***Melbourne Housing***

**Project Initiation Document**

**Himanshu Khatri**

**Mohit**

**Ankur Srivastava**

**Wangjun Shen**

**Haiyue Wang**

**Pei-Yi Lui**

TABLE OF CONTENTS

1. **Purpose of the Project Initiation Document ……………………………………………..………….. 3**
2. **Melbourne Housing …………………………………………………………………..………………..……….. 4**
   1. Background ………………………………………………………………………………………….………….. 4
   2. Goals ……………………………………………………………………………………………………….………. 4
   3. Objectives ………………………………………………………………………………………….……………. 5
   4. Literature Review ……………………………………………………………………………….……………. 6
   5. Critical Success Factors ……………………………………………………………………………………. 7
3. **Scope …………………………………………………………………………………………..………………………. 8**
   1. In scope ……………………………………………………………………………….………….………………. 8
   2. Out of Scope ……………………………………………………………………..…………………….………. 8
4. **Project Plan ……………………………………………………………………………………………….……….… 9**
   1. Project Approach …………………………………………………………..………………….…………….. 9
   2. Project Schedule & Milestones (Stages) ………………………………………….……………….. 9
   3. Tools & Deliverables ……………………………………………………………………..………………….. 9
   4. Communication Plan ……………………………………………………………………..……….………. 10
5. **Project Risks & Known Issues …………………………………………………………..………………….. 11**
6. **Project Constraints, Assumptions, and Interfaces with other Projects ……………….... 12**
7. **Project Organization ………………………………………………………………………………….………... 13**
   1. Steering Committee ………………………………………………………………………………………...13
8. **Project Governance ………………………………………………………………………………..…………… 14**
   1. Meeting Structure …………………………………………………………………………………………... 14
   2. Progress Tracking …………………………………………………………………………..…………….…. 14
   3. Risk & Issue Management ………………………………………………………………………………. 14
9. **References …………………………………………………………………………………………………….….… 15**

Purpose of the Project Initiation Document

The Project Initiation Document for the Melbourne Housing Data Science Project serves as a foundational guide, outlining essential details about the project. It provides a clear orientation, framework, and set of guidelines for the project's execution.

The Project Initiation Document serves to outline key project details, including:

* Project Deliverables: What the project aims to achieve.
* Project Methodology: How the project will accomplish its goals.
* Project Timeline: When the project is expected to be completed.

This document includes crucial information like:

* Project Objectives & Goals: The overarching purpose and specific targets of the project.
* Project Scope: What is included and excluded from the project's scope.
* Project Approach & Schedule: The strategy and timeline for project execution.
* Project Organization & Responsibilities: Roles and responsibilities of team members.
* Project Governance & Processes: The structure for decision-making and the methodologies used in the project.

Melbourne Housing

## Background

**The Melbourne Housing Dataset is a collection of data pertaining to residential properties in Melbourne, Australia. This dataset encompasses a wide range of information, including the number of bedrooms, bathrooms, land size, location, age of property, and most importantly, property prices. With thousands of entries, it provides a comprehensive view of the city's housing market, making it a valuable resource for real estate professionals, data enthusiasts, and researchers alike. Through meticulous exploration and analysis of this dataset, we can gain insights into the dynamics of Melbourne's property market, uncover trends, and even develop predictive models to estimate property values.**

Preliminary data exploration of the Melbourne Housing Dataset can unveil a wealth of initial insights. It can reveal the distribution of property prices across different neighbourhoods, identifying potential hotspots or areas with more affordable housing. Exploration may also uncover correlations between various property attributes, such as the positive influence of additional bedrooms or bathrooms on property prices. Additionally, it can shed light on potential outliers or anomalies, indicating unique or exceptional property listings.

This dataset comprises a diverse array of information, such as property attributes (e.g., number of bedrooms, bathrooms, land size), geographical data (location and proximity to amenities), and, crucially, property prices making it a good candidate for exploration using machine learning techniques like linear regression which can be employed to predict property prices based on various features such as location, size, room count etc. Feature importance Algorithms like Random Forest or Gradient Boosting can also reveal the most influential factors affecting property prices.

Using clustering algorithms like K-means, we can identify clusters of houses with similar attributes like size, location, and room count. Or we can employ some anomaly detection method to identify outliers or properties that significantly deviate from the norm. For instance, these techniques can help pinpoint overpriced or under-priced properties in the Melbourne Housing Dataset.

