**C964 WGU Capstone**

**Daniel Roberts**

**08/27/2023**

**Table of Contents**

1. **Letter of Transmittal**
2. **Executive Summary**

**B1. Opportunity**

**B2. Customers**

**B3. Gaps**

**B4. Data**

**B5. Methodology**

**B6. Deliverables**

**B7. Planning and Outcomes**

**B8. Validation**

**B9. Programming and Costs**

**B10. Timeline**

1. **Data Product**
2. **Project Documentation**

**D1. Vision**

**D2. Data Sets**

**D3. Code for Data**

**D4. Assessment of Hypothesis**

**D5. Visualizations**

**D6. Accuracy**

**D7. Results**

**D8. Source Code**

**D9. Quick Start Guide**

**D9. Summation of Learning Experience**

1. **Sources**

**E1. Sources that were Cited**

**A. Letter of Transmittal**

Daniel B. Roberts

1130 Enterprise Dr.

Los Angelos, California 89506

Drob839@wgu.edu

(555) 555-0415

August 17, 2023

Robert Y. Daniels

CEO

Property Inc.

Robert Daniels,

I am pleased to present to you an extensive project proposal for the development and implementation of a pricing predictor application for homes, which utilizes advanced data analysis techniques to estimate housing prices based on multiple property features. This innovative data product provides valuable insights and support for informed decision-making within the real estate industry.

The challenge in the real estate industry lies in accurately determining property values. Traditional methods often lack precision and fail to consider the subtle relationships between various property features. The Housing Pricing Predictor addresses this issue by leveraging data-driven analysis to generate accurate and reliable housing price estimates.

The Housing Pricing Predictor offers tangible benefits to both real estate professionals and potential buyers. By harnessing the power of multivariable linear regression and K-means clustering, this data product provides a sophisticated tool for estimating property values. Real estate professionals can make more informed investment decisions, while potential buyers can gain insights into the importance of prospective properties. The data product empowers users to navigate the real estate market confidently.

The Housing Pricing Predictor is a software application that will integrate data analytics and statistical techniques. It takes input features related to a property and produces an estimated price as output. The user-friendly interface allows for seamless interaction and intuitive access to valuable insights.

The data used for this project was collected from Kaggle.com. The data used to construct the Housing Pricing Predictor is sourced from reliable and reputable sources. It includes a wide range of property features such as location, size, condition, and amenities. This comprehensive data set ensures accurate predictions.

The primary objective of this project is to create a predictive model that accurately estimates housing prices. The hypotheses include the assumption that specific property features significantly influence the overall price and that K-means clustering can reveal distinct property clusters based on shared attributes.

The Housing Pricing Predictor methodology involves collecting, preprocessing, and cleaning housing data. K-means clustering is applied to identify inherent property relationships, followed by the development of a multivariable linear regression model. The model is trained in historical housing data and evaluated using established metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The proposed project requires resources for data acquisition, software development, and model evaluation. An estimated budget of $5,000 will cover necessary expenses, including data procurement, software development tools, and labor hours.

Impact on Stakeholders:

The Housing Pricing Predictor can potentially revolutionize the real estate industry by enhancing decision-making for both professionals and buyers. Real estate agencies can benefit from more accurate property assessments, while buyers can confidently navigate the market landscape. The impact of the Housing Pricing Predictor on stakeholders within the real estate industry is significant and far-reaching, promising to bring about transformative changes that enhance decision-making processes and empower various parties involved. The potential outcomes of the predictor's implementation extend to both real estate professionals and potential property buyers, creating a positive effect throughout the entire industry. A few examples of those effects include:

• Price properties more competitively and strategically.

• Optimize the listing process by setting appropriate initial prices.

• Identify undervalued or overvalued properties and adjust their marketing strategies accordingly.

• Enhance negotiations with buyers and sellers by providing transparent and evidence-backed pricing information.

We recognize the importance of handling sensitive data ethically and responsibly. The project team will adhere to data privacy regulations and ensure transparent communication regarding data usage and processing. Some key factors include but are not limited to:

• Bias and Fairness: We will ascertain any unfair biases during our data cleansing stage regarding factors such as race, ethnicity, or neighborhood.

• Transparency: Potential home buyers have a right to know just how our predictions are made. We will provide clear documentation outlying our application’s features.

• Reliability and accuracy: We will ensure our application is regularly maintained and updated to ensure accurate and reliable results.

My team and I have expertise in data analytics, statistical modeling, and software development. My proficiency in Python programming and experience with libraries such as NumPy, ` Pandas, and Matplotlib ensure the successful implementation of the Housing Pricing Predictor.

