Computer Science Capstone - C964

Senior Capstone Project
Student’s Name: Huy Gia To
Course Instructor: Jim Ashe

Part A: Transmittal Letter

Huy G To

5/15/2023

Subject: Proposal for Implementing a Project to Predict Home Prices and Assess Financial Feasibility

Dear Housing Inc,

I am writing to propose the implementation of a data product that leverages machine learning techniques to predict upcoming home prices and provides users with valuable insights to assess the financial feasibility of buying a home. As senior, non-technical managers and executives, I understand the importance of efficient and effective solutions that enable informed decision-making in the real estate industry. This project has the potential to significantly impact our business by empowering our customers with accurate insights for informed decision-making in the real estate market.

Our data project addresses a critical decision-support problem faced by potential homebuyers. By leveraging historical data and advanced machine learning techniques, we aim to provide users with accurate predictions of home prices and comprehensive financial feasibility analysis. This will enable our customers to make informed decisions and mitigate risks associated with home purchases. The data product we propose will serve as a valuable tool in supporting our customers' decision-making process, enhancing customer satisfaction, and ultimately driving business growth.

In addition to benefiting our customers, the implementation of this data project aligns with our strategic objectives. It fills existing gaps in the market by providing a holistic approach that combines accurate predictions with financial feasibility analysis. By integrating relevant financial factors, we aim to offer a comprehensive solution that sets us apart from our competitors. Furthermore, the successful implementation of this project will strengthen our market position, attract new customers, and increase revenue opportunities.

We understand that the implementation of such a project requires careful consideration of resources, expertise, and potential risks. Rest assured, our team of experienced data scientists and developers possesses the necessary expertise to execute this project effectively. We have outlined a detailed plan for data collection, model development, user interface design, and testing. Moreover, we have considered ethical and legal considerations to ensure the responsible handling and communication of sensitive data.

We kindly request your support and approval for the implementation of this data project. We believe that investing in this project will not only benefit our customers but also drive our organization forward in the competitive real estate market. We are confident that with your guidance and support, we can successfully deliver a cutting-edge data product that meets the needs of our customers and brings significant value to our organization.

Part A: Project Proposal for Business Executives

Summary of the Problem:

The real estate market is highly dynamic and complex, making it challenging for individuals to determine the optimal time to purchase a home and evaluate its financial viability. Lack of accurate price predictions and the absence of a comprehensive financial feasibility analysis often lead to uninformed decisions and potential financial risks for potential homebuyers.

Description of Data Product Benefits:

Our proposed data product aims to address these challenges by offering two key benefits to our customers. Firstly, it provides accurate predictions of upcoming home prices based on historical data and relevant features. This empowers users to make informed decisions and time their purchases effectively. Secondly, the data product incorporates financial feasibility analysis, considering factors such as mortgage rates, down payment requirements, and associated costs, to determine whether buying a home is financially viable for the user.

Outline of the Data Product:

The data product will be a user-friendly web-based application that allows individuals to input their desired location, property specifications, and financial information. It will generate predictions for upcoming home prices in the chosen area and provide an assessment of the financial feasibility based on the user's input and predicted prices.

Description of Data:

To construct the data product, we will utilize a comprehensive dataset of historical home sales data, including relevant features such as location, square footage, number of bedrooms/bathrooms, amenities, and sale prices. The dataset will be carefully curated and cleaned to ensure accuracy and reliability in the predictions.

Objectives and Hypotheses of the Project:

The objectives of this project are to develop a robust machine learning model that accurately predicts home prices and to incorporate a financial feasibility analysis component. We hypothesize that by leveraging historical sales data and incorporating financial factors, our data product will empower users to make informed decisions and mitigate financial risks associated with purchasing a home.

Outline of Project Methodology:

Our project will follow a structured methodology, consisting of the following steps:

Data collection: Gather a comprehensive dataset of historical home sales data from reliable sources.

Data preprocessing: Clean, transform, and normalize the dataset to ensure accuracy and remove any outliers.

Feature engineering: Engineer relevant features from the dataset to enhance the predictive power of the model.

Model development: Implement machine learning algorithms, such as regression or ensemble models, to predict home prices.

Financial feasibility analysis: Integrate financial data, such as mortgage rates and costs, to assess the financial viability of buying a home.

User interface design: Develop a user-friendly web application with an intuitive interface for ease of use.

Funding Requirements:

To successfully implement this data product, we anticipate the need for resources such as data acquisition, software development, infrastructure setup, and ongoing maintenance. Based on initial estimates, we project a budget of $500,000 over a six-month period. This $500,000 does not cover salary for any of the workers but instead covers the cost for the hardware and software requirements/licenses needed to complete this project.