The Melbourne Housing Dataset offers a chance to discover more about the city's real estate. By studying this data, we aim to find important information, like popular neighborhoods and what determines the property prices. With machine learning and careful study, we hope to reveal Melbourne's housing market secrets.

## Goals

In the dynamic real estate market of Melbourne, staying informed about housing trends, price shifts, and area preferences is crucial. We aim to find the important insights of the properties. For these objectives to be realized, we aim to:

1. **Data Exploration:**

* Understand the Melbourne Housing Dataset thoroughly by conducting comprehensive data exploration, including statistical analysis and visualization, to identify patterns and trends in the housing market.
* Uncover key factors influencing property prices, such as location, property attributes, and market dynamics, through data analysis.

1. **Machine Learning Model Development:**

* Split the dataset into two distinct groups: a training dataset and a testing dataset, ensuring that the model's performance is evaluated on unseen data.
* Build initial machine learning models, including regression model, using the training dataset to predict property prices accurately based on the dataset's features.
* Evaluate the performance of these initial models on the testing dataset to assess their ability to generalize to new, unseen data.
* Identify areas for model improvement based on the testing dataset's performance, aiming to enhance model accuracy and robustness.

1. **Model Improvement:**

* Refine and optimize machine learning models by fine-tuning hyperparameters, feature selection, or engineering, with the aim of improving model accuracy and robustness.
* Explore advanced machine learning techniques, such as ensemble methods or deep learning, to potentially achieve even better predictive performance.

## Objectives

**The primary objectives of the Melbourne Housing Data Science Project are to conduct a data exploration and analysis of the Melbourne Housing Dataset, identify key factors influencing property prices, develop and optimize machine learning models for accurate price prediction, and get valuable insights. Through these objectives, the project aims to empower individuals with a better understanding of Melbourne's housing market.**

The following tasks provide a detailed breakdown of the project's activities, organized according to the project's objectives. They serve as a roadmap for carrying out the Melbourne Housing Data Science Project effectively.

**Data Exploration**

In the initial phase of the Melbourne Housing Data Science Project, our primary objective is to conduct an exploration of the Melbourne Housing Dataset. This exploration involves delving into the data, performing statistical analysis to understand the distribution of property prices, as well as the range and variation of property attributes. Through the use of data visualization techniques, we aim to uncover meaningful patterns within the housing market. These insights may include spatial distributions and variations based on the suburb, land size, number of rooms, other basic amenities and comparing them over price, distance from CBD and other suitable factors. Also, we can detect notable outliers that could represent unique or exceptional property listings.

**Factors Influencing Property Prices**

Our next step is to find the factors that influence property prices within the Melbourne housing market. Through in-depth data analysis, we intend to investigate the impact of various variables on property prices. This section is about the relevant feature selection for our machine learning models. This includes examining the geographical influence of location, the significance of property attributes such as bedrooms, bathrooms, and land size, and the role of broader market dynamics such as economic conditions. We will employ statistical methods to quantify the relationships between these factors and property prices, providing clarity on factors have the biggest impact on property prices in Melbourne.

**Machine Learning Model Development**

The third objective is to use the power of machine learning to develop predictive models that can accurately estimate property prices. We will start by partitioning the dataset into distinct training and testing subsets to ensure proper model evaluation. Using these subsets, we will construct initial machine learning models, including regression models, to predict property prices based on the dataset's features and a classification model, to assign the properties to their respective categories. Our objective is to assess the performance of these initial models using the testing dataset, thereby determining their ability to generalize to unseen data. Areas for model enhancement will be identified based on the performance metrics generated during this evaluation.

**Model Improvement**

Objective four focus on the iterative process of model refinement and optimization. We will fine-tune machine learning models by adjusting hyperparameters, selecting pertinent features, or derive new ones. Additionally, we will explore machine learning techniques such as ensemble methods or deep learning to potentially achieve predictive performance. Continuous monitoring and updates will ensure that the models remain accurate and current as new data becomes available. Recommendations for further model enhancements will be developed based on ongoing performance evaluations.