To conclude, the Housing Pricing Predictor offers a revolutionary solution for addressing the challenges of property valuation. This project proposal outlines an approach to creating a powerful and insightful data product. I am confident that this product would greatly benefit Property Inc. and contribute to more informed decision-making within the real estate industry.

Thank you for considering this proposal. I look forward to discussing this project further and collaborating on its successful implementation.

Sincerely,

Daniel B. Roberts

**B. Executive Summary**

**B1. The decision-support problem or opportunity you are solving for:**

The real estate industry is faced with a critical challenge - accurately estimating property values in a constantly evolving market. Traditional methods often lack the precision required for informed decision-making, leading to potential financial losses, and missed opportunities. Our solution aims to bridge this gap by introducing a data-driven housing price predictor that leverages advanced analytics to provide reliable and precise property value estimations. Homeowners and real-estate companies alike can have an error rate as high as 6%-8% of the value of the home.

**B2. A description of the customers and why this product will fulfill their needs:**

Our primary focus will be real estate professionals, agents, property investors, and potential buyers. These stakeholders require a dependable tool that can offer insights into property values, enabling them to make confident decisions in an uncertain market landscape. By having a dependable estimation tool, real estate professionals can enhance their credibility and strengthen their client relationships. Armed with accurate estimations, investors can identify properties with promising appreciation potential and make informed decisions about their investments. Potential buyers often face challenges in understanding whether a property's asking price aligns with its true market value. Our solution empowers buyers to make better evaluations, negotiate better terms, and make well-informed purchasing decisions. The real estate market is dynamic and subject to fluctuations influenced by various factors, including economic conditions, local trends, and property attributes. Our solution’s ability to offer precise property valuations helps mitigate the uncertainty inherent in the market.

**B3. Existing gaps in the data products you are replacing or modifying:**

From Zillow to Realtor.com, most major real estate sites offer a tool to estimate a property’s value. These existing products often fall short of accurately capturing the complex dynamics of real estate markets. Our data-driven approach seeks to overcome these limitations, offering a comprehensive solution that factors in a wide array of variables to provide a more accurate and reliable property value estimation. Our innovative approach leverages the power of machine learning techniques to revolutionize property valuation. Unlike conventional methods that rely on a limited set of variables, our data-driven solution taps into a vast array of relevant factors that influence property values. By incorporating variables such as location, property size, condition, neighborhood trends, and local economic indicators, our approach paints a more complete picture of a property's true worth.

**B4. The data available or the data that needs to be collected to support the data product lifecycle:**

To support the data product lifecycle and ensure the accuracy and reliability of our housing pricing predictor, we require a diverse dataset that captures the complex factors influencing property values. This dataset will serve as the foundation for training our machine learning model, allowing it to learn and adapt to the changing aspects of real estate markets. We will need data such as property characteristics such as size, number of bedrooms, bathrooms, and with or without a basement. Location aspects can greatly impact a property’s value such as access to water views, distance to schools, etc. Historical sales records and a property’s current condition can also play a major role in the overall value of the home. These sorts of data sets can be constructed or pulled from a source such as Kaggle.com. For our initial model, we will use an existing data set from Kaggle.com.

**B5. The methodology you use to guide and support the data product design and development:**

Our project will adhere to the Iterative Development Methodology, which aligns flawlessly with the iterative nature of predictive model creation. Our decision to adopt the Iterative Development Methodology stems from its proven track record in similar projects within the field. Notably, this methodology promotes a continuous cycle of improvements, enabling us to fine-tune the model's accuracy as new insights emerge. To execute this methodology effectively, we will adhere to the following workflow process:

* Initial Problem Definition and Scope
* Data Collection and Exploration
* Feature Engineering and Preprocessing
* Initial Model Development
* Model Evaluation and Feedback
* Iteration and Refinement
* Validation and Verification
* Interpretability and Visualization
* Documentation and Reporting
* Deployment and Monitoring

By supporting an Iterative Development Methodology, we ensure that the predictive housing pricing model evolves alongside the real estate market. Our feedback loop, flexibility, and commitment to improvement will result in a data product that meets the evolving needs of our stakeholders while maintaining a high level of accuracy.

**B6. Deliverables associated with the design and development of the data product:**

Our main deliverable will be the finished application model of the Housing Pricing Predictor. Documentation for the predictive model as well as documentation for our clean and raw data sets, data visualizations, and results from testing.

