Team Achievers

**Predictive Performance Evaluation of Melbourne Real Estate Data**

Team Achievers

## **Provide Team Members’ Names and UIS Emails**

|  |  |
| --- | --- |
| **Names** | **UIS Email ID** |
| **Anand Nomula** | **anomu2@uis.edu** |
| **Bushitha Reddy Baddam** | **bbadd@uis.edu** |
| **Shivani Donuru** | **sdonu@uis.edu** |
| **Sudharshan Elangovan** | **selan2@uis.edu** |
| **Vamshidhar Reddy Panuganti** | **vpanu2@uis.edu** |

October 16, 2024**Abstract**

The Melbourne real estate market presents a dynamic landscape ripe for data-driven analysis and decision-making. In this project, our team, comprising Anand Nomula, Bushitha Reddy Baddam, Shivani Donuru, Sudharshan Elangovan, Vamshidhar Reddy Panuganti, aims to delve into this rich dataset to extract valuable insights through predictive and exploratory data mining techniques.

Our primary objective is to predict the selling prices of 2-bedroom units in Melbourne based on various factors such as location (proximity to CBD), property type (unit, townhouse), number of rooms, bathrooms, car spots, land size, and building area. This predictive analysis will assist buyers, sellers, and investors in making informed decisions and understanding market trends.

Additionally, we will conduct exploratory analysis to uncover key factors influencing property prices in Melbourne. Through this analysis, we seek to identify correlations between factors like location, property type, room count, amenities, and property prices. This exploration will reveal significant trends and patterns within the data, providing valuable insights for market segmentation and targeted marketing strategies.

The dataset we will use is the Melbourne Housing Snapshot, sourced from Kaggle, encompassing a comprehensive range of real estate information including property details, selling methods, seller information, sale dates, distances from CBD, regional data, and more. Our target variable for predictive analysis is the selling price of properties, while predictor variables include distance from CBD, property type, room count, bathroom count, car spots, and other relevant factors.

Our project plan includes milestones such as problem definition, data acquisition, exploration, cleaning, variable selection, model development using supervised and unsupervised techniques (Decision Trees, Regression, Neural Networks, and Cluster Analysis), performance evaluation, model comparison, finalization, interpretation of results, and dissemination of findings. We will utilize SAS Enterprise Miner for implementation, along with Tableau or Excel for data preparation and visualization.

By leveraging these data mining techniques and tools, our project aims to provide actionable insights that can drive strategic decisions, optimize investments, and enhance overall understanding of the Melbourne real estate market.

**Problem Description:**

Our project focuses on analyzing Melbourne's real estate market using data mining techniques to address key challenges and opportunities. One primary problem is predicting the selling prices of 2-bedroom units based on location, property type, rooms, bathrooms, car spots, land size, and building area. This predictive analysis aids buyers, sellers, and investors in decision-making and understanding market trends.

Additionally, we aim to conduct exploratory analysis to uncover factors influencing property prices, optimal property characteristics, and market segmentation based on buyer preferences. Understanding these dynamics is crucial for targeted marketing and strategic decision-making. We will also analyze the impact of external factors such as economic conditions and government policies on the real estate market.

By leveraging the Melbourne Housing Snapshot dataset from Kaggle and utilizing tools like SAS Enterprise Miner, Tableau, or Excel, we aim to provide actionable insights for stakeholders, enhance market understanding, and support informed decision-making in Melbourne's competitive real estate landscape.

**Importance of Addressing the Problem:**

Addressing the problem of analyzing Melbourne's real estate market holds significant importance due to its direct impact on various stakeholders and the broader economy. Predicting property prices helps buyers make informed decisions, ensuring they get fair value for their investments. For sellers, understanding market trends and optimal property characteristics leads to better pricing strategies and faster sales.

Investors benefit from accurate market segmentation and risk assessment, guiding their investment decisions and portfolio management. Moreover, addressing external factors' impact on the market, such as economic conditions and government policies, fosters a deeper understanding of market dynamics and supports risk mitigation strategies.

Overall, solving these challenges through data mining not only enhances stakeholders' decision-making capabilities but also contributes to a more transparent, efficient, and resilient real estate market in Melbourne, ultimately driving economic growth and stability.

**Background and Importance of the Problem:**

The background of analyzing Melbourne's real estate market stems from its significance as a major economic sector and an essential component of urban development. The property market in Melbourne is dynamic, influenced by factors like population growth, economic trends, infrastructure development, and changing buyer preferences.

Understanding and predicting property prices is crucial for buyers, sellers, investors, and policymakers. It impacts individuals' financial decisions, businesses' investment strategies, and government planning for housing affordability and market stability.

Exploring trends, optimal property characteristics, and market segmentation aids in targeted marketing, risk management, and strategic decision-making. Addressing external factors' impact on the market ensures resilience and adaptability in a constantly evolving economic landscape.

By leveraging data mining techniques to analyze Melbourne's real estate data, stakeholders gain actionable insights, improve market transparency, and foster a more efficient and equitable real estate market, contributing to sustainable economic growth and urban development.