**Insights and Documentation**

The final objective is to extract meaningful insights from the project's findings. We will summarize the insights acquired through data exploration and analysis, highlight key discoveries pertaining to Melbourne's housing market. The project's results will be communicated effectively through clear and informative reports, and visualizations, ultimately contributing to a better understanding of housing trends and dynamics in Melbourne. And finally documenting insights and details in a data science project is essential for preserving knowledge and facilitating communication. Organized in a consistent format and accessible to team members enhances transparency in the project's outcomes.

## Literature Review

Housing price prediction has become an essential area of research, with considerable implications for stakeholders in real estate, potential homebuyers, and policy makers. Machine learning and data-driven methods have increasingly been employed to enhance the accuracy and efficiency of these predictions. This literature review seeks to understand the landscape of research in the field, drawing on recent articles and papers focused primarily on the Melbourne housing market.

***Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia***

Property valuation is a multifaceted process, and this paper emphasizes the importance of considering sub-areas within a city, such as statistical areas (SAs), for more accurate price predictions. It investigates various techniques, including traditional hedonic price models and machine learning approaches like Random Forest and Gradient Boosting. The study indicates that these machine learning methods outperform other approaches, especially when considering smaller geographical regions. [1].

**Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data**

The use of machine learning techniques, specifically linear regression, and logistic regression, in analyzing housing price data from the Melbourne real estate market is the central theme of this paper. The study demonstrates the efficiency of these algorithms in predicting average housing prices and classifying houses sold in different councils into distinct categories. [2].

**Prediction of House Pricing using Machine Learning with Python**

Focusing on the use of Python libraries and regression methods, this paper offers insights into predicting house costs. It emphasizes the importance of refining the aspects used for calculating house prices to enhance prediction accuracy. Additionally, the paper outlines various graphical and numerical techniques employed in the house pricing model. [3].

**Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia**

House price forecasting is a central concern in real estate, and this study utilizes machine learning techniques to analyze historical property transactions in Australia, focusing on Melbourne. Notably, it highlights a significant variation in house prices between the city's most expensive and most affordable suburbs. The paper introduces a combination of Stepwise and Support Vector Machine (SVM) techniques to predict house prices, demonstrating competitive results based on mean squared error measurements. [4].

***Explainable housing price prediction with determinant analysis***

The study utilizes regression-based machine learning models to predict housing prices and identify key factors affecting them using publicly available datasets (Ames and Melbourne Housing datasets). The research finds that in the Ames dataset, living area, basement size, and remodeling age are top determinants, while in the Melbourne dataset, more rooms/bathrooms, larger land size, and proximity to the CBD drive higher prices. Using explainable SHAP analysis, these factors are highlighted for informed decision-making by buyers and sellers. Limitations include skewed price distributions, requiring conversion to categorical data, and categorization methods' effectiveness remains a concern. [5].

In summary, these papers collectively emphasize the significance of data exploration and machine learning in the context of the Melbourne Housing Dataset. They illustrate the potential for developing accurate predictive models for Melbourne housing market. These studies also highlight the importance of considering sub-area analysis, classifying areas by councils, and advanced techniques for improving the accuracy of housing price predictions.

## Critical Success Factors

Critical success factors (CSFs) are elements or conditions that are essential for the successful achievement of a project's objectives. In the context of the Melbourne Housing Data Science Project, here are some critical success factors to consider:

1. The success of the project hinges on the reliability and accuracy of the Melbourne Housing Dataset. If the data is not accurate, insights and predictions derived from it might mislead from good results.
2. Before finalizing the machine learning algorithms, their understanding and functioning is very important to use them in the project for better outcomes.
3. Identifying and using the most impactful variables can significantly improve the accuracy and utility of prediction models. It's essential to invest time in understanding which features drive the best predictions.
4. Housing data can have outliers due to unique property sales or market anomalies. Properly handling or accounting for these in the models is crucial to avoid skewed predictions.
5. Define appropriate evaluation metrics for the machine learning models, such as mean squared error or R-squared, and regularly assess model performance against these metrics.
6. Experiment with different hyperparameters for the machine learning algorithms used. Fine-tuning these hyperparameters can significantly impact model performance.
7. Implement cross-validation techniques to assess model performance more accurately and to avoid overfitting, ensuring that the model works well to unseen data.

