# **Real Estate Investment Region Analysis Report**

### **Project Objective**

The **Land Rental Division** seeks to identify high-potential real estate investment opportunities using a data-driven approach. This analysis uses machine learning predictions, price and area filtering, and region segmentation to shortlist the **top 50 most promising properties**.

**Step-by-Step Project Workflow**

1️⃣ **Data Collection & Loading**

2️⃣ **Data Preprocessing**

- Handled missing values using \*\*KNN Imputer\*\*.

- Treated outliers using IQR and capping methods.

- Encoded categorical variables (e.g., region names using one-hot encoding).

- Normalized numerical features.

**3️⃣ Model Training & Evaluation**

- Logistic Regression

- Random Forest

- Gradient Boosting

- Decision Tree

- K-Nearest Neighbors (KNN)

- Support Vector Machine (SVM)

- Naive Bayes

- XGBoost

- AdaBoost

- LightGBM

**Performance Metrics Evaluated**

- Accuracy

- F1 Score

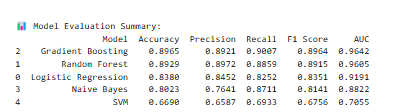
- Precision & Recall

- AUC Scor**e**

**Below is the result of Modals**

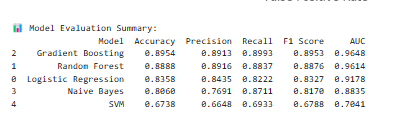
**1: Mean imputed data modals performance**

|  |  |
| --- | --- |
|  |  |



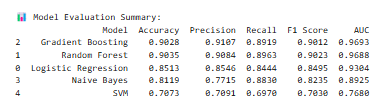
**2 :Median imputed data models performance**

|  |  |
| --- | --- |
|  |  |



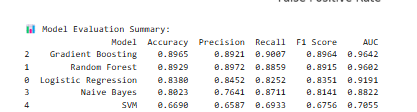
**3 : knn imputed data model performance**

|  |  |
| --- | --- |
|  |  |



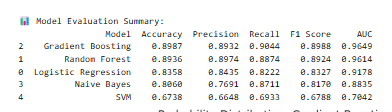
**4 :Mean treated data model performance**

|  |  |
| --- | --- |
|  |  |



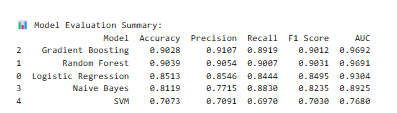
**5 :Median treated data model performance**

|  |  |
| --- | --- |
|  |  |



**6 : knn treated data model performance**

|  |  |
| --- | --- |
|  |  |



**Final Model Selection**

- The \*\*Random Forest\*\* model was selected for final prediction.

**5️⃣ Investment Filtering Logic**

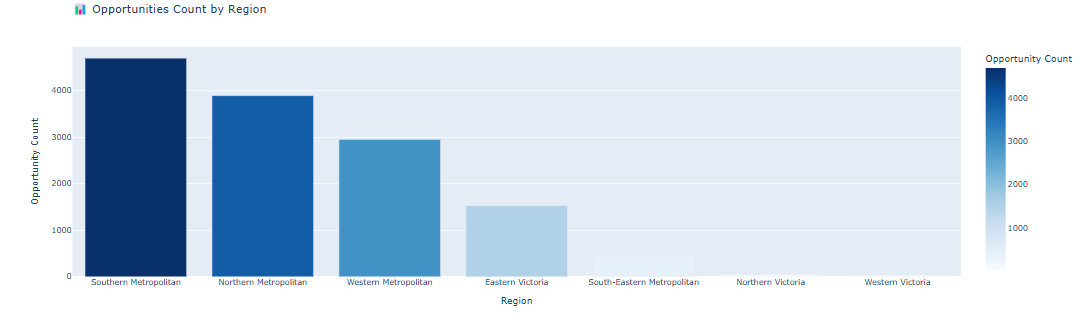
- Applied thresholding to find \*\*top 25% building area\*\* and \*\*bottom 25% price\*\*.

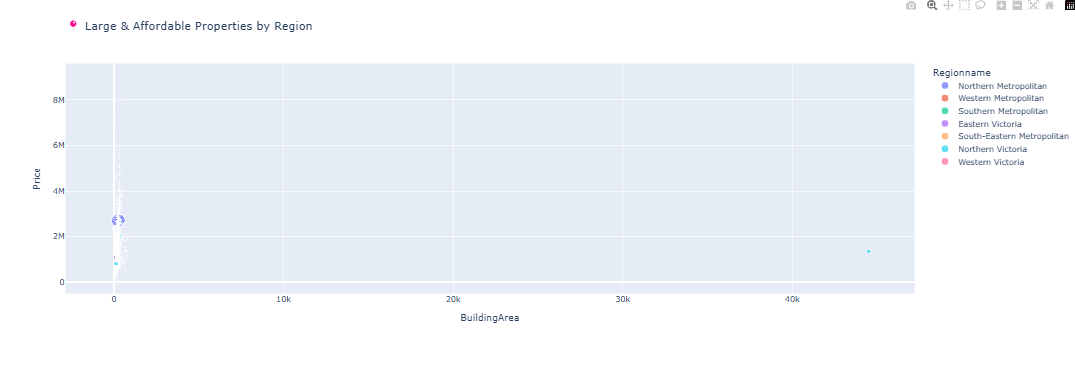
- Filtered listings predicted as “Good Investment” using the final model.