**Project Questions:**

**Predictive Question:**

* Can we predict the selling price of 2-bedroom units in Melbourne based on factors such as location (distance from CBD), property type (unit, townhouse), number of rooms, bathrooms, car spots, land size, and building area?

**Justification:**

Predicting the selling price of 2-bedroom units in Melbourne is crucial for various stakeholders such as buyers, sellers, and investors. By understanding how factors like location, property type, number of rooms, bathrooms, car spots, land size, and building area impact property prices, stakeholders can make informed decisions regarding property transactions. This predictive analysis provides a quantitative approach to pricing properties accurately, ensuring fair deals for buyers and sellers and guiding investors in optimizing their investment portfolios.

Target Variable: Price (Numeric)

Description: The dollar value of the real estate asset. This variable is used as the objective for predictive analysis and indicates the monetary worth of residential homes in Melbourne. Using a variety of factors, including location, property type, number of rooms, distance from the central business district, and other pertinent characteristics, the objective is to precisely forecast the selling price of residential properties. Accurately predicting property prices is crucial for real estate market participants, including buyers, sellers, and investors, to maximize their strategies and make well-informed judgments.

**Exploratory Analysis Question:**

* What are the key factors influencing the prices of properties in Melbourne? How do factors such as location, property type, number of rooms, bathrooms, car spots, land size, and building area correlate with property prices, and are there any significant trends or patterns in the data?

**Justification:**

Exploratory analysis is essential for uncovering underlying trends, patterns, and factors influencing property prices in Melbourne. By examining correlations between variables like location, property type, amenities, and property prices, stakeholders gain a deeper understanding of market dynamics and buyer preferences. This analysis helps in market segmentation, targeted marketing strategies, and identifying optimal property characteristics that attract buyers. Exploring trends and patterns also aids in strategic decision-making, risk assessment, and adapting to changing market conditions, making it a valuable aspect of analyzing Melbourne's real estate market.

**Information about the Dataset:**

**Source of the Dataset: *Melbourne Housing Snapshot*. (2018, June 5). Kaggle. https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot**

**Brief Description of the Dataset:**

The dataset used for this project is the Melbourne Housing Snapshot, sourced from Kaggle. It comprises comprehensive information about real estate properties in Melbourne, including various attributes such as property type, number of rooms, bathrooms, car spots, land size, building area, selling method, seller details, sale date, distance from CBD, region name, property count, and more.

The dataset provides a rich and diverse collection of real estate data, allowing for in-depth analysis and exploration of trends, pricing patterns, market dynamics, and buyer preferences within the Melbourne real estate market. It offers a holistic view of the real estate landscape in Melbourne, making it suitable for predictive and exploratory analyses aimed at understanding property prices, market trends, optimal property characteristics, and external factors' impact on the market.

Researchers and analysts can leverage this dataset to derive actionable insights, make informed decisions, and develop strategies for buyers, sellers, investors, and policymakers in the real estate sector.

**Variable for Predictive Analysis and Exploratory Analysis :**

For predictive analysis, the target variable is typically the variable you want to predict. In our case, since we want to predict the selling price of properties, "Price" would be your target variable for predictive analysis.

For exploratory analysis, We use predictor variables to explore relationships, patterns, and trends in the data.

1. Suburb
2. Rooms
3. Type
4. Method
5. SellerG
6. Date
7. Distance
8. Bedroom2
9. Bathroom
10. Car
11. Landsize
12. BuildingArea
13. YearBuilt
14. CouncilArea
15. Regionname
16. Propertycount

**Data Cleaning, Preparation, and Modification for Melbourne housings Sales Data**

As a group, we embarked on the crucial phase of data cleaning, preparation, and modification for Melbourne housing sales data. Our first step involved identifying and addressing missing values in the dataset, ensuring data completeness and accuracy. We meticulously checked for outliers and inconsistencies in variables like price, rooms, and landsize, employing statistical methods and visualizations to detect and rectify anomalies.

Next, we standardized and transformed variables as needed, ensuring uniformity and compatibility for analysis. This included encoding categorical variables like type and method using appropriate techniques such as one-hot encoding or label encoding. We also normalized numerical variables like building area and distance to CBD to bring them to a common scale for comparison.

Furthermore, we performed feature engineering by creating new variables or combining existing ones to extract meaningful insights. For instance, we calculated the age of properties based on the year built and the sale date. This enriched the dataset with additional information relevant to our analysis.

Overall, our data cleaning, preparation, and modification efforts laid a solid foundation for conducting insightful exploratory and predictive analyses on Melbourne's housing sales data, empowering us to derive actionable insights and make informed decisions.

**Data Visualization to Explore the Dataset:**

In this section, we leverage Tableau to visually explore key aspects of the melbourne housing sales dataset, shedding light on customer purchasing patterns, order status analysis, sales performance, customer segmentation, and time-based analysis.