Scope

## In Scope

* **Visualization Tools:** Implement tools for visual representation of data, like heat maps showing property price concentrations.
* **Correlation Analysis**: Investigate correlations between different features (e.g., correlation between property price and proximity to CBD) to identify potential factors influencing property prices.
* **Geospatial Analysis**: Utilize geospatial techniques to map property locations, examine spatial patterns, and assess the impact of location on property values.
* **Price Prediction:** Predict property prices based on current market trends, historical data, and other influencing factors.
* **Clustering**: Implement clustering algorithms (e.g., K-means) to group similar properties, helping to identify market segments or neighborhoods with common characteristics.
* **Feature Importance:** Determine the most influential factors affecting property prices using algorithms like Random Forest or Gradient Boosting, which can reveal feature importance scores.
* **Anomaly Detection:** Apply anomaly detection techniques (e.g., Isolation Forest, One-Class SVM) to identify outliers or unusual property listings that may require special attention.
* **Popularity:** Analyze which Melbourne areas are having higher frequency of sales or what kind of properties attract higher price tags.

## Out of Scope

* **Data Beyond Melbourne:** Collection or analysis of housing data from regions outside Melbourne.
* **Regular Data Updates:** Updating the dataset frequently to reflect the most recent market conditions cannot be done here.
* **Exact Long-Term Predictions:** As the dataset is of only one financial year it’s not possible to make predictions for an extended period, like multiple years into the future, as it can be highly unreliable.
* **External factors:** Effect of big external factors like change in RBA cash rate or change in migration rates cannot be encompassed in models built using the chosen dataset.
* **Neighborhood Growth Potential:** Predict which Melbourne areas or suburbs are poised for growth in terms of infrastructural development, population influx, and other factors.
* **Rental Yield Predictions:** Predicting rental yield will require additional data like previous leases and proximity to amenities.
* **Future price predictions:**  As the dataset is of only one year time series analysis cannot predict future price trends.

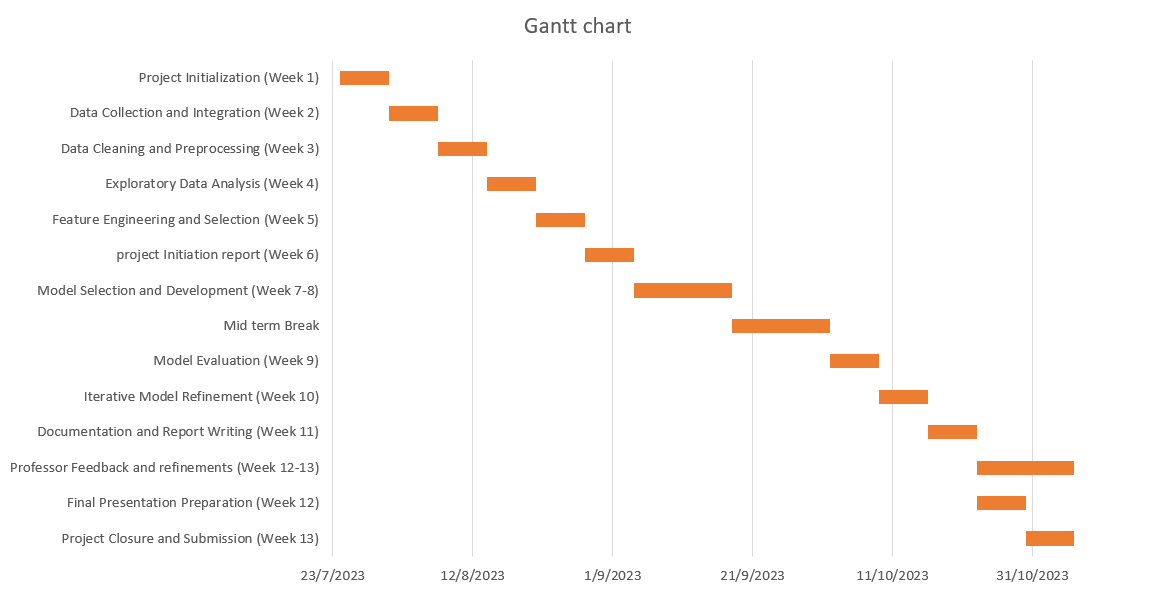
Project Plan

## Project Approach

In the Melbourne Housing Data Science Project, our approach comprises two essential phases. Initially, we prepare and explore the dataset, ensuring data accuracy and highlights the important insights. This involves data cleaning, statistical analysis to understand factors influencing property prices. Feature selection enhances the dataset for machine learning.