**B7. The plan for implementation of your data product, including the anticipated outcomes from this development:**

The plan for the implementation of the housing pricing predictor data product involves a structured approach to deploying the developed solution, ensuring continuous integration, user accessibility, and ongoing monitoring. The outcomes of the development encompass a range of benefits that align with the project's goals and objectives. The implementation plan and anticipated outcomes are as follows:

Model Deployment: The predictive model, developed using multivariable linear regression, will be deployed in a user-friendly environment. A web-based dashboard will eventually be created to allow users to input property features and receive estimated property values.

Integration: Integrate the model with Property Inc.’s existing real estate platforms and websites that stakeholders frequently use, ensuring easy access.

User Training: Provide training materials and guides to help stakeholders navigate the application effectively and understand the interpretability of the predictions.

Testing and Quality Assurance: Conduct thorough testing to identify and resolve any issues related to model accuracy, dashboard functionality, and user experience.

Security Measures: Implement security measures to protect sensitive user information and ensure data privacy throughout the application's usage. Our plan will be to implement authentication measures, data encryption, input validation measures, monitoring and logging practices, continuous security patches, and disaster and recovery plans.

Scalability Considerations: Design the solution to be scalable, allowing for potential future enhancements, such as incorporating additional features, integrating with more datasets, or expanding to new geographical regions.

Enhanced Decision-Making: Real estate professionals, agents, and investors will have a reliable tool to assist in making informed decisions about property investments, pricing, and market analysis.

Improved Property Assessments: Potential buyers will benefit from accurate property value estimations, aiding them in evaluating properties and negotiating effectively.

Confident Market Navigation: Users will be able to navigate the real estate market with confidence, backed by data-driven insights.

Time and Cost Savings: The application's efficiency will save time for real estate professionals by streamlining property valuation processes and minimizing manual estimates.

Business Growth Opportunities: Real estate agencies can leverage the tool to attract potential clients, showcase their expertise, and stand out in a competitive market.

Data-Driven Insights: The solution will generate insights into feature importance and relationships, offering stakeholders a deeper understanding of factors influencing property values.

Continuous Improvement: User feedback will provide opportunities for continuous improvement and refinement of the application, ensuring its long-term relevance.

**B8. The methods for validating and verifying that the developed data product meets the requirements and subsequently the needs of the customers:**

Our team will initialize a 3-step validation process:

* Testing: This will include unit testing on the underlying code of the application, integration testing to verify the data and application integrates smoothly, performance-based testing to ensure our application can respond quickly during peak usage, and security testing to identify any vulnerabilities and ensure data is properly protected.
* Accuracy Assessment: We will compare our data model to the actual market prices to measure the accuracy of our predictive application.
* Feedback Collection: We will receive feedback from Property Inc. as well as conduct continuous surveys of users on satisfaction. This will ensure our product meets the needs and requirements of stakeholders.

**B9. The programming environments and any related costs, as well as the human resources that are necessary to execute each phase in the development of the data product:**

Data Collection and Preprocessing:

Programming Environment: Jupyter Notebook with Python

Human Resources: Data Analyst

Cost: $1,000

Initial Data Exploration:

Programming Environment: Jupyter Notebook with Python

Human Resources: Data Analyst

Cost: $1,000

K-means Clustering Implementation:

Programming Environment: Jupyter Notebook with Python

Human Resources: Data Scientist

Cost: $2,000

Multivariable Linear Regression Development:

Programming Environment: Jupyter Notebook with Python

Human Resources: Data Scientist

Cost: $2,000

Model Training and Evaluation:

Programming Environment: Jupyter Notebook with Python

Human Resources: Data Scientist

Cost: $1,000

Visualization Creation:

Programming Environment: Jupyter Notebook with Python, Data Visualization Tools (Matplotlib, Seaborn)

Human Resources: Data Analyst

Cost: $1,000

Testing:

Programming Environment: Jupyter Notebook with Python

Human Resources: Real Estate Users, Data Analyst

Cost: $1,000

Deployment and Implementation:

Programming Environment: Jupyter Notebook with Python

Human Resources: DevOps Team

Cost: $500

User Training and Support:

Programming Environment: Jupyter Notebook with Python

Human Resources: Project Team

Cost: $500

**B10. A projected timeline, including milestones, start and end dates, duration for each milestone, dependencies, and resources assigned to each task:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Event** | **Start** | **End** | **Hours** | **Dependencies** | **Resources** |
| Kickoff | 08/20/23 | 08/20/23 | 2 | None | Project Team |
| Data Collection and Preprocessing | 08/21/23 | 08/25/23 | 20 | Kickoff | Data Analyst |
| Data Exploration | 08/26/23 | 08/28/23 | 16 | Data Collection and Preprocessing | Data Analyst |
| K-Means Clustering Implementation | 08/29/23 | 08/30/23 | 8 | Data Exploration | Data Scientist |
| Multivariable Linear Regression Development | 08/31/23 | 09/02/23 | 8 | Data Exploration | Data Scientist |
| Model Training and Evaluation | 09/03/23 | 09/06/23 | 20 | K-Means & ML Regression Implementation | Data Scientist |
| Visualization | 09/07/23 | 09/08/23 | 8 | Model Training | Data Analyst |
| Testing | 09/09/23 | 09/12/23 | 30 | Visualization | Property Inc. Team & Project Team |
| Deployment | 09/13/23 | 09/15/23 | 16 | Testing | DevOps Member |
| Training | 09/16/23 | 09/17/23 | 8 | Deployment | Project Team |

**C. Data Product**

See the attached Jupyter Notebook file.

**D. Project Documentation**

**D1. Business Requirements**

Project Name: Housing Pricing Predictor

Version: 1.0

Date: August 20, 2023

The business requirements for this project have been outlined through other documentation already provided. Below is a synopsis of those requirements.

Business Vision:

Through data-driven insights and advanced predictive modeling, we aim to provide accurate property value estimates that transform decision-making processes.

Business Objectives:

* Deliver accurate property value estimates.
* Enhance decision-making for Real Estate Professionals.
* Empower buyers with reliable information.

Business Goals:

* Develop an advanced housing pricing predictor.
* Provide a user-friendly interface.
* Support continuous monitoring and enhancements.

Functional Requirements:

* Data
* Predictive model – K-Means Clustering & Multilinear Regression
* Visualization

Non-Functional Requirements:

* Performance testing measures
* Security measures
* Usability and user feedback measures

Stakeholder Roles and Responsibilities:

* Real Estate professionals with Property Inc.
* Property Investors
* Potential Buyers

Potential Risks:

* Data security
* Model inaccuracy

Implementation Timeline:

* Documented in the Executive Summary.

Budget:

$10,000

Project Sign-off:

* Completed by Property Inc. management team.

**D2. Data Cleansing**

**Raw Data:** See attachment – “CapstoneData.csv”

**Cleansed Data:** See attachment – “cleaned\_data.csv”

**Test Data for predictions –** See attachment – “test\_data\_with\_predictions.csv”

**Cleansing Data Code:**

Raw = Raw.dropna()

Raw = Raw.drop(['sqft\_living15', 'sqft\_lot15', 'lat', 'long'], axis=1)

Raw['id'] = range(1, len(Raw) + 1)

Raw['date'] = pd.to\_datetime(Raw['date'])

data = Raw.iloc[:100]

cleaned\_data\_path = 'cleaned\_data.csv'

data.to\_csv(cleaned\_data\_path, index=False)

**D3. Analysis of Data**

**Linear Regression:**

features = ['id', 'bedrooms', 'bathrooms', 'sqft\_living', 'sqft\_lot', 'floors', 'waterfront', 'view', 'condition', 'grade', 'sqft\_above', 'sqft\_basement', 'yr\_built', 'yr\_renovated', 'zipcode']

X = data[features]

y = data['price']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

model = LinearRegression()

model.fit(X\_train\_scaled, y\_train)

y\_pred = model.predict(X\_test\_scaled)

mse = mean\_squared\_error(y\_test, y\_pred)

X\_test['Predicted\_Price'] = y\_pred

X\_test.to\_csv('test\_data\_with\_predictions.csv', index=False)

D3.

K-Means Clustering:

scaler2 = StandardScaler()

X\_scaled = scaler2.fit\_transform(X)

num\_clusters = 3

kmeans = KMeans(n\_clusters=num\_clusters, n\_init=10, random\_state=42)

data['cluster'] = kmeans.fit\_predict(X\_scaled)

D4. Assessment of Hypothesis

D5. Visualizations

**The top 3 features that correlate with the pricing of the home are the square footage, grade, and number of bathrooms each home has.**

**A diagram of heatmap

Description automatically generated**

**Most of the homes in our sample set last sold between the prices of $250k and $750k.**

**A green and black graph

Description automatically generated**

**The year the home was built does not necessarily dictate a positive reaction to the price of the home. Our highest-priced home in our sample data was built in 1968 and last sold for $2 million.**

**A graph of scatters of house prices

Description automatically generated**

**D6. Product’s Accuracy**

When discussing the accuracy of our product, it is important to consider that we are currently operating in version 1.0. The complex nature of property valuations is influenced by several factors. Variables such as the locality, neighboring residences, amenities like pools and fireplaces, and the extent of property enhancements can all have a significant influence on a property's value.

