Housing Price Prediction with the Charasteristics of Properties in Melbourne

Ali Aytuğ Tok
Middle East Technical University

Ankara, Turkiye

Abstract— **This study aims to construct and compare machine learning models to classify housing prices in Melbourne regarding being "cheap" and "expensive" considering many features. To be able to handle the data in more capability, EDA process was applied first via packages GGally and ggplot2 in R Programming environment, regarding to the research questions raised before the study. After EDA stage, CDA was applied to ensure the claims about the data is satisfied via statistical tests. After preprocesses such as outlier deletion, imputation, one-hot encoding, transformations with best possible option, and scaling, the target value (Price) had been binarized using the median of the logarithmic price values because of the normality problem. After this procedure, several models were built such as Binary Logistic Regression, ANN, SVM, Random Forest and XGBoost. The model performances were evaluated with cross-validation and metrics as accuracy, F1 Score, precision, recall and Kappa score. Moreover, these models were tuned via GridSearchCV and Optuna in the environment of Python.**

Keywords— Price Classification, K-Fold Cross-Validation, Machine Learning, Binary Logistic Regression, ANN, SVM, Random Forest, XGBoost

I. Introduction

Housing prices in cosmopolitan regions like Melbourne show differences significantly depending on several factors such as type and physical attributes of the property. Categorizing these properties into meaningful factors as "cheap" or "expensive" can assist for giving decisions more accurately for buyers, sellers, and also policymakers. In statistical perspective, since the housing prices are continuous, converting them into binary classes based on statistical threshold, in this project the median, enables for effective classification and modelling. To satisfy robustness, this transformation was anticipated by data preprocessing, including exploratory data analysis (EDA), statistical validation over confirmatory data analysis (CDA), outlier deletion, imputation, encoding, transformation.

Afterwards, several machine learning models were built to make predictions for price classes, including Binary Logistic Regression, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, and XGBoost. These models were built with the specific cross-validation method named Stratified K-Fold Cross-Validation. Even this method is used for unbalanced data, it is still recommended to show robustness and consistency. The built models are consisting Binary Logistic Regression, ANN, SVM, RF and lastly XGBoost.

II. Literature Review

In tasks involving the prediction of home prices, recent studies have demonstrated that machine learning (ML) algorithms provide notable benefits over conventional statistical techniques. Using the popular Ames Housing dataset, Zhang, Lyu, and Yu (2023) carried out a thorough assessment of several machine learning models, such as Ridge Regression, K-Nearest Neighbors (KNN), Random Forest, and XGBoost. According to their research, when it comes to prediction performance, XGBoost regularly produces the most accurate outcomes. Additionally, the study emphasized the benefits of interpretability provided by SHAP (SHapley Additive exPlanations), which enable a clearer comprehension of feature contributions. In a more general setting, Ritu (2023) carried out a comprehensive analysis of over 20 research papers concentrating on machine learning methods for home price prediction. According to evaluation metrics like Root Mean Square Error (RMSE) and R-square (R2), this review consistently demonstrated the superiority of community learning techniques, particularly Random Forest and XGBoost. The author has underlined how these models' adaptability and flexibility make them particularly well-suited for intricate, nonlinear models, which are frequently present in real estate datasets. These studies support the expanding body of evidence showing that incorporating cutting-edge machine learning techniques can greatly increase the precision and dependability of home price projections.

III. Methodology

A. Dataset

The dataset used in this study was retained from Kaggle.com and it is publicly available real estate data. It involves more than 27,000 property records with large features in terms of physical, geographical, and transactional features. In details, the data includes suburb, price, number of rooms-bedrooms-bathrooms-car spaces, type of the property, sales method, council area and seller agency. In order to perform classification tasks and normalize the original target variable (Price), it is log-transformed and binarized. Despite being transformed, the target variable did not become normal, thus the binarization process was based on the median value of the Price variable and distinguished between “cheap” and “expensive”. For accurate analysis, missing value analysis and imputation, one-hot-encoding, scaling, and outlier deletion have been applied.