In the second phase, we develop machine learning models, including careful data splitting into training and testing set, model selection, and optimization. In our Melbourne Housing Data Science Project, we're really careful about checking how good our predictions are. We use some tricks to understand why our models make certain predictions. We will work to make our models better.

## Project Schedule & Milestones (Stages)



## Tools & Technologies

In the Melbourne Housing Data Science Project, a set of tools and technologies is harnessed for data analysis and machine learning. Python serves as the core programming language, with libraries like pandas and NumPy for data manipulation and Matplotlib, Seaborn for visualization. Jupyter Notebooks enable interactive coding and documentation interface. Docker and Kubernetes assist in containerization and orchestration, and Flask or Django are used for model deployment. Collaboration and documentation rely on Microsoft Office.

## Communication Plan

**General**

**The target of project in the course is to analyze the Melbourne ‘housing dataset. All the projects will be completed by every member of group. In order to achieve the ultimate goal, the efforts of each member are very important, and the communication of the team is also very important, which even directly determines the efficiency and success of the entire project implementation. Good team communication can lay a good foundation for efficient and effective completion of project goals, and at the same time can achieve experience sharing, reduce or even avoid detours to reduce the workload of members and ensure quality project implementation. Therefore, a team communication plan is specified to help the team efficiently complete the project goals.**

The plan is divided into 3 stages according to the project implementation process, and the focus of each stage specifies different communication priorities and plans.

There are three main communication methods for the team, based on the project management page defined by Notion, the WhatsApp project group discussion group for instant communication, and online or offline meetings held at important nodes. WhatsApp allows all members to communicate in real time and solve problems encountered in real time. Notion can track the progress of all project nodes and task modules to control the overall project. Important node meetings can share, summarize, and sort out the phased results, and at the same time, make instant adjustments to the currently uncertain or inappropriate parts.

**Introduction (initializing communicative effort)**

According to the requirements of the course project, the main goal of this stage is to select the appropriate project research data. In this stage, all members will search for valuable open-source data sets that can be researched according to the defined characteristic requirements. Each member will select at least one data set and indicate the reason for choosing the data set and share the collected data and annotations with others. Personnel, project members finally select the appropriate data set through voting, and data sharing is carried out through the designated Notion-based project management within the group.

**Engagement (concluding conceptual phase, preparing development phase)**

The main objective of this phase is to specify the final research objectives based on the defined research data. All members were familiarized with the research data. And discuss the final research goals. On the basis of determining the research objectives, the task is subdivided, and the task modules are divided according to the subdivision results, and then all the task modules are assigned to specific members for implementation. The communication at this stage mainly shares the members’ understanding of the data, and all results are shared through the Notion project management page.

**Implementation (concluding development phase, preparing deployment)**

Members implement according to the assigned task modules, and all implementation steps and current results should be shared on the project management page.

Project Risks & Known Issues

|  |  |  |
| --- | --- | --- |
| Risk Description | Impact | Mitigation Strategies |
| Lack of Data Quality and Consistency | This risk could result in inaccurate analysis and incorrect conclusions drawn from the dataset due to inconsistencies or errors in the data. | Implement data preprocessing and cleaning steps to ensure data quality. Regularly validate and clean the dataset. |
| Insufficient Domain Knowledge | This risk may lead to misinterpretation of analysis results and incorrect conclusions due to inadequate understanding of the real estate market and variables in the dataset. | Collaborate with domain experts or conduct research to gain a better understanding of the real estate market and variables in the dataset. |
| Group Member Availability Issues | This risk could cause project delays and incomplete work if team members are unavailable, or communication breaks down. | Set clear communication channels, establish a timeline with milestones, and encourage regular updates among group members. |
| Technical Challenges in Analysis | This risk may result in the inability to perform complex analyses, leading to suboptimal or incomplete results. | Allocate time for learning and troubleshooting. Seek guidance from instructors, online resources, or peers. |
| Unforeseen External Factors | This risk could cause disruptions due to unexpected external events, such as changes in data sources or other unforeseen circumstances. | Have contingency plans in place to adapt to unexpected changes. Regularly check the dataset source for updates. |
| Lack of Statistical Significance | This risk could cause disruptions due to unexpected external events, such as changes in data sources or other unforeseen circumstances. | Determine minimum data requirements for statistically valid conclusions. Consider seeking additional data sources if needed. |
| Ineffective Group Collaboration | This risk could result in miscommunication and duplication of efforts, potentially hampering the project's progress. | Establish clear roles and responsibilities. Use collaboration tools and maintain consistent communication within the group. |