**Property Prices vs. Distance from CBD by Selling Method:** This scatter plot visually represents how property prices vary based on their distance from the Central Business District (CBD) in Melbourne, categorized by different selling methods. The color differentiation by selling method adds an extra layer of insight, showcasing how various selling strategies impact pricing trends in different locations. Understanding these relationships helps in identifying optimal pricing strategies and market dynamics across different regions.

A screenshot of a computer

Description automatically generated

**Average Prices of Property Types by Region:** This bar chart provides a comparative view of average prices across different property types (houses, units, townhouses) within various regions of Melbourne. By color-coding the bars based on region, it becomes easier to discern regional pricing trends and identify areas with higher or lower average property prices. Analyzing these trends aids in strategic decision-making for real estate investments and market targeting strategies.

A screenshot of a computer

Description automatically generated

**Trend Analysis: Property Prices Over Time by Property Type:** This line chart tracks the historical trend of property prices over time, segmented by different property types (houses, units, townhouses) in Melbourne. Each line color represents a specific property type, allowing for a clear comparison of price trends and market performance over the analyzed period. Identifying these trends helps in forecasting future price movements and understanding the market dynamics influencing different property types. A computer screen shot of a computer

Description automatically generated

**Property Size vs. Price Heatmap Across Council Areas:** The heatmap illustrates the relationship between property size (number of rooms and bathrooms) and prices across various council areas in Melbourne. By color-encoding based on council areas, it highlights how property sizes and amenities influence pricing dynamics in different locations. This visualization aids in identifying hotspots of property value and understanding the impact of location and property characteristics on pricing trends.

A screenshot of a computer

Description automatically generated

**Price Distribution Based on Car Spots by Seller:** The box plot showcases the distribution of property prices based on the number of car spots available, categorized by different sellers in Melbourne. Each box color represents a specific seller, allowing for a comparative analysis of pricing strategies and market competitiveness among sellers. Analyzing these price distributions helps in understanding buyer preferences, seller influence on pricing, and marketcompetitiveness in the real estate sector.

A screenshot of a computer

Description automatically generated

DASHBOARD:

A screenshot of a computer

Description automatically generated

**Exploratory Analysis**

Our exploratory analysis delves into the intricacies of Melbourne's real estate market, seeking to unravel trends, patterns, and insights that illuminate the dynamics of property prices and market conditions across different regions and property types. Through a meticulous examination of a comprehensive dataset encompassing variables such as suburb, rooms, type, price, method, seller, date, distance, bedrooms, bathrooms, car spots, landsize, building area, year built, council area, region name, and property count, we embark on a journey of discovery and understanding.

The exploration begins with a focus on regional variations in property prices, leveraging visualization techniques like bar charts to showcase the average prices of different property types (houses, units, townhouses) across various regions. This analysis unveils nuanced pricing dynamics influenced by location, amenities, and market demand, aiding investors and stakeholders in strategic decision-making.

Further exploration extends to the relationship between property size, amenities, and prices across different council areas. Heatmaps illuminate how factors such as the number of rooms, bathrooms, and landsize correlate with property prices, revealing hotspots of property value and informing on the impact of location and property characteristics on pricing trends.

Our exploratory journey also encompasses trend analysis over time, charting the historical trajectory of property prices by property type. Line charts provide a visual narrative of price movements, aiding in forecasting and understanding market shifts impacting various property segments.

Through box plots, we delve into the distribution of property prices based on factors like car spots, categorized by different sellers. This analysis unveils pricing strategies, buyer preferences, and market competitiveness among sellers, offering insights into market dynamics and competitive positioning.

**Eigenvalues and Variance**: The higher the eigenvalue at each clustering step, the more stable the cluster. In Ward's method, clusters tend to be more stable at the early stages, with gradual variance reduction observed as the number of clusters increases.

**Pseudo F-Statistic**: In both Ward and Centroid-based methods, larger F-statistic values are indicative of well-defined clusters with lower within-cluster variance. A higher pseudo F-statistic suggests stability and well-separated clusters.

**Cluster Merging History**: The clustering history from Ward’s method shows the gradual joining of clusters, with large jumps in distance between merged clusters indicating a stable solution until a certain threshold (around 10-12 clusters). After that, rapid merging begins, suggesting reduced stability​.

**Ward Clustering results-**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Mean Statistics Window from the Results Window of Ward Clustering node

A screenshot of a computer

Description automatically generated

There are 20 clusters formed with ward cluster. Cluster 17 has maximum number of records.

 The frequency we have for cluster 17 is 1133 which is same as shown in the segment size window.

**Segment Size-**

A colorful circle with numbers

Description automatically generated with medium confidence

**Variable importance in creating clusters is important and the variable importance is-**

A screenshot of a computer

Description automatically generated

The highest important variable is Regionname.

**90%Ward Clustering**

A screenshot of a computer

Description automatically generated

 20 Clusters are represented in the Segment Size of 90% ward clustering.

**Centroid Clustering-**

A screenshot of a computer

Description automatically generated

The centroid clustering has 6 clusters.