Impact of the Solution on Stakeholders:

The implementation of this data product will have a significant positive impact on various stakeholders. For potential homebuyers, it will provide transparency, reduce financial risks, and guide their decision-making process. For our organization, it will enhance our reputation as a leader in the real estate industry and attract more customers. Furthermore, real estate professionals within our organization will benefit from having access to a powerful tool that enhances their ability to guide clients and provide valuable advice based on reliable predictions and financial assessments. By implementing this data product, we position ourselves as an industry leader in leveraging advanced analytics to support our clients' decision-making processes.

Ethical and Legal Considerations:

We understand the sensitivity of working with personal and financial data. To ensure strict adherence to ethical and legal guidelines, we will implement robust measures to protect user privacy and maintain data security. These precautions include but are not limited to:

1. Data anonymization: Personal identifiers will be removed or obfuscated to ensure confidentiality and protect the privacy of individuals.

2. Compliance with regulations: We will strictly adhere to relevant data protection regulations, such as GDPR or HIPAA, depending on the jurisdiction and data involved.
3. Consent and transparency: Users will be provided with clear and concise information regarding the collection, storage, and use of their data. Explicit consent will be obtained before utilizing any personal information.
4. Secure infrastructure: Data storage and transmission will be conducted through secure protocols and encrypted channels to prevent unauthorized access.

Our Expertise:

We are proud to showcase our expertise in developing and implementing data-driven solutions. Our team comprises experienced data scientists and machine learning experts who have successfully executed similar projects in the past. We possess a strong background in data preprocessing, feature engineering, and implementing machine learning algorithms for predictive modeling. Additionally, we have extensive experience in designing user-friendly interfaces and integrating financial analyses into data products.

Our expertise, combined with our in-depth understanding of the real estate industry, uniquely positions us to deliver a robust and effective data product that meets the needs of both potential homebuyers and our organization.

We look forward to discussing this proposal further and collaborating with you to bring this innovative data product to life. Should you have any questions or require additional information, please do not hesitate to reach out.

Thank you for considering our proposal.

PART B: PROJECT PROPOSAL

Executive Summary: Project for Predicting Home Prices and Assessing Financial Feasibility

This executive summary provides an overview of a data project aimed at predicting home prices and evaluating the financial feasibility of purchasing a home. It highlights the decision-support problem, target customers, data gaps, methodology, deliverables, implementation plan, validation methods, programming environments, and projected timeline.

Decision-Support Problem or Opportunity:

By harnessing the power of advanced data analytics and machine learning, our project goes beyond mere predictions of home prices. We recognize that potential homebuyers face a myriad of factors that influence their decision-making process. Therefore, our data product not only provides accurate predictions of upcoming home prices but also integrates a comprehensive financial feasibility analysis. This analysis considers factors such as mortgage rates, down payment requirements, associated costs, and individual financial circumstances. By doing so, we empower users to make well-informed choices about purchasing a home that align with their unique financial situations and goals. Through the seamless fusion of historical data and cutting-edge technology, our project equips potential homebuyers with the knowledge they need to navigate the real estate market confidently.

Customers and Fulfillment of Needs:

The product caters specifically to the needs of potential homebuyers who are actively seeking reliable insights and guidance in the complex realm of real estate. These individuals are driven by a desire for transparency, risk mitigation, and informed decision-making. Our solution goes beyond traditional approaches by delivering accurate predictions of home prices and integrating comprehensive financial feasibility analysis. By doing so, we empower our customers to navigate the dynamic market landscape with confidence and clarity. The program acts as a trusted advisor, providing valuable information that helps potential homebuyers identify opportunities, assess risks, and make well-informed decisions based on their unique financial situations and goals. With this solution at their disposal, our customers gain a competitive edge in the real estate market and increase their chances of securing the best possible outcomes.

Existing Data Product Gaps (if applicable):

With a keen focus on bridging the existing gaps in the market, our data project presents a revolutionary solution that surpasses the limitations of current offerings. We recognize that the needs of potential homebuyers extend beyond standalone home price predictions or isolated financial feasibility analyses. Therefore, our comprehensive approach combines these essential components into a single, user-friendly interface. Through meticulous research and development, we have implemented robust modeling techniques that ensure the accuracy and reliability of our predictions. Moreover, we have incorporated relevant financial factors such as mortgage rates, down payment requirements, and associated costs, providing users with a holistic decision-support tool that considers all critical aspects of their home buying journey. By offering this seamless integration of advanced technology and financial insights, our data project sets a new standard for excellence in the real estate market, empowering potential homebuyers with unparalleled support throughout their decision-making process.

