.

## C.5 Practical Significance

Practical significance can be assessed if the model provides clear guidance on property valuations. The model should streamline the process of evaluating real estate prices, yielding faster decisions without significantly compromising accuracy. For example, a real estate firm may use the model to prioritize renovations that the model indicates has the highest impact on property values.

## C.6 Visual Communication

Two types of visualizations will be included to communicate the model’s findings. The visuals will be created using Matplotlib and Seaborn in Python.

1. Scatterplot of Residuals: This graph will show the residuals (differences between observed and predicted values) of the model. This is crucial to diagnosing and understanding the model’s performance. The visual will help determine if the model systematically under or overpredicts certain price points.
2. Barplot of Feature Performance: This chart will show the importance of each feature in the model, by gain and split. It will aid in understanding which features are most influential in determining price.

# D. Description of Dataset

## D.1 Source of Data

The dataset that will be used for this project is from Kaggle.

* USA Real Estate Dataset - Ahmed Shahriar Sakib
  + Real Estate listings (1.4M+) in the US broken by State and zip code
  + <https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data>

## D.2 Appropriateness of Dataset

The dataset is appropriate for the research question for several reasons:

1. Comprehensive Coverage: It includes properties across the USA and offers a diverse and representative sample.
2. Relevant Features: The dataset includes critical variables such as location, number of bathrooms and bedrooms, and property size, which are hypothesized to influence real estate prices.
3. Structured Format: The data is in CSV format which is suitable for analysis using Python and LightGBM.

## D.3 Data Collection Methods

## The dataset was downloaded in CSV format from Kaggle at <https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset/data>.

## D.4 Observations on Quality and Completeness of Data

Quality:

* Consistency: The dataset displays a high degree of consistency in key variables such as bed, bath, acre\_lot, city, state with valid entries showing categorized and consistent data.
* Data Integrity: Variables don’t show any mismatched entries, indicating that the data types are consistent with their expected values for each column.

Completeness:

* Missing Values: Some columns, like bed, bath, and acre\_lot, exhibit missing data, with approximately 15% to 26% of records incomplete. The column prev\_sold\_date is missing in almost half of the dataset.
* Overall Dataset Health: Despite the missing values, most of the data is complete with critical price and feature information.

Accommodations:

* Given the dataset’s large number of entries (~1.4 million), excluding incomplete rows and outliers is unlikely to impact data integrity. Exclusion is preferable over imputation as filling critical features could skew the model’s ability to predict price accurately, and the dataset is likely to remain diverse and robust after exclusion.

## D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

1. Data Governance: Data Governance ensures our data integrity and managed responsibly. Use standardized policies on data access, usage, sharing, and handling to maintain data integrity. Ensure data changes are clearly recorded.
2. Privacy: The dataset contains information on properties, not individuals and no addresses so privacy concerns are minimized. As a result, no precautions are necessary.
3. Security: Safeguarding against data breaches is critical in all environments. The computer is password protected to prevent unauthorized access.
4. Ethical, Legal and Regulatory: Ethical use of data should avoid bias or misrepresentation. The dataset is from Kaggle and may have additional terms of use. Findings should be clearly communicated and neutral. Review and adhere to the terms of use from Kaggle and the dataset.

# References

Asaftei, G. M., Doshi, S., Means, J., & Sanghvi, A. (2018, October 8). *Getting ahead of the market: How big data is transforming real estate*. McKinsey & Company. <https://www.mckinsey.com/industries/real-estate/our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate>

*Factors that influence real estate prices*. (2023). Schwab Brokerage. <https://www.schwab.com/learn/story/factors-that-influence-real-estate-prices>

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, *2021*, 1–9. <https://doi.org/10.1155/2021/7678931>