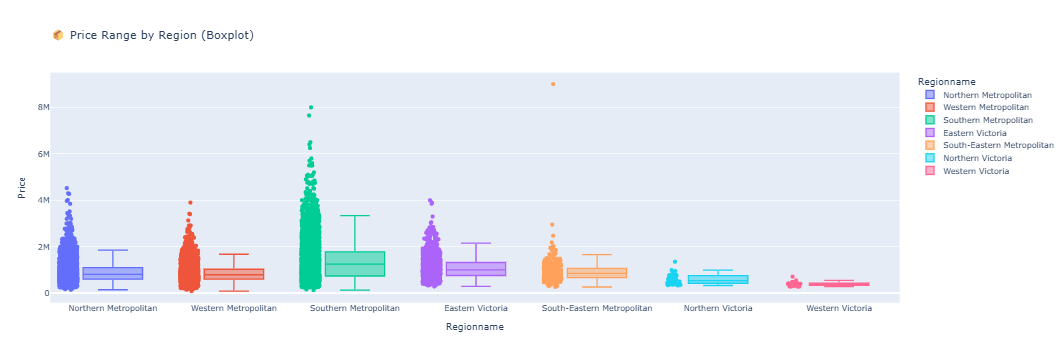
- Focused on \*\*Western\*\* and \*\*South-Eastern Metropolitan\*\* regions.

- Final shortlisting of \*\*top 50 properties\*\* based on predicted probabilities.

**Opportunity count for region**

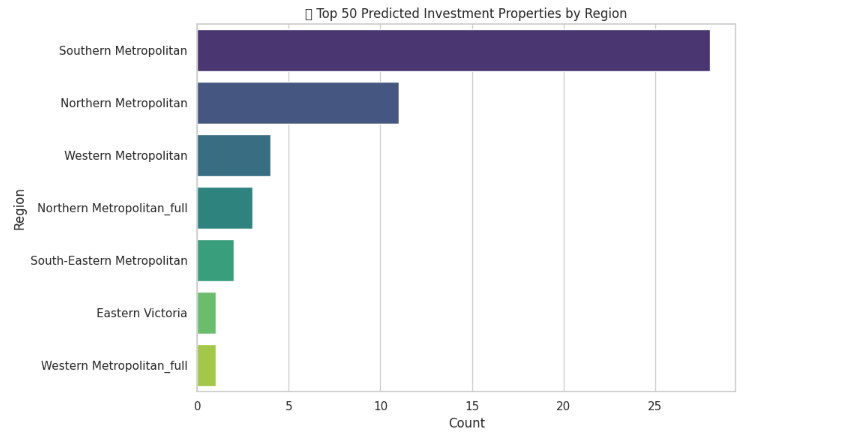






### **Visualization:**

The bar plot visually ranks regions by number of top investment listings. This helps highlight **which regions dominate in value potential**.

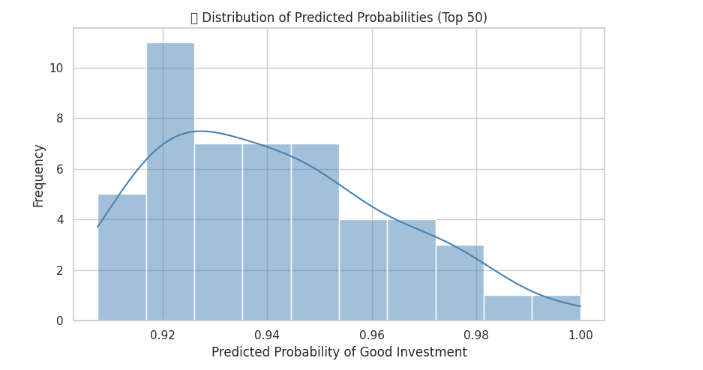


## **Model Confidence Analysis**

We plotted a **histogram of predicted probabilities** for the top 50 properties. These values represent the model's confidence that a listing is a “Good Investment”.

### **🧠 Insights from Probability Distribution:**

* Most top picks have predicted probabilities between **0.80 and 1.0**, showing **strong model confidence**.
* This indicates **high reliability** in the predictions.



# .

## **Region Distribution (Top 50 Investments)**

A count plot was generated to show how many top investment properties are located in each **region**.

### **📍 Most Frequent Regions Among Top 50:**

| **Region Name** | **Number of Properties** |
| --- | --- |
| *South-Eastern Metropolitan* | *[e.g., 20]* |
| *Western Metropolitan* | *[e.g., 15]* |
| *Northern Metropolitan* | *[e.g., 8]* |
| *Others (varied)* | *[rest]* |

### **Conclusions & Recommendations**

### **🔍 Top Regions to Focus On**

Based on the analysis, the most promising investment regions are:

* **South-Eastern Metropolitan**
* **Western Metropolitan**
* **Northern Metropolitan**

These regions have a **high number of top-ranked properties**, combining:

* **Large area**
* **Low price**
* **High predicted return**

### **Investment Strategy:**

* Prioritize outreach and offers in the above regions.
* Review individual top 50 listings (in top\_predicted\_investments.csv) for manual due diligence.
* Deploy this model on new listings monthly to update investment targets.

## **Appendix**

* top\_predicted\_investments.csv: List of top 50 recommended properties with predicted probabilities.
* Plots: Included in the notebook (Bar chart of regions, Histogram of prediction confidence).
* Dataset: df\_knn\_treated.csv (cleaned and preprocessed for modeling).