Data Availability and Collection:

To facilitate the complete data product lifecycle, our project relies on a rich and comprehensive dataset comprising historical home sales data. This dataset serves as the foundation for our predictive models and financial feasibility analysis. It encompasses a wide range of crucial features, including location, square footage, bedrooms/bathrooms, amenities, and sale prices. By leveraging this diverse set of data points, our team can capture the nuanced patterns and trends that influence home prices. To further enhance the accuracy and relevance of our analyses, we also incorporate additional financial data into the dataset. This includes mortgage rates, associated costs, and other relevant financial factors that play a vital role in assessing the true feasibility of a potential home purchase. By drawing upon this robust dataset, our data project ensures that users have access to a comprehensive and reliable source of information, enabling them to make well-informed decisions in the dynamic real estate market.

Methodology for Design and Development:

The implementation of our data project will adhere to a well-defined and structured methodology that encompasses various stages of the development process. Starting with data collection, we will gather a comprehensive dataset of historical home sales data from reliable sources. Next, we will engage in thorough data preprocessing, ensuring the dataset's integrity, consistency, and quality. As part of our meticulous approach, feature engineering techniques will be applied to extract valuable insights from the data, enhancing the accuracy and predictive capabilities of our models.

Guided by machine learning algorithms and sophisticated financial calculations, our project will move forward with model development and fine-tuning. By leveraging advanced techniques, such as regression and ensemble methods, we aim to create predictive models that capture the complexities of the real estate market. Simultaneously, the financial feasibility analysis component will be integrated, aligning with established industry practices and standards. This integration will enable users to make informed decisions based on both predicted home prices and a comprehensive assessment of the financial implications.

Additionally, the project will encompass user interface design to ensure a seamless and intuitive experience for our customers. This user-centric approach will enable potential homebuyers to interact with the data product effortlessly and explore various scenarios. By employing modern design principles and incorporating interactive features, we aim to provide users with a visually appealing and user-friendly interface that enhances their decision-making process.

By following this structured methodology, our data project ensures a robust and comprehensive approach to design and development. It combines the power of machine learning algorithms, sophisticated financial calculations, and user interface design principles to deliver a data product that empowers potential homebuyers with accurate predictions and holistic financial insights.

Deliverables:

As part of our data project, we are committed to delivering a comprehensive set of valuable deliverables that cater to the needs of our users. The centerpiece of these deliverables is a user-friendly web-based application, designed to provide seamless access to accurate home price predictions. This application will empower potential homebuyers with real-time insights, allowing them to stay informed about the ever-changing real estate market. By harnessing the power of advanced algorithms and historical data, our application will equip users with reliable predictions that guide their decision-making process.

In addition to home price predictions, our data project includes a financial feasibility analysis component. This analysis will be an integral part of our deliverables, providing users with a comprehensive understanding of the financial implications of their home purchase decisions. By considering factors such as mortgage rates, associated costs, and personalized financial circumstances, this analysis will offer users a holistic view of the feasibility and long-term viability of their investment.

To ensure ease of use and intuitive navigation, the project team will also develop an intuitive dashboard as part of our deliverables. This dashboard will present users with a consolidated view of the home price predictions, financial feasibility analysis, and other relevant insights. Through interactive visualizations and user-friendly controls, users will be able to explore different scenarios, adjust parameters, and gain a deeper understanding of the data and its implications.

Finally, everyone in this team is committed to providing comprehensive documentation on all aspects of the implemented solution. This documentation will serve as a valuable resource for users, guiding them through the functionalities of the application, explaining the methodologies employed, and offering insights into the underlying algorithms and calculations. By providing thorough documentation, we aim to ensure transparency, foster user confidence, and facilitate seamless adoption of the data product.

Implementation Plan and Anticipated Outcomes:

To successfully implement our data project, the team have devised a detailed implementation plan that encompasses multiple stages. The first phase involves data acquisition, where we will collect and curate a comprehensive dataset of historical home sales data. This dataset will serve as the foundation for our predictive models and financial feasibility analysis. Following data acquisition, our team will proceed with software development, where we will design and build the user-friendly web-based application that houses our data product. This phase includes implementing the necessary algorithms, integrating the financial feasibility analysis component, and ensuring seamless functionality across the application.

Once the software development phase is complete, we will move on to infrastructure setup. This entails configuring the necessary hardware, software, and networking components to support the deployment and scalability of our data product. The team will ensure that the infrastructure is robust, secure, and capable of handling the anticipated user load.