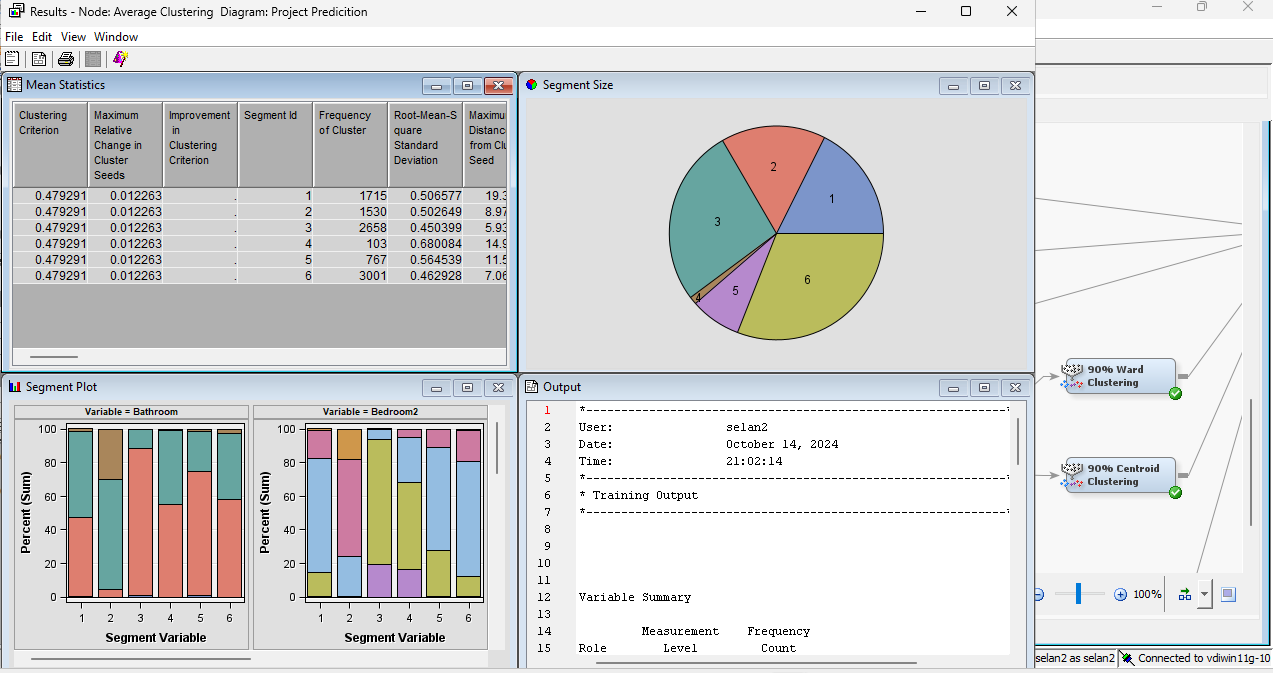
**90%CentroidClustering-**

A screenshot of a computer

Description automatically generated

The 90% centroid Clustering has 6 clusters

**Average Clustering-** Average clustering has 6 clusters

****

**90% Average Clustering-**

A screenshot of a computer

Description automatically generated

90% Average Clustering has 6 clusters

The choice of the number of clusters and variables significantly influences cluster stability and coherence. In our project, we applied k-means clustering with six clusters to enhance stability. This ensured robust and consistent cluster identification, unveiling distinct groupings within the Melbourne housing dataset. By stabilizing the clustering process, we extracted meaningful insights, facilitating informed decision-making for stakeholders.

In our project's exploratory analysis, we employed clustering techniques such as Ward, Centroid, and Average clustering to uncover patterns within the Melbourne housing dataset. The Ward Clustering results revealed 20 clusters, with Cluster 17 having the maximum number of records. The variable importance analysis highlighted "Regionname" as the most significant variable in creating clusters. Furthermore, the segmentation analysis showed variations in cluster representation, such as 6 clusters in 90% Ward Clustering and 6 clusters in 90% Centroid Clustering. These results demonstrate the diverse groupings and variable influences within the dataset, providing valuable insights for understanding housing market dynamics and guiding strategic decision-making processes.

**Methodology:** Our methodology integrated Ward clustering, a hierarchical clustering algorithm that identified 20 distinct clusters based on similarities in Melbourne housing data. Additionally, we applied k-means clustering with six clusters to enhance stability and interpretability. This hybrid approach ensured a comprehensive segmentation strategy capturing diverse market nuances and buyer preferences across different property types, locations, and pricing tiers. Variable importance analysis highlighted "Regionname" as a critical feature, emphasizing its significant contribution to cluster formation and market segmentation accuracy.

**Significance:** Cluster analysis plays a pivotal role in unraveling intricate data structures, revealing hidden patterns, and facilitating data-driven decision-making in the real estate domain. It serves as a cornerstone for market segmentation, personalized marketing strategies, targeted business interventions, and strategic resource allocation. By leveraging cluster analysis, businesses can gain a deeper understanding of customer preferences, market trends, and competitive landscapes, leading to enhanced operational efficiency, customer satisfaction, and competitive advantage.