1. **Lack of Data Quality and Consistency:**

* **Impact**: This risk could result in inaccurate analysis and incorrect conclusions drawn from the dataset due to inconsistencies or errors in the data.
* **Mitigation Strategies:** To address this risk, the project team will implement comprehensive data preprocessing and cleaning procedures. This involves identifying and rectifying data inconsistencies, outliers, and missing values. Regular validation and cleaning of the dataset will be performed to maintain data quality over time.

1. **Insufficient Domain Knowledge:**

* **Impact:** This risk may lead to misinterpretation of analysis results and incorrect conclusions due to inadequate understanding of the real estate market and variables in the dataset.
* **Mitigation Strategies:** The team will collaborate closely with domain experts who possess specialized knowledge in the real estate market. Additionally, the team will conduct thorough research to gain a comprehensive understanding of the variables and factors that influence the housing market, ensuring more accurate analysis and interpretation.

1. **Group Member Availability Issues:**

* **Impact:** This risk could cause project delays and incomplete work if team members are unavailable, or communication breaks down.
* **Mitigation Strategies:** Clear communication channels will be established within the team. A timeline with well-defined milestones will guide the project's progress, and regular updates will be encouraged among group members. This approach aims to maintain consistent progress and minimize the impact of any availability issues.

1. **Technical Challenges in Analysis:**

* **Impact:** This risk may result in the inability to perform complex analyses, leading to suboptimal or incomplete results.
* **Mitigation Strategies:** The team will allocate dedicated time for learning and troubleshooting technical challenges that arise during the analysis process. Instructors, online resources, and peers will be consulted to seek guidance and solutions to overcome technical hurdles.

1. **Unforeseen External Factors:**

* **Impact:** This risk could cause disruptions due to unexpected external events, such as changes in data sources or other unforeseen circumstances.
* **Mitigation Strategies:** Contingency plans will be established to adapt to unexpected changes. Regular monitoring of the dataset source will ensure that the team remains informed about any updates or modifications that might affect the project's analysis.

1. **Lack of Statistical Significance:**

* **Impact:** This risk may lead to drawing conclusions based on insufficient data, reducing the validity of the analysis.
* **Mitigation Strategies:** The team will determine the minimum data requirements necessary for statistically valid conclusions. If the dataset does not meet these requirements, the team will consider seeking additional data sources to supplement the analysis and ensure more reliable insights.

1. **Ineffective Group Collaboration:**

* **Impact:** This risk could result in miscommunication and duplication of efforts, potentially hampering the project's progress.
* **Mitigation Strategies:** Roles and responsibilities will be clearly defined within the group. Collaboration tools will be utilized to facilitate effective communication. Consistent updates and regular check-ins will be encouraged among team members to maintain efficient collaboration.

Project Constraints, Assumptions, and Interfaces with other Projects

The project is subject to the following constraints:

1. **Limited Data Duration:** The project relies solely on data from the span of 2016, July to 2017, June, thus lacking a broader historical context. This limitation might hinder the model's ability to capture long-term trends accurately.
2. **Inadequate Data Volume:** The available dataset may not contain a sufficient quantity of data points to construct a prediction model with a high degree of accuracy. Insufficient data can lead to unreliable predictions.
3. **Future Generalization Concerns:** Due to the reliance on a relatively short historical timeframe, any predictions generated by the model might not effectively generalize to future periods. The model's predictive power beyond the observed data range could be compromised.
4. **Impact on Accuracy and Reliability:** The shortage of data, both in terms of timeframe and volume, could significantly affect the accuracy and reliability of the prediction model's outcomes. Predictions might not be robust enough to handle real-world scenarios.
5. **District Definition Variability:** The predefined districts used for analysis may not perfectly align with the actual geographical distribution. Variability in district definitions could introduce biases or inaccuracies into the model's predictions.