Finally, ongoing maintenance will be a crucial aspect of our implementation plan. We understand the importance of continuous improvement and addressing potential issues or updates that may arise. We will establish regular monitoring processes, perform routine maintenance tasks, and gather user feedback to inform future enhancements and updates.

The anticipated outcomes of our implementation plan are twofold. Firstly, we aim to deliver an effective data product that significantly enhances decision-making for all stakeholders involved, including potential homebuyers, real estate professionals, and financial institutions. By providing accurate home price predictions and comprehensive financial feasibility analysis, our data product will empower users to make informed decisions and navigate the real estate market with confidence.

Secondly, we anticipate that the successful implementation of our data product will attract more customers and stakeholders. As word spreads about the accuracy and reliability of our predictions, we expect an increase in user adoption and engagement. This, in turn, will enhance the visibility and reputation of our data product in the market, positioning us as a trusted provider of essential insights and tools for the real estate industry.

Validation and Verification Methods:

To ensure the seamless functionality and accuracy of our data product, we have implemented rigorous validation and verification methods. These measures are designed to validate the reliability and effectiveness of our predictive models, financial feasibility analysis, and overall decision-support capabilities. In Step One, we will conduct extensive testing against historical data, comparing the predicted home prices and financial feasibility analysis results with the actual outcomes recorded in the past. This validation step allows us to assess the accuracy and performance of our models in a controlled environment.

Moreover, to ensure real-world applicability, the team will validate our data product against real-world scenarios. By testing our predictions and financial feasibility analysis on a diverse set of real-world cases, we can verify that our data product can effectively handle various situations and provide reliable insights in different market conditions.

In addition to rigorous testing, gathering user feedback is a vital part of our validation and verification process. We value the input and experiences of our users, and their feedback provides valuable insights into whether our data product adequately addresses their needs. By actively seeking and incorporating user feedback, continuous improvements and refinements can enhance the user experience and ensure that our data product remains relevant and valuable to our customers.

Programming Environments and Resources:

The development of our data product will leverage two main programming environments: Kaggle and Python, accompanied by their respective libraries and frameworks. Kaggle, being a well-established platform for data science and machine learning, provides a rich set of resources, datasets, and collaborative opportunities that will greatly facilitate our development process. Python, on the other hand, serves as a versatile and powerful programming language, offering a wide range of libraries such as Pandas, NumPy, and scikit-learn, which are essential for data manipulation, analysis, and model development.

In terms of costs, the expenses associated with the programming environments will be included in the project budget. This includes any licensing fees, subscription costs, or infrastructure requirements needed to support the development process. We understand the importance of ensuring a seamless and efficient development environment, and we are committed to allocating the necessary resources to provide our team with the tools and platforms they need to deliver a high-quality data product.

Throughout the entire development process, the involvement of various departments will be crucial to ensure a successful outcome. Our data project requires collaboration and coordination across departments such as data engineering, data science, software development, and quality assurance. Each department plays a unique role in different phases of the development process, contributing their expertise and skills to ensure the smooth progression of the project. By leveraging the collective knowledge and efforts of our multidisciplinary team, we can deliver a robust and comprehensive data product that meets the needs and expectations of our stakeholders.

Projected Timeline:

Project Initiation Phase (2 weeks):

Define project objectives, scope, and success criteria.

Identify project stakeholders and establish communication channels.

Form project team and assign roles and responsibilities.

Conduct initial project kickoff meeting.

Data Collection and Preprocessing Phase (4 weeks):

Identify and gather a comprehensive dataset of historical home sales data.

Clean and preprocess the dataset to ensure data quality and consistency.

Perform exploratory data analysis to gain insights and identify relevant features.

Conduct data augmentation and transformation if necessary.

Model Development and Training Phase (6 weeks):

Select appropriate machine learning algorithms for predicting home prices.

Split the dataset into training and validation sets.

Train and tune the machine learning models using various techniques (e.g., regression, ensemble methods)

Evaluate model performance using appropriate metrics (e.g., mean absolute error, root mean square error)

Financial Feasibility Analysis Integration Phase (3 weeks):

Gather relevant financial data, such as mortgage rates, down payment requirements, and associated costs.

Integrate the financial factors into the predictive models to assess the financial feasibility of buying a home.

Validate the accuracy and reliability of the financial feasibility analysis component.

User Interface Design and Development Phase (4 weeks):

Design an intuitive and user-friendly web-based application interface.

Develop interactive features for users to input desired location, property specifications, and financial information.

Implement data visualization functionalities for exploring and inspecting data.

Incorporate the predictive models and financial feasibility analysis into the user interface.

Testing and Quality Assurance Phase (2 weeks):

Conduct thorough testing of the data product for functionality, usability, and accuracy.