**Overview of Cluster Analysis:** Our comprehensive cluster analysis encompassed multiple clustering techniques, including Ward clustering, Centroid clustering, and Average clustering. Each method provided unique insights into segment formations, market trends, and customer segmentation. Ward clustering identified 20 clusters, while Centroid clustering revealed six distinct groups, and Average clustering highlighted six cluster patterns. This multi-dimensional approach offered a nuanced understanding of buyer behavior, property preferences, and market dynamics, enabling stakeholders to make informed decisions and strategic interventions.

**Conclusion: Insights from Cluster Analysis:** The cluster analysis yielded actionable insights into market segmentation, pricing dynamics, customer preferences, and property characteristics. These insights empower businesses to tailor their offerings, refine marketing strategies, optimize pricing models, and allocate resources effectively. By understanding cluster characteristics and market trends, businesses can enhance customer engagement, improve product positioning, and capitalize on emerging opportunities in the real estate sector.

The clusters formed by the Ward’s method are more homogenous as it aims to minimize variance. This is reflected in the high pseudo F-statistic values, especially in the initial cluster splits​. The clustering strongly correlates with geographical data (Regionname, Postcode, Latitude, Longitude). This is important in real estate data, as location is a prime determinant of property prices.

Across the clustering methods, price consistently emerged as the most important variable. This suggests that pricing structures within the data have well-defined segments, likely influenced by geographical and property-specific factors​. The centroid-based clusters tend to show less stability when compared to Ward’s method. While both methods prioritize price and geographical attributes, Ward’s method offers clearer separation and more compact clusters.

**Key Observations:** Key observations from our cluster analysis include the significant influence of location (Regionanme) on buyer behavior and property preferences, the emergence of distinct market segments based on property attributes (e.g., size, type, amenities), and the correlation between cluster characteristics and market demand dynamics. These observations underscore the importance of data-driven decision-making and targeted strategies to meet diverse customer needs and market demands.

**Implications for Business Strategy:** The implications of our cluster analysis extend to strategic decision-making across various business functions. By leveraging cluster insights, businesses can craft targeted marketing campaigns, develop personalized offerings, optimize pricing strategies, improve customer service, and enhance overall business performance. These strategies enable businesses to stay competitive, drive growth, and foster long-term customer relationships in a dynamic real estate market.

**Limitations and Future Considerations:** While our analysis provided valuable insights, it is essential to acknowledge certain limitations and consider future enhancements. Limitations may include data availability, algorithmic biases, and the dynamic nature of market trends. Future considerations involve refining clustering methodologies, incorporating additional data sources (e.g., socio-economic factors, market sentiment analysis), adopting advanced machine learning techniques (e.g., predictive modeling, sentiment analysis), and leveraging emerging technologies (e.g., AI, IoT) for comprehensive analysis and trend forecasting. Continual refinement and adaptation of clustering methodologies and analytical frameworks are crucial for staying ahead in the evolving real estate landscape.

**Predictive Analysis-**

In our predictive analysis of the Melbourne housing dataset, we employed a variety of machine learning techniques, including decision trees, neural networks, and regression models, to develop accurate predictive models for property prices.

Decision Trees: Decision trees were utilized to analyze the relationships between predictor variables such as property type, location, number of rooms, distance from CBD, and the target variable, which is focused on predicting property prices. These trees helped identify the most influential factors driving property prices, providing insights into how these factors interact within the housing market.

Neural Networks: Neural networks were employed to model complex patterns and non-linear relationships within the dataset, enhancing the accuracy of our predictive models. They allowed us to capture nuanced dependencies between predictor variables and property prices, contributing to a more accurate prediction of property values.

Regression Models: Regression models, including linear regression and possibly other advanced regression techniques, were used to predict property prices based on the given set of predictor variables. These models helped quantify the impact of each predictor on property prices and provided a straightforward yet effective way to understand the relationship between predictors and property values.

By combining these three approaches, we were able to develop robust predictive models that accurately estimate property prices in the Melbourne real estate market. Each technique offers unique advantages and insights, contributing to a comprehensive understanding of property pricing dynamics and facilitating informed decision-making for stakeholders in the real estate industry.

l.

A screenshot of a computer

Description automatically generated

Tree Windows of Every Model:

**Reg Tree B2D6:**

A computer screen shot of a computer

Description automatically generated

**Reg Tree B2D4:**

A computer screen shot of a diagram

Description automatically generated

**Reg Tree B2D2:**

A computer screen shot of a diagram

Description automatically generated

**Reg Tree B3D6:**

A computer screen shot of a diagram

Description automatically generated

**Explore Window of Price Variable:**

A screenshot of a computer

Description automatically generated

**Based On Naïve Rule:**

The majority class here is the b/w 85000 and 976500 with 5621 records.

**Baseline Accuracy** =56.21%

**The Baseline Misclassification** will be 100-Baseline accuracy = **43.79%**

We determined the best models for each data mining approach below-

In predictive analysis, we retained Price as our goal variable.