The dataset contains locational and structural information about the houses sold in Melbourne with several categorical and continuous variables related to the property traits. They capture crucial attitudes such as physical size, accessibility etc. These characteristics are listed as:

- Suburb – categorical, location of the property in Melbourne
- Rooms – Discrete, number of rooms
- Bedroom – Discrete, number of bedrooms
- Bathroom – Discrete, number of bathrooms
- Car – Discrete, number of car spaces
- Distance – Continuous, distance to the center of Melbourne in kilometers
- BuildingArea – Continuous, area of the property in square meters
- YearBuilt – Discrete, year the property was built
- Method – Categorical, method of sale
- SellerG – Categorical, selling real estate agency
- CouncilArea – Categorical, the municipal council of the property
- Regionname – Categorical, broader region classification
- Price – Continuous, selling price of the house in Australian Dollars.
- BuildingArea_log – Continuous, log-transformed version of BuildingArea for scaling
- Price_log – Continuous, log-transformed version of Price for normalization
- Price_Class – cheap = 0, expensive = 1

B. *Descriptive Statistics*

Descriptive statistics has an important role to illustrate the insights for the data. It is conducted to summarize the central tendencies and distributions of the variables in the data.

Continuous variables Price_log, BuildingArea_log, Distance were searched for Min., 1st Quantile Range, Median, Mean, 3rd Quantile Range and Max. and the results are obitaned as following:

TABLE I. SUMMARY OF CONTINUOUS VARIABLES

	Price_log	BuildingArea_log	Distance
Min.	11.35	0.010	0.00
1st Qu.	13.36	4.635	6.40
Median	13.68	4.920	10.30
Mean	13.72	4.917	11.18
3rd Qu.	14.07	5.242	14.00
Max.	16.23	16.23	48.10

Discrete variables Rooms, Bedroom, Bathroom, Car, YearBuilt also were searched for Min., 1st Quantile Range, Median, Mean, 3rd Quantile Range and Max. and the results are obitaned as following:

TABLE II. SUMMARY OF DISCRETE VARIABLES

	Rooms	Bedroom	Bathroom	Car	YearBuilt
Min	1.000	0.000	0.000	0.00	1196
1st Qu	2.000	2.000	1.000	1.00	1940
Median	3.000	3.000	2.000	2.00	1970
Mean	3.031	3.085	1.625	1.73	1965
3rd Qu	4.000	4.000	2.000	2.00	2000
Max	16.000	30.000	12.000	26.0	2019

Table 2: Summary of discrete variables

As mentioned, Price variable has been binarized for constructing Logistic Regression Model and Machine Learning algorithms. A summary for this variable can be observed at figure 1.

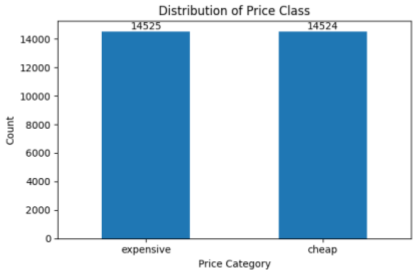
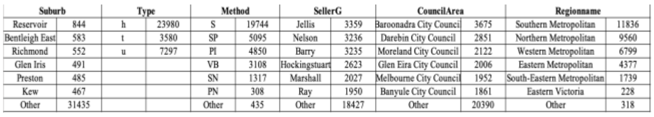


Fig. 1. Bar Plot

For the categorical variables in the dataset, the frequencies have been captured as seen at the table 3:

TABLE III. SUMMARY OF CATEGORIC VARIABLES



Suburb		Type		Method		SellerG		CouncilArea		Regionname	
Reservoir	844	h	23980	S	19744	Jellis	3359	Baroonadra City Council	3675	Southern Metropolitan	11836
Bentleigh East	583	t	3580	SP	5095	Nelson	3236	Darebin City Council	2851	Northern Metropolitan	9560
Richmond	552	u	7297	PI	4850	Barry	3235	Moreland City Council	2122	Western Metropolitan	6799
Glen Iris	491			VB	3108	Hockingstuart	2623	Glen Eira City Council	2006	Eastern Metropolitan	4377
Preston	485			SN	1317	Marshall	2027	Melbourne City Council	1952	South-Eastern Metropolitan	1739
Kew	467			PN	308	Ray	1950	Banyule City Council	1861	Eastern Victoria	228
Other	31435			Other	435	Other	18427	Other	20390	Other	318

There are so many observations among categorical data. The sum of the frequencies of each variable is not equal each other. The reason is that there are NA values in the preprocessed data. Therefore, as will be mentioned following, the imputation techniques and also outlier deletion were applied. Moreover, some of these variables were not used in the analysis for being redundant.