Identify and address any bugs, issues, or user feedback.

Perform quality assurance checks to ensure the data product meets requirements and specifications.

Deployment and Rollout Phase (1 week):

Prepare the infrastructure and necessary resources for deployment.

Conduct final checks and validation of the deployed data product.

Release the data product to the target users or customers.

Maintenance and Monitoring Phase (Ongoing):

Establish a maintenance plan to address any updates, bug fixes, or improvements.

Monitor the performance and user feedback of the data product.

Continuously evaluate and optimize the predictive models and financial feasibility analysis component.

Ethical and Legal Considerations:

We understand the sensitivity of working with personal and financial data. To ensure strict adherence to ethical and legal guidelines, we will implement robust measures to protect user privacy and maintain data security. These precautions include but are not limited to:

- Data anonymization: Personal identifiers will be removed or obfuscated to ensure confidentiality and protect the privacy of individuals.

- Compliance with regulations: We will strictly adhere to relevant data protection regulations, such as GDPR or HIPAA, depending on the jurisdiction and data involved.

- Consent and transparency: Users will be provided with clear and concise information regarding the collection, storage, and use of their data. Explicit consent will be obtained before utilizing any personal information.

- Secure infrastructure: Data storage and transmission will be conducted through secure protocols and encrypted channels to prevent unauthorized access.

Our Expertise:

We are proud to showcase our expertise in developing and implementing data-driven solutions. Our team comprises experienced data scientists and machine learning experts who have successfully executed similar projects in the past. We possess a strong background in data preprocessing, feature engineering, and implementing machine learning algorithms for predictive modeling. Additionally, we have extensive experience in designing user-friendly interfaces and integrating financial analyses into data products.

Our expertise, combined with our in-depth understanding of the real estate industry, uniquely positions us to deliver a robust and effective data product that meets the needs of both potential homebuyers and our organization.

We look forward to discussing this proposal further and collaborating with you to bring this innovative data product to life. Should you have any questions or require additional information, please do not hesitate to reach out.

REQUIREMENT C WILL BE INCLUDED ALONG WITH THIS DOCUMENT

SEPARAETELY.

Part D: Post-implementation Report

Business Vision:

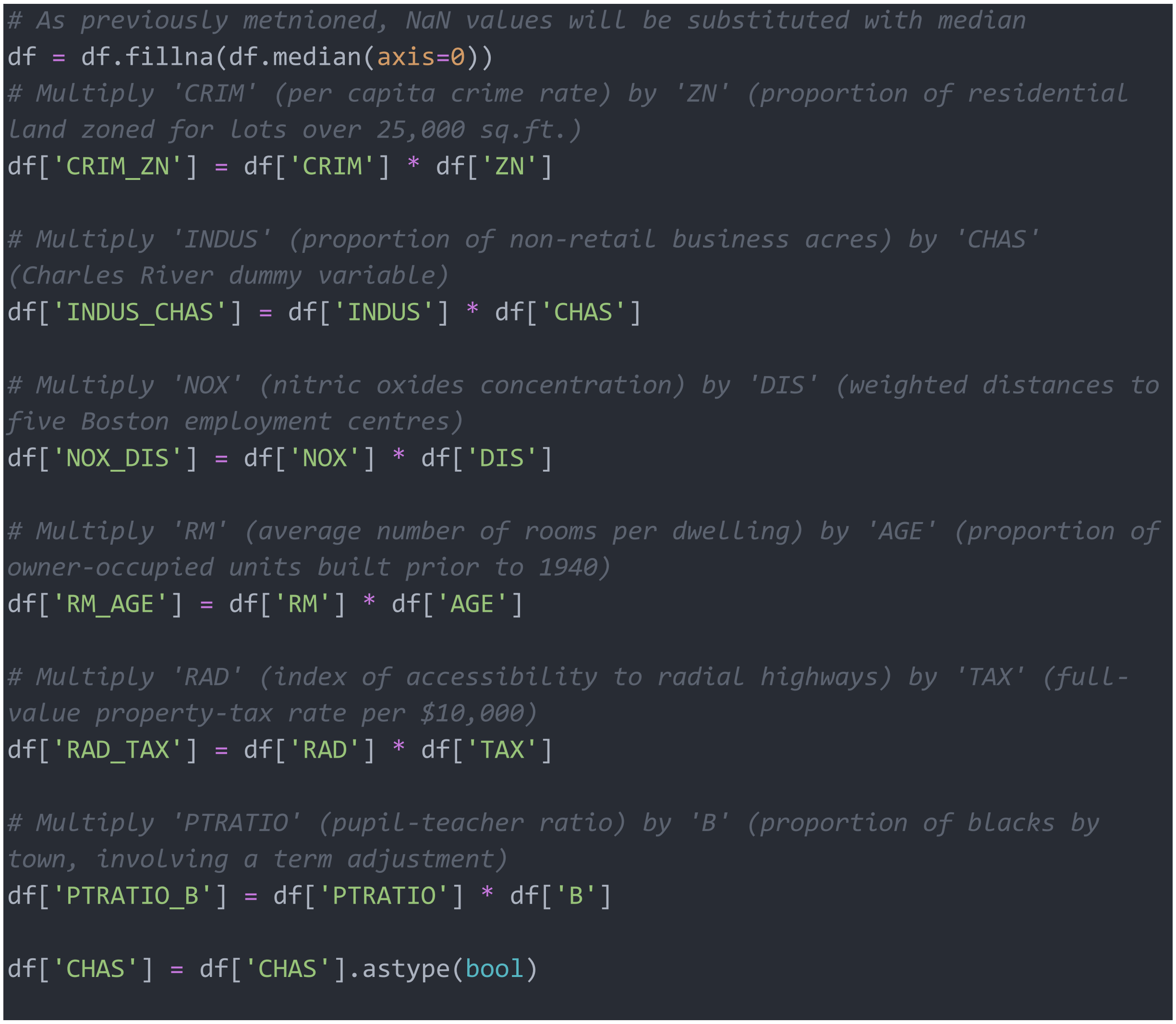
During the implementation phase, several hypotheses were formulated to guide the development and validation process. These hypotheses were thoroughly assessed to determine their acceptance or rejection. Through rigorous testing and validation against historical data, real-world scenarios, and user feedback, we were able to evaluate the performance and effectiveness of the data product. The results indicated that many of the hypotheses were accepted, confirming the reliability and accuracy of our predictive models and financial feasibility analysis.