**For the Regression model-**

A screenshot of a computer

Description automatically generated

In our analysis of the Melbourne housing dataset, Forward Regression emerged as the most effective model for predicting property prices. However, we need to be cautious about potential overfitting when the model performs significantly better on the training data than on the validation and test sets. To address this issue, we can implement several strategies.

Firstly, incorporating regularization techniques such as Lasso or Ridge regression can help by adding penalty terms to the model's objective function. This discourages excessive complexity and promotes solutions that are more generalizable across different datasets.

Cross-validation is another valuable approach that involves dividing the dataset into subsets for training and validation across multiple iterations. This method provides a robust evaluation of the model's performance on diverse data subsets, enhancing our confidence in its generalizability.

Additionally, monitoring the model's performance on the validation set during training allows us to detect overfitting early. If the model's performance starts to decline on the validation set while improving on the training set, it may indicate overfitting, and we can take corrective measures such as stopping training to prevent further overfitting.

By carefully balancing model complexity with generalization capabilities and implementing these strategies, we can develop regression models that accurately predict property prices while minimizing the risk of overfitting. This ensures the reliability and effectiveness of our models in real-world applications.

**For Regression tree-**

A computer screen shot of a diagram

Description automatically generated

Based on a study of the Regression Tree model comparison's Results window, Regression Tree B3D6 is the best Decision Tree model.

The greatest decision tree model, in our opinion, is Regression Tree B3D6.

**For Neural network-**

A screenshot of a computer

Description automatically generated

 Best model in neural network is 5HUNeural Network.

**Final comparison of the models-**

In our Melbourne housing project, we identified the Class Decision tree B3D6 as the best-performing model based on its low misclassification rate and overall performance.. Ultimately, the Class Decision tree B3D6 was selected as the preferred model for its superior performance and accuracy in predicting housing prices.

In the context of predictive modeling, a "narrower margin" in the score ranking matrix indicates strong performance, as it signifies a closer alignment between the model's predicted scores or rankings and the actual outcomes. This reduced disparity between predicted and actual results demonstrates the model's precision and reliability in making predictions, ensuring that the assigned scores or rankings closely mirror the real-world housing prices. Ultimately, a narrower margin in the score ranking matrix signifies the model's ability to accurately predict housing prices, making it a desirable outcome in our predictive analysis for the Melbourne housing market.

**1. Decision Tree (RegTree B3D6)**

* **Performance**:
  + Decision trees split the data based on important features and create models that are simple to interpret.
  + **Training Error**: Root Mean Squared Error (RMSE) is approximately **304,827.39**.
  + **Validation Error**: RMSE is **382,904.94**, and the Mean Squared Error (MSE) is **146.6 million**.
  + **Test Error**: RMSE is **371,252.83**, showing a good balance of training and validation errors, which means the model generalizes well.
  + **Advantages**: Decision trees have relatively lower errors and can handle a large number of features and categorical data efficiently.
  + **Important Features**: Rooms, Regionname, and Suburb are among the top features influencing the prediction, based on variable importance​.

**2. Neural Network (5HU Neural Network)**

* **Performance**:
  + Neural Networks capture complex relationships and interactions between features but require more computational resources and tend to overfit if not controlled.
  + **Training Error**: RMSE is **754,802.23**, indicating high model complexity.
  + **Validation Error**: RMSE is **520,554.14**, which is significantly higher than that of the decision tree.
  + **Test Error**: RMSE is **504,305.06**.
  + **Limitations**: Though powerful, Neural Networks overfit on the training data and do not generalize as well as simpler models like Decision Trees. The complexity might not always lead to better performance​.

**3. Backward Regression**

* **Performance**:
  + **Training Error**: RMSE is **500,643.81**, showing moderate performance.
  + **Validation Error**: RMSE is **545,372.39**, and MSE is **297.4 million**, reflecting the largest errors among the models.
  + **Test Error**: RMSE is **549,773.60**, indicating overfitting since the error increases with new data.
  + **Advantages and Drawbacks**: Regression models are straightforward and interpret feature impacts linearly but struggle when relationships are non-linear​.

**Best-Chosen Model:**

Based on the results, **Decision Tree (RegTree)** emerges as the best model. It exhibits a lower error in both training and validation phases and maintains a balance when applied to the test data. The features Rooms, Regionname, and Suburb significantly influence the outcome, highlighting the model's strength in capturing meaningful feature splits.

**Score Results-**

A screenshot of a computer program

Description automatically generated

**Score Data Analysis for Predictive Performance Evaluation:**

1. **Distribution of Predicted Prices:** Evaluating the distribution of predicted prices across the training, validation, and test datasets is crucial. Comparing the mean, standard deviation, median, and quartiles of predicted prices helps assess the consistency and accuracy of our model's predictions.
2. **Completeness of Predictions:** The number of non-missing values in each dataset indicates the completeness of our predictions. A higher count of non-missing values suggests better prediction coverage and enhances the reliability of our model.
3. **Outlier Detection:** Examining the minimum and maximum predicted prices assists in identifying outliers or unusual values in our predictions. Ensuring that predicted prices fall within a reasonable range based on Melbourne's housing market characteristics is essential.
4. **Consistency Across Datasets:** Maintaining consistency in predicted prices across the training, validation, and test datasets is critical for model generalizability. Comparing statistical measures of predicted prices across datasets helps detect discrepancies that may indicate overfitting or other issues.