C. *Exploratory Data Analysis*

In this part of the study, 5 research questions have been raised to handle the structure of the data and relationship between the variables.

1) Is there a significant difference in the sales price among different type of the properties?

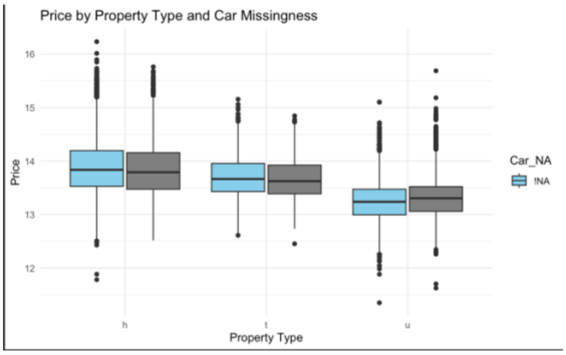


Fig. 2. Box Plot

In the plot, the distribution of the *Price* is seen with and without missing values of the *Car*. X-axis denotes the property type of the houses: *house (h)*, *townhouse (t)* and *unit (u)*. Both distributions of the missing and non-missing values can be seen on the plot. Considering non-missing values, there are visible extreme outliers for *house (h)* variable, compared to *townhouse (h)* and *unit (u)*. *t* and *u* variables show tighter distributions with lower medians compared to *h* and less extreme outliers as mentioned.

Since the transformed Price is still not normal, for comparison, Kruskal-Wallis test is applied instead of ANOVA.

TABLE IV. KRUSKAL-WALLIS TEST RESULT

Kruskal-Wallis Rank Sum Test		
Price_log v. Type		
Chi-Sq =7598.2	df = 2	p-value = <2.2e-16

According to the test results, low p-value indicates that type of the property significantly has an impact on price.

TABLE V. PAIRWISE COMPARISON TEST SUMMARY

Comparison	Z	P.unadj	P.adj
h-t	25.40075	2.47e-142	7.422e-142
h-u	86.71754	0.000	0.000
t-u	35.67237	1.06e-278	3.18e-278

The post-hoc Dunn test indicates that median prices show difference between all types with all adjusted p-values which are very low compared to significance level.

2) Does the age of the properties have an impact on housing price?

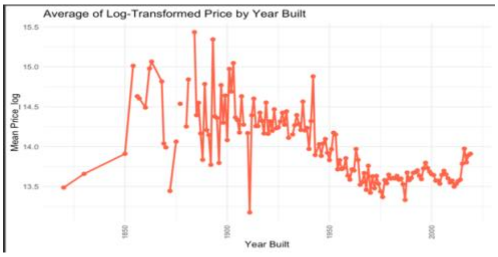


Fig. 3. Line Chart

The plot tells properties built before 1900s show dramatically high average prices. It can be associated with heritage or properties with special characteristics. These properties may be gained their values according to their architectural charm, larger land sizes or special locations.

After 1900s until mid-century, there is a significant decline in average price. From the 1980s until today, the stabilization is seen. The reason may be standardization and suburban expansion. Moreover, there is a small increase in recent years due to growing value in modern and developed properties with the advancement of technology.

There are splits in the plot, indicates there are missing values. This problem had been solved with imputation methods, which is mentioned in the following of the report.

TABLE VI. SPEARMAN'S RANK CORRELATION RHO

Spearman's Rank Correlation Rho	
Data: YearBuilt and Price_log	
S = 5.6656e+12	p-value < 2.2e-16
Rho = -0.3865276	

The test shows a significant negative relationship between YearBuilt and Price_log. It means that older homes yield to have higher average price compared to new ones.

3) How does the building area of the property effect its market price?