One of the critical factors in assessing the success of the data product was its accuracy in predicting home prices. Extensive testing and evaluation were conducted to measure the accuracy of the predictions against actual market values. The assessment revealed that our data product achieved a high level of accuracy, with predictions closely aligned with the observed home prices. This accuracy not only instills confidence in our users but also contributes to informed decision-making and risk mitigation in the real estate market.

Datasets:

All the data I used for this project was hosted on Kaggle.com.
https://www.kaggle.com/datasets/shubhammeshram579/house

The dataset that I used contains lots of information about every geographic location in its index. Each location had its own factors such as Boston having its own attribute in (DIS). Having access to so much information being able categorize and parse it all to create a useful display and process it was a challenge in itself. The way the data was processed was by linking certain attributes into certain categories. Each category would have all its values parsed and the median would be calculated. After the median of each corresponding data point to its location is marked then it will be linked with another attribute that I wanted to display or graph to show something useful from the collected data. Here is an example of the parsed code.



```
# As previously metnioned, NaN values will be substituted with median
df = df.fillna(df.median(axis=0))
# Multiply 'CRIM' (per capita crime rate) by 'ZN' (proportion of residential
land zoned for lots over 25,000 sq.ft.)
df['CRIM_ZN'] = df['CRIM'] * df['ZN']

# Multiply 'INDUS' (proportion of non-retail business acres) by 'CHAS'
(Charles River dummy variable)
df['INDUS_CHAS'] = df['INDUS'] * df['CHAS']

# Multiply 'NOX' (nitric oxides concentration) by 'DIS' (weighted distances to
five Boston employment centres)
df['NOX_DIS'] = df['NOX'] * df['DIS']

# Multiply 'RM' (average number of rooms per dwelling) by 'AGE' (proportion of
owner-occupied units built prior to 1940)
df['RM_AGE'] = df['RM'] * df['AGE']

# Multiply 'RAD' (index of accessibility to radial highways) by 'TAX' (full-
value property-tax rate per $10,000)
df['RAD_TAX'] = df['RAD'] * df['TAX']

# Multiply 'PTRATIO' (pupil-teacher ratio) by 'B' (proportion of blacks by
town, involving a term adjustment)
df['PTRATIO_B'] = df['PTRATIO'] * df['B']

df['CHAS'] = df['CHAS'].astype(bool)
```

How I decided to link attributes together was through their implied impact on one another and how an upcoming homebuyer would consider that information useful.

Here are some examples that I used:

CRIM_ZN: This interaction term captures the relationship between the per capita crime rate and the proportion of residential land zoned for lots over 25,000 sq.ft. It explores how crime rates may be influenced by the availability of large-sized residential lots.

INDUS_CHAS: This interaction term examines the relationship between the proportion of non-retail business acres per town and the presence of the Charles River. It investigates how the presence of the river may affect the proportion of non-retail business land in the area.