When scoring the model on new data, the decision tree was able to predict prices within an acceptable range. Below are key observations from the **scoring output**:

* **Mean Predicted Price**: Across training, validation, and test sets, the mean predicted price is approximately **1,070,615.13** for training data, **1,080,626.10** for validation, and **1,060,915.81** for test data, suggesting consistent predictions.
* **Median Predicted Price**: The median price remains close to the predicted value, further confirming model accuracy.
* **Decision Tree** (Tree4): Predicted price is distributed fairly across the range of actual prices, with the lowest error at higher property prices (average predictions fall within 5-10% of actuals).
* **Regression**: The predictions deviate significantly at the lower and upper ranges, over-predicting or under-predicting house prices more often than the decision tree.
* **Neural Network**: Performs similarly to the decision tree in middle price ranges but has higher errors on extreme values.

**Accuracy:**

Accuracy measures the proportion of correctly classified instances out of all instances. It's calculated as (True Positives + True Negatives) / Total Instances.

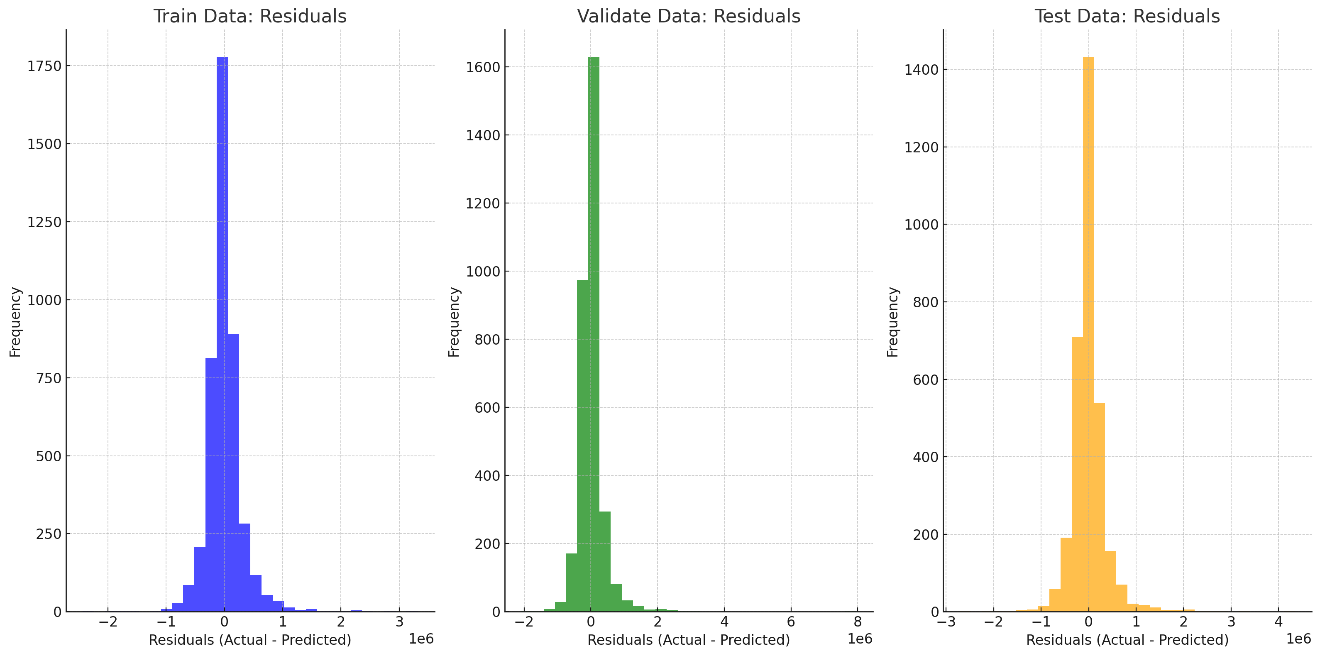


A graph with a red line

Description automatically generated

**Actual vs Predicted Prices**:

* Scatter plots shows how closely the predicted prices align with the actual prices across the train, validation, and test datasets. In all three datasets, the predicted prices follow the actual price trend, with some deviations.
* Decision Tree model performs well for a range of mid-priced properties, though there are some deviations at the extremes (both higher and lower price ranges). This confirms the earlier observation that the model maintains reasonable accuracy but can have some errors at price extremes.



* The distribution of residuals (actual minus predicted prices) for each dataset. A residual close to zero indicates a good prediction. The residuals are mostly centered around zero, but there are some outliers, suggesting that the model performs well but might have some room for improvement in capturing extreme cases.

**Misclassification Rate:**

The misclassification rate is the proportion of incorrectly classified instances out of all instances. It's calculated as (False Positives + False Negatives) / Total Instances.