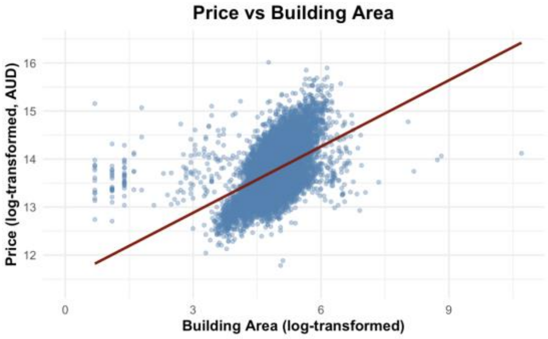


Fig 3. Scatter Plot with Regression Line

This scatter plot investigates the relationship between building area and property price. Both variables are log-transformed to be scaled and also for normality for the next process of the study. Each dot indicates a property.

As building area increases, there is an obvious upward trend in price, illustrated by the fitted regression line. In other words, larger homes usually present higher prices as expected. However, there is a surprising insight, which is the tight clustering in the middle of the chart. It gives a hint about the core of the Melbournian Market: standard-sized properties where price differences are less about raw size, but more about other features such as location, condition and so on.

When the outliers considered, tiny properties are surprisingly expensive than the large ones. Some of the features may led this outcome for the market price.

TABLE VII. SPEARMAN'S RANK CORRELATION RHO

Spearman's Rank Correlation Rho	
Data: BuildingArea_log and Price_log	
S = 1.4177+12	p-value < 2.2e-16
Rho = 0.6529818	

According to the test, significant positive relationship between building area and price has been observed. It means, larger homes tend to have higher price.

4) How do the building area and distance to the city center effect price classification of the property, moreover what is the role of the counts of the room in this relation?

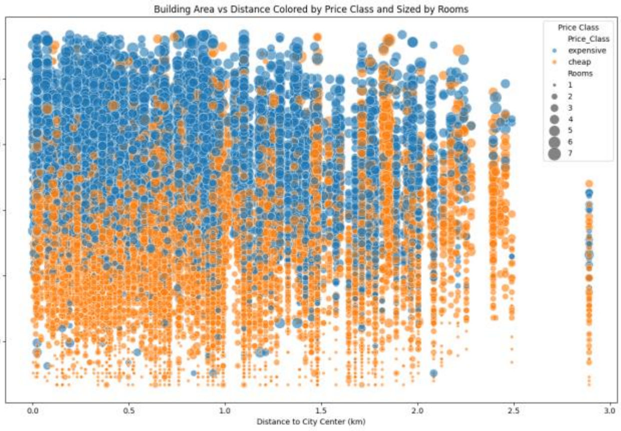


Fig. 4. Bubble Plot

For the x-axis of the plot, the absolute value of the Distance has been taken to use Distance as a comparison, not direction. Bubble size denotes the number of the rooms, and the color indicates whether the house is expensive or cheap.

It can be concluded that larger homes with more rooms are dominantly classified as expensive, especially the ones are close to the city center. Depending, cheaper properties tend to be smaller and far from the city center. But, the overlap tells that distance solely is not explaining the whole pattern. Size and number of rooms are the key features for this case.

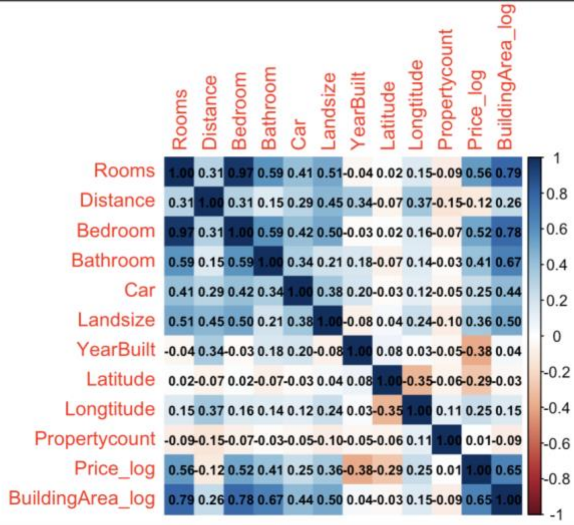
Although transforming continuous variables, they did not become normal according to the Shapiro-Wilk test. Thus, non-parametric tests should be taken into account. In the light of this fact, Aligned Rank Test is conducted for confirmation:

TABLE VIII. ALIGNED RANK TEST RESULTS

Analysis of Variance of Aligned Rank Transformed Data				
Response: art(Distance)				
	Df	Df.res	F value	Pr(<F)
Price_Class	1	29035	9.2097	0.0024
as.factor(Rooms)	6	29035	423.4455	2.22e-16
Price_Class: as.factor(Rooms)	6	29035	30.4078	2.22.e-16

The test shows that Price_Class, Rooms and the interaction effect show that they are significantly effecting the distance, means that number of rooms influences the location different for cheap and expensive properties.

For the following process, the assumptions should be held. In the light of this goal, correlation matrix is an important tool to build models. It can help to avoid multicollinearity, which can lead misleading results.



	Rooms	Distance	Bedroom	Bathroom	Car	Landsize	YearBuilt	Latitude	Longitude	Propertycount	Price_log	BuildingArea_log
Rooms	1.00	0.31	0.97	0.59	0.41	0.51	-0.04	0.02	0.15	-0.09	0.56	0.79
Distance	0.31	1.00	0.31	0.15	0.29	0.45	0.34	-0.07	0.37	-0.15	-0.12	0.26
Bedroom	0.97	0.31	1.00	0.59	0.42	0.50	-0.03	0.02	0.16	-0.07	0.52	0.78
Bathroom	0.59	0.15	0.59	1.00	0.34	0.21	0.18	-0.07	0.14	-0.03	0.41	0.67
Car	0.41	0.29	0.42	0.34	1.00	0.38	0.20	-0.03	0.12	-0.05	0.25	0.44
Landsize	0.51	0.45	0.50	0.21	0.38	1.00	-0.08	0.04	0.24	-0.10	0.36	0.50
YearBuilt	-0.04	0.34	-0.03	0.18	0.20	-0.08	1.00	0.08	0.03	-0.05	-0.38	0.04
Latitude	0.02	-0.07	0.02	-0.07	-0.03	0.04	0.08	1.00	-0.35	-0.06	-0.29	-0.03
Longtitude	0.15	0.37	0.16	0.14	0.12	0.24	0.03	-0.35	1.00	0.11	0.25	0.15
Propertycount	-0.09	-0.15	-0.07	-0.03	-0.05	-0.10	-0.05	-0.06	0.11	1.00	0.01	-0.09
Price_log	0.56	-0.12	0.52	0.41	0.25	0.36	-0.38	-0.29	0.25	0.01	1.00	0.65
BuildingArea_log	0.79	0.26	0.78	0.67	0.44	0.50	0.04	-0.03	0.15	-0.09	0.65	1.00

Fig. 5. Correlation Matrix

According to the correlation matrix, there are some variables showing very high correlation, such as Bedroom and Rooms. For building models, this case has to be considered.

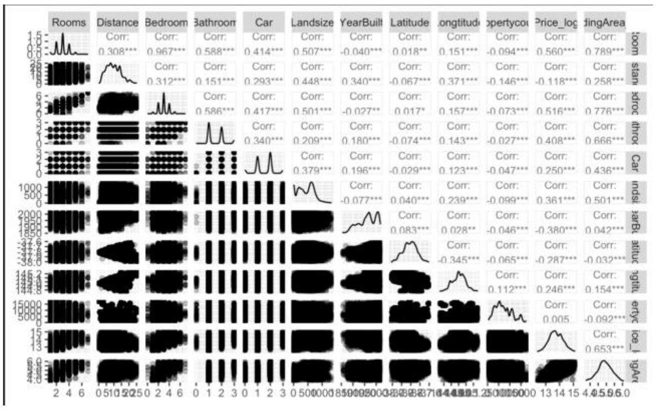


Fig. 6. Correlogram of the dataset

As can be seen from the correlogram, there are some variables highly correlated. To prevent any kind of problem, dimension reduction can be considered as a solution.

D. Missingness

The data has missing values in its default. The ratio observed as 16.3% missing and 83.7% present. For determining the pattern of the missing values, MCAR test were conducted:

TABLE . MCAR TEST SUMMARY

statistic	df	p.value	missing.patterns
3685.734	390	0	48

Since the p-value is smaller than the significance level, MCAR case has to be dismissed, the pattern has to MAR (Missing at Random) or MNAR (Missing Not At Random). With the help of the visualization, the pattern can be observed much better.

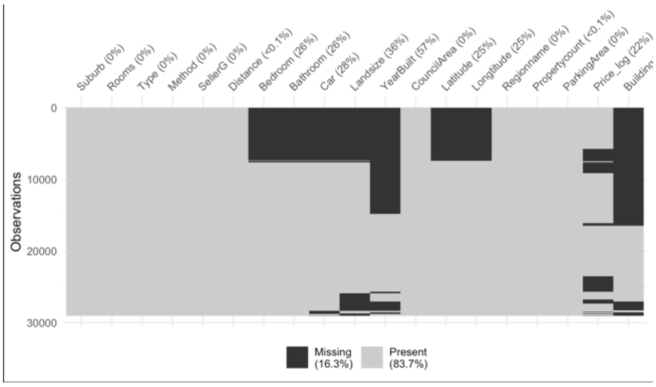


Fig. 7. Missingness Plot

According to the plot, missing values are located within specific variables, however not randomly splitted across observations. Depending on, the mechanism is very close to Missing at Random (MAR).

After deciding the pattern mechanism, the accurate imputation techniques were applied. For numeric variables, PMM method was conducted. The nominal variables were imputed with polynomial regression technique.

After the imputation method the summary statistics of preprocessed data and imputed data were compared. At the result of this comparison, the summary statistics were remained very close to each other. As a result, the imputation process was successful. Here is an example of the outcome:

TABLE IX. BEFORE AND AFTER IMPUTATION COMPARISON

Variable	Metric	Preprocessed	Imputed
Price_log	Mean	13.72	13.78
	Median	13.68	13.75
Building_Area_log	Mean	4.917	4.860
	Median	4.920	4.860

To ensure if the imputation was successfully held, missingness pattern had been checked again.

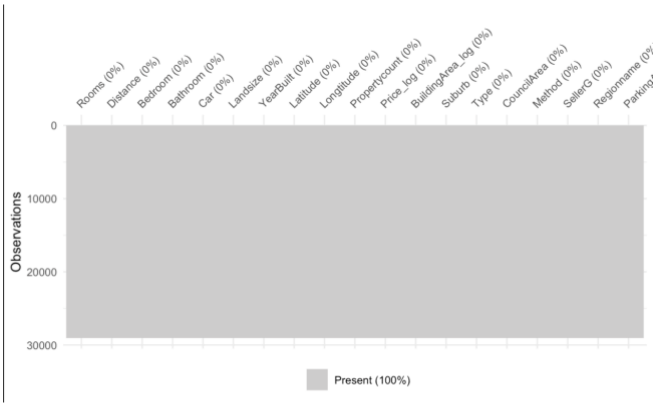


Fig. 8. Missingness Plot after imputation

E. Modelling

1. Principal Component Analysis (PCA)

Due to having highly correlated variables in the dataset, dimension reduction can be applied. To perform the task, PCA was applied. It is a technique that transforms highly correlated variables into a smaller set of uncorrelated features while saving as much variance in the dataset.

To find the optimum number of components for the process, scree plot is a helpful tool to investigate the cut off point for the next stage.

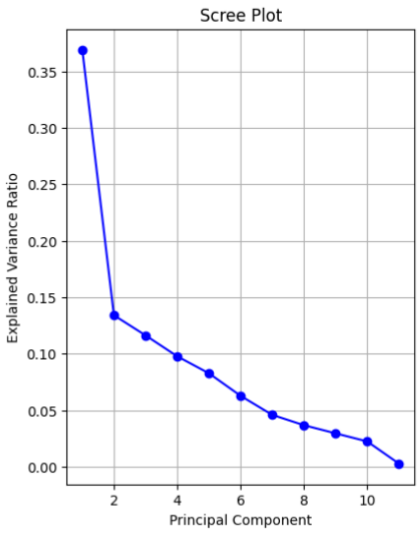


Fig. 8. Scree Plot

According to the plot, the elbow point is 2. The slight decrease begins after PC5. To ensure, cumulative explained variance table is analyzed.

TABLE X. EXPLAINED VARIANCE TABLE

Explained Variance Table	
Up to PC1	0.3690
Up to PC2	0.5031
Up to PC3	0.6192
Up to PC4	0.7170
Up to PC5	0.7997
Up to PC6	0.8624
Up to PC7	0.9084
Up to PC8	0.9451
Up to PC9	0.9747
Up to PC10	0.9972
Up to PC11	1.0000

Since the %86.24 can be explained with PC6, 6th component is chosen for PCA. The feature importance plot for this procedure can be seen in the next figure:

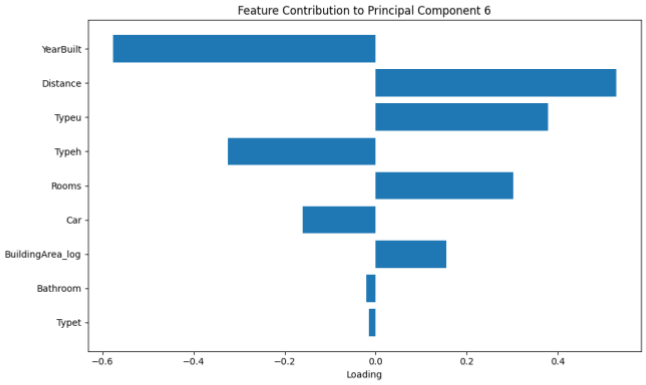


Fig. 9. Feature Importance for PC6

According to the figure, the most influential features are YearBuilt and Distance but in the opposite directions. It can be concluded that PC6 captures opposition between older houses and their proximity to the city.

TABLE XI. CONTRIBUTION TABLE FOR PC6

Variables	Contribution to PC6
Typeh	0.041
Typet	0.000
Typeu	-0.047
YearBuilt	0.045
Distance	0.052
BuildingArea_log	0.121
Bathroom	0.002
Car	0.326
Rooms	0.029

Cross – Validation

Since the target variable is not normally distributed, the threshold has been chosen as the median. When lower and higher values corresponding to the median considered, the distribution was almost equal. Thus, the data is balanced.

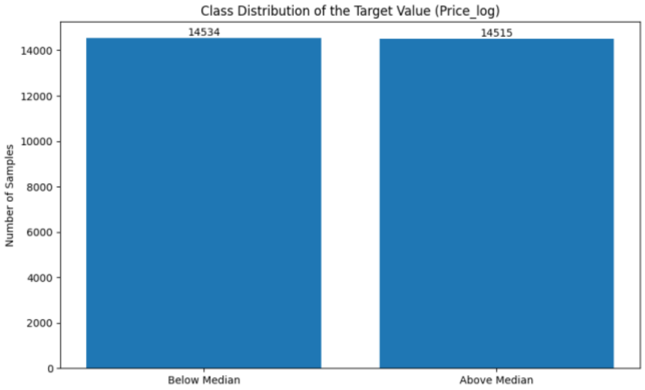


Fig. 10. Distribution Chart of the Target Value

At the following stage, since the classification process had been applied, the target value classified and binarized as lower values as “cheap” and encoded as 0, higher proportions labeled as “expensive” and encoded as 1. Regarding to this operation, Stratified K-Fold Cross-Validation technique has been applied to the data. After the process, the distribution of the train and test data has been captured as seen on the next figure:

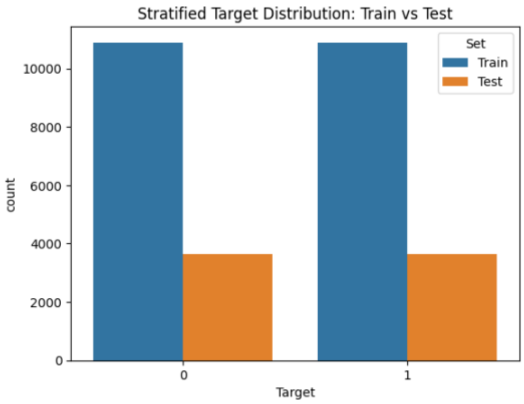


Fig. 11. Clustered Bar Plot

2. *Binary Logistic Regression*

As mentioned previously, to conduct this model, target value has been binarized and classified. Since the variables are not normal distributed, this model was accurate to build. After the PCA process, the multicollinearity problem was removed. Thus, the model