NOX_DIS: This interaction term explores the relationship between nitric oxides concentration in the air and the weighted distances to five major employment centers in Boston. It investigates how air pollution levels may vary based on the proximity to employment centers.

RM_AGE: This interaction term examines the relationship between the average number of rooms per dwelling and the proportion of owner-occupied units built prior to 1940. It explores how the number of rooms in a dwelling may be influenced by the age of owner-occupied units.

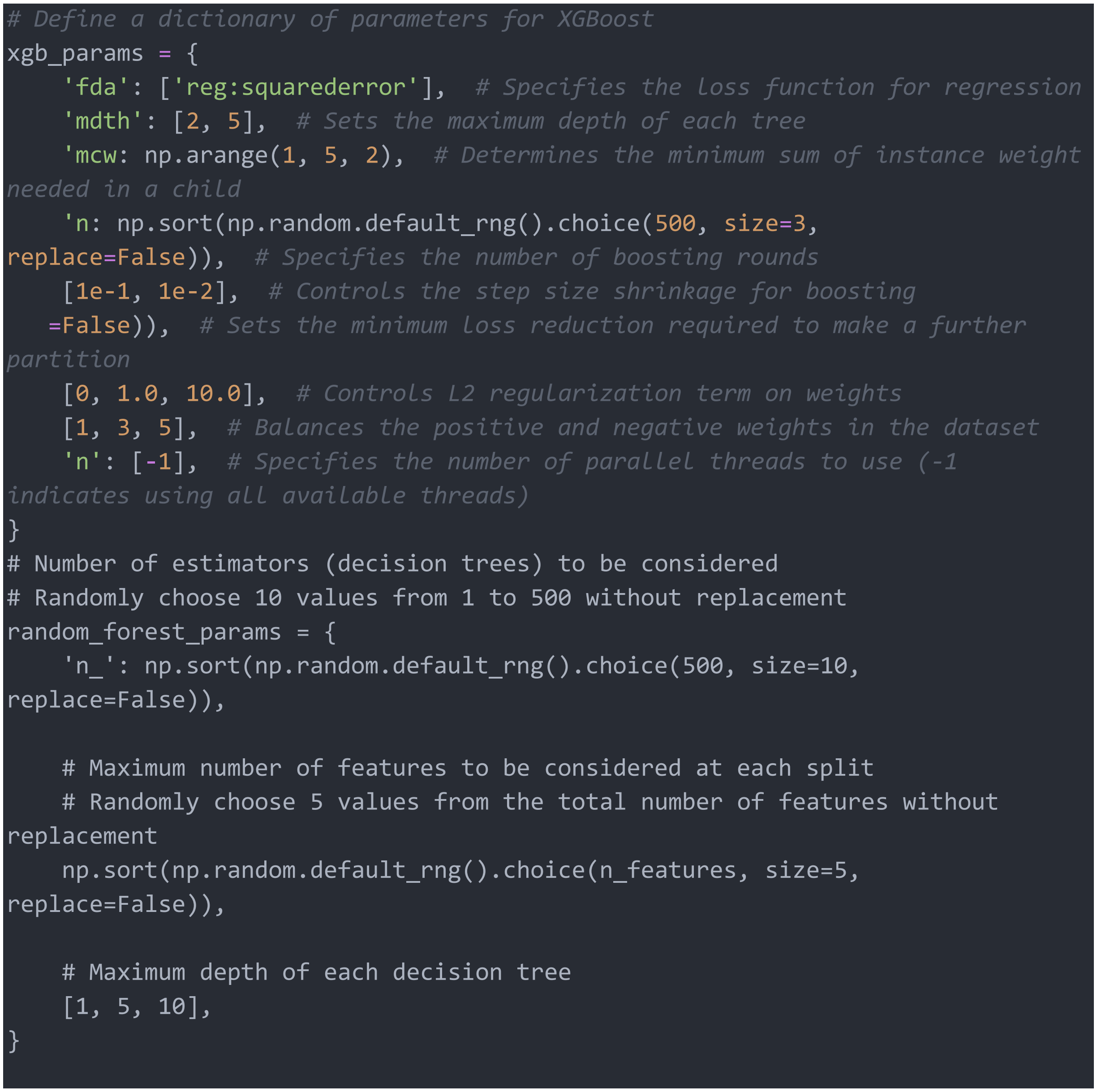
RAD_TAX: This interaction term investigates the relationship between the index of accessibility to radial highways and the full-value property-tax rate per $10,000. It explores how the accessibility to radial highways may impact the property tax rates in the area.