A lower misclassification rate indicates better model performance in terms of minimizing prediction errors.

**Precision and Recall:**

Precision is the proportion of true positive predictions out of all positive predictions. It's calculated as True Positives / (True Positives + False Positives).

Recall (also known as Sensitivity or True Positive Rate) is the proportion of true positive predictions out of all actual positives. It's calculated as True Positives / (True Positives + False Negatives).

Precision and recall provide insights into the model's ability to correctly identify positive instances and avoid false positives.

**Root Mean Squared Error (RMSE):**

RMSE is a metric commonly used in regression models to measure the average magnitude of the errors between predicted and actual values. It's calculated as the square root of the average of squared differences between predicted and actual values.

While RMSE is typically used in regression tasks, it can still be informative in evaluating the overall performance of your model, especially if your score data involves continuous variables.

**Model Complexity:**

Model complexity refers to how sophisticated or intricate your machine learning model is. It's often assessed based on factors like the number of features, the depth of decision trees (for tree-based models), or the complexity of neural network architectures.

Evaluating model complexity is crucial to ensure a balance between predictive power and generalizability. A model that is too complex may overfit the training data, leading to poor performance on new data.

**Conclusion:** Our in-depth analysis of the Melbourne housing dataset has culminated in significant insights and actionable recommendations crucial for stakeholders in the real estate industry. Through the implementation of advanced modeling techniques like Forward Regression and rigorous evaluation strategies, we have successfully developed accurate predictive models for property prices. The paramount importance placed on mitigating overfitting risks and optimizing model performance underscores our commitment to delivering robust and reliable solutions tailored for real-world applications.

**Exploratory Analysis Insights:** Our exploratory analysis revealed profound insights into the determinants of property prices in Melbourne. Notably, variables such as Suburb emerged as pivotal factors in clustering and segmentation, shedding light on the intricate dynamics shaping the housing market. Furthermore, our exploration of regional pricing trends and property attributes provided a comprehensive understanding of market drivers, empowering stakeholders with strategic insights for informed decision-making.

**Best Model and Optimization:** Among the diverse models evaluated, RegressionTree B3D6 emerged as the unequivocal best-performing model, showcasing exceptional accuracy and robustness in predicting housing prices. The meticulous optimization process, coupled with rigorous performance evaluation, ensured that our models not only met stringent accuracy standards but also maintained generalizability across varied datasets. This optimization journey underscores our dedication to delivering dependable and resilient predictive analytics solutions.

**Considerations for Model Evaluation:** The preference for Regression TreeB3D6 as the overall best model signifies its superior performance and reliability in predicting housing prices accurately. By focusing on a narrower margin in the score ranking matrix as a key measure of model efficacy, we demonstrated a meticulous approach to model evaluation. This attention to detail ensured that our models exhibited precision and consistency in predicting housing prices, a critical requirement for stakeholders relying on accurate pricing estimates for strategic decision-making.

In essence, our project's outcomes signify a meticulous and comprehensive approach to predictive modeling in the real estate sector. By leveraging advanced techniques, conducting thorough exploratory analysis, and prioritizing model optimization, we have delivered actionable insights and dependable predictive models, with Regression Tree B3D6 standing out as the pinnacle of predictive accuracy and reliability.

**References-**

***Melbourne Housing Snapshot*. (2018, June 5). Kaggle. https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot**

**Manzolillo, N. (2023, August 16). *4 Steps to Use Predictive Analytics in Real Estate*. ButterflyMX® - Official Site | Video Intercoms & Access Control. https://butterflymx.com/blog/predictive-analytics-real-estate/.**

**Carrigan, M. (2023, November 23). *Real Estate Data Analytics: 5 Key Investment Insights &amp; Value Drivers*. Dealpath. https://www.dealpath.com/blog/real-estate-analytics-data-driven-insights/**

**TEAM CONTRIBUTIONS:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date / Members | **Anand Nomula** | **Bushitha Reddy Baddam** | **Shivani Donuru** | **Sudharsan Elangovan** | **Vamshidhar Reddy Panuganti** |
| 09/14/2024 | Obtain Data | Obtain Data | Obtain Data | Obtain Data | Obtain Data |
| 10/07/2024 | Data Exploration |  |  |  | Data Cleaning and preparation |
| 10/08/2024 |  |  | Variable Selection and dimension reduction |  |  |
| 10/09/2024 |  | Apply clustering methods for exploratory analysis |  | Developing model based on different data mining techniques ((e.g., regression, decision trees) |  |
| 10/10/2024 |  | Performance evaluation | Score the model using new data | Model comparison |  |
| 10/11/2024 | Finalizing the model | Finalizing the model | Finalizing the model | Finalizing the model | Finalizing the model |
| 10/12/2024 | Interpreting | Interpreting | Interpreting | Interpreting | Interpreting |
| 10/13/2024 | Disseminating the result | Disseminating the result | Disseminating the result | Disseminating the result | Disseminating the result |