The primary objective of this project is to develop a prediction model utilizing historical data on Melbourne housing prices spanning from July 2016 to June 2017, encompassing an entire financial years’ worth of data. Despite this clear goal, the project faces various challenges. These include limitations stemming from the scarcity of data covering a broader timeframe, which hampers the construction of a comprehensive model. Furthermore, the accuracy of the prediction model is compromised due to insufficient data, and the validity of results is influenced by the fact that the predefined districts may exhibit slight disparities in relation to the actual geographical distribution. It's essential to note that the available dataset only pertains to the years 2016 and 2017.

Consequently, it's inappropriate to assume that predictions generated by the model will accurately mirror future scenarios, as the lack of extended data undermines the model's reliability and precision. Moreover, the district divisions are based on our own criteria, utilizing longitude and latitude information to define them, which could potentially introduce deviations from the real geographical area distribution. These constraints collectively shape the project's scope and the expectations for the predictive model's capabilities and limitations.

Project Organization

## Steering Committee

The committee will review and monitor the project plan and progress. The project committee supervises the project on a weekly basis and urges each task module to update the status and organize discussions if necessary.

Project Governance

Meeting Structure

1. Scheduled weekly meetings for members to describe their task modules.
2. The meeting will consist of two parts, the first part is a description, another part is a discussion.
3. For the description, each member should talk about their own tasks. All they talked about should consist of four parts.

* Describe the status of your own task module.
* Share what’s the problem you come across.
* What’s the next plan?
* Other related issues.

1. For the discussion, all members could talk based on the content from the description.

Progress Tracking

All the tasks of the project will be tracked using the project management page in Notion.

Page Url: <https://ember-beret-8ce.notion.site/Project-Initiation-Document-Part2-Chapter4-7-8-aa036918d67d41fe89c08f910c2321d9?pvs=4>

Risk & Issue Management

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Impact | | | | |
| Probability | Negligible | Low | Medium | Very high | Extreme |
| Rare | LOW | LOW | LOW | MEDIUM | HIGH |
|  |  | Group member Availability Challenges  Unforeseen External Factors | Ineffective Group Collaboration |  |  |
| Unlikely | LOW | LOW | MEDIUM | MEDIUM | HIGH |
|  |  |  | Insufficient Domain Knowledge  Technical Analysis Challenges  Lack of Statistical Significance |  |  |
| Possible | LOW | MEDIUM | MEDIUM | HIGH | HIGH |
| Likely | MEDIUM | MEDIUM | HIGH | HIGH | EXTREME |
|  |  |  |  | Data Quality and Consistency Issues |  |
| Almost Certain | MEDIUM | MEDIUM | HIGH | EXTREME | EXTREME |

|  |  |
| --- | --- |
| Risk Description | Mitigation Plan |
| Data Quality and Consistency Issues | Implement rigorous data preprocessing and cleaning. Regularly validate data to ensure accuracy. |
| Insufficient Domain Knowledge | Collaborate with domain experts to understand real estate market intricacies. Conduct thorough research on relevant variables. |
| Group Member Availability Challenges | Establish clear communication channels. Define roles, responsibilities, and milestones. Encourage regular updates among group members. |
| Technical Analysis Challenges | Allocate dedicated time for learning and troubleshooting. Seek guidance from instructors, peers, and online resources. |
| Unforeseen External Factors | Develop contingency plans to adapt to unexpected changes. Monitor dataset source for updates and changes. |
| Lack of Statistical Significance | Define minimum data requirements for meaningful conclusions. Consider additional data sources if needed. |
| Ineffective Group Collaboration | Establish clear roles and communication protocols. Utilize collaboration tools for consistent updates within the group. |

References

[1] Gao, Q, Shi, V, Pettit, C & Han, H 2022, ‘Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia’, Land Use Policy, vol. 123, p. 106409–.

[2] He, K & He, C 2021, ‘Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data’, in 2021 IEEE International Conference on Computing (ICOCO), IEEE, pp. 241–246.

[3] Jain, M, Rajput, H, Garg, N & Chawla, P 2020, ‘Prediction of House Pricing using Machine Learning with Python’, in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), IEEE, pp. 570–574.

[4] Phan, TD 2018, ‘Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia’, in 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), IEEE, pp. 35–42.

[5] Teoh, EZ, Yau, W-C, Ong, TS & Connie, T 2022, ‘Explainable housing price prediction with determinant analysis’, International Journal of Housing Markets and Analysis, vol. 16, no. 5, pp. 1021–1045.