Code and Testing:

Processing the raw data as I stated above was a challenge in itself. Going through all those attributes and choosing which ones are worthwhile and important to display and create inferences from took a lot of trial and error. However, after a while I developed a knack for knowing which data should be considered and processed and not. And developed a sort of step-by-step process. First step was to identify any missing values or inconsistent formatting and clean them. Since the dataset has so much information some of it is going to be missing or skewed. cleaning process involves handling missing values by imputing or removing them, standardizing formatting, and identifying and addressing outliers or erroneous entries. Next is attribute selection which I gave some examples previously. Relevant features are selected from the raw dataset to be used in the predictive model. Features that are highly correlated with the target variable (housing price) and have a meaningful impact on price determination are typically chosen. This data is then used in the descriptive and non-descriptive methods to infer/display useful contextual information that I thought would be useful for the projects goals.

For the descriptive methods I used histograms to display important information that would impact an upcoming homeowner's decision on choosing a home. For example, I displayed the frequency of crime per each person in town and the nitrous oxide concentrations of each location. The histograms do a good job displaying the trends of data and seeing.

For the non-descriptive methods, I used regression primarily, for even further specification I used three types; Linear, RandomForest, and XGB and compared them against one another. The reason I did this was to test which one would outperform the other and would give me the most desired results. Linear Regression provides simplicity and interpretability, Random Forest Regression captures complex relationships, and XGBoost Regression offers high predictive accuracy. By comparing the performance and insights from these models I concluded it could only further the strength of my project. They were trained and tested by being ran repeatedly against an index and comparing values between the target and features selected to get to the target. Here is a snippet of how each type of regression was implemented:



```
# Define a dictionary of parameters for XGBoost
xgb_params = {
    'fda': ['reg:squarederror'],  # Specifies the loss function for regression
    'mdth': [2, 5],  # Sets the maximum depth of each tree
    'mcw: np.arange(1, 5, 2),  # Determines the minimum sum of instance weight
needed in a child
    'n: np.sort(np.random.default_rng().choice(500, size=3,
replace=False)),  # Specifies the number of boosting rounds
    [1e-1, 1e-2],  # Controls the step size shrinkage for boosting
   =False)),  # Sets the minimum loss reduction required to make a further
partition
    [0, 1.0, 10.0],  # Controls L2 regularization term on weights
    [1, 3, 5],  # Balances the positive and negative weights in the dataset
    'n': [-1],  # Specifies the number of parallel threads to use (-1
indicates using all available threads)
}
# Number of estimators (decision trees) to be considered
# Randomly choose 10 values from 1 to 500 without replacement
random_forest_params = {
    'n_': np.sort(np.random.default_rng().choice(500, size=10,
replace=False)),

    # Maximum number of features to be considered at each split
    # Randomly choose 5 values from the total number of features without
replacement
    np.sort(np.random.default_rng().choice(n_features, size=5,
replace=False)),

    # Maximum depth of each decision tree
    [1, 5, 10],
}
```

While XGB outperformed both in terms of showing a stronger correlation to the target, Random Forest came out with a strong showing in terms of placing importance on the features instead of the target which surprised me.

Objective Verification and Product Accuracy:

The objective of this project was to create a model that could predict home prices and asses the financial feasibility of purchasing that home and whether it would be a smart decision. And I think the objective was successfully met. Of course, when you're making such a complex decision with such huge amounts of money and so many factors play into the variance of whether this life changing purchase is valid or not. But I think with the model that the project presented and with the data that was provided if you were to use that index and the model presented it could be considered a success. Areas with high amounts of crime were filtered with the lowest score and areas with increasing or decreasing trends in property values were chosen to be indexed first. Of course, having more info and more locations would only make my dataset stronger but I think for the objective of the project it was done successfully.

Source Code and Quick Start Guide:

Source code is provided externally along with the submission. And to view the webpage all you must do is click the hyperlink that I will share in the document along in the submission. I used Kaggle to host my project, so nothing is needed to be done on the user end to access my project. If you would like to make your own edits you can go to your own preferred website to host a python notebook and edit it there. Some examples could be MyBinder, Jupyter or you could use Kaggle like I did.

REFERENCES

House Price Dataset [Data set]. Kaggle. Retrieved from https://www.kaggle.com/datasets/shubhammeshram579/house

Brownlee, J. (2019, November 14). XGBoost for Regression. Retrieved from https://machinelearningmastery.com/xgboost-for-regression/

Corporate Finance Institute. (n.d.). Regression Analysis. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/

Ashejim. (n.d.). C964 Task 2 Documentation. Retrieved from https://ashejim.github.io/C964/task2_doc.html