For our initial product launch, we intentionally adopted a simplified approach to our dataset and testing features. This simplicity allowed us to lay a foundation for accurate predictions while acknowledging the potential for variations in results. Some of our predictions closely aligned with the most recent sales prices, highlighting the model's capability, while others displayed inequalities.

To reach a more precise reflection of a property's potential value, a thorough and detailed dataset concerning each property is essential. As we continue to enhance our product, we remain committed to incorporating a deeper dataset that captures the complexities of each property, thus elevating the accuracy of our predictions.

**D7. Results**

**The image shows a scatter plot with points representing our test properties. The horizontal axis measures the living area size, while the vertical axis shows the property prices. Colors indicate different property clusters, and hovering over points reveals property IDs. This test data demonstrates a general trend: as square footage increases, property prices tend to rise.**

**A graph showing different colored dots

Description automatically generated**

**This resulting graph demonstrates the relationship between square footage and property prices. It displays two components: individual data points representing actual property prices against square footage, and a predictive line showing the model's forecasted price trends based on square footage. As mentioned earlier, certain predictions from our model closely align with the property's recent sale price, whereas others exhibit notable differences. To validate our model's predictions effectively, a comprehensive analysis of the home's features through deeper data exploration would be essential.**

**A graph with lines and dots

Description automatically generated**

**We incorporated a query code that reads data from our test CSV file. It enables interactive exploration by presenting a dropdown menu to select a property ID of the data that was tested. Upon selecting an ID and clicking the "Execute" button, the code retrieves and displays information about the chosen property, including its associated features, actual price, and predicted price. This tool improves the user's ability to research property details and predictions effectively.**

**A screenshot of a computer

Description automatically generated**

**D8. Source Code**

See the attached Capstone Project Jupyter Notebook file for all source code.

**D9. Quick Start Guide**

This guide will help you get started quickly and efficiently. Follow these steps to install and use the product:

System Requirements: Make sure your system meets the following requirements:

Operating System: Windows 10 or 11

Python: Version 3.7 or higher

Jupyter Notebook: Installed and functional

Download the Application: Download the application files from the provided source.

Install Dependencies: Open your terminal or command prompt and navigate to the application directory. Run the following command to install the required libraries:

pip install numpy pandas matplotlib seaborn scikit-learn

Usage Steps:

Open Jupyter Notebook: Launch Jupyter Notebook on your system.

Navigate to the Application: Use Jupyter Notebook to navigate to the directory where you downloaded the Housing Pricing Predictor files.

Open the Notebook: Open the Jupyter Notebook file named "Capstone Project.ipynb."

Execute the Cells: Execute each cell in the notebook by clicking on it and pressing Shift + Enter. Follow the instructions within the notebook to load the data and run the predictive model.

Interact with the Application: Once the model is loaded, you can input property features to get estimated housing prices. Follow the prompts and input the required data.

View Results: The application will display the estimated housing price based on the input features.

Note: This is a basic guide. Refer to the documentation for detailed explanations and usage.

**D9. Summation of Learning Experience**

**Previous Experience:**

My previous academic experiences have provided me with a strong foundation that prepared me for the challenges of this project. I have acquired a strong understanding of data analysis, modeling, and programming languages such as Python through previous coursework. These newly acquired skills laid the framework for me to begin this project.

**Additional Learning and Resources:**

While my previous academic experiences equipped me with a strong foundation, I was able to recognize that I would need additional resources to be able to successfully complete this project. I utilized several UDEMY courses on Jupyter Notebook and Machine Learning to assist in my understanding of the project requirements. I also utilized several YouTube videos to help my understanding of K-means clustering and Multivariable Linear Regression.

**Lifelong Learning:**

This project has helped to reinforce the importance of lifelong learning in my personal and professional development. Technology is ever evolving and skills such as machine learning and data science are desperately needed. This project has helped me understand the value of self-directed learning and cultivating a growth mindset. With technology forever changing, those qualities and skills are essential.

To conclude, my previous experiences, supplemented learning resources, and the principles of lifelong learning have prepared me to take on any project and be successful.

**E1. Sources**

No external sources were quoted or used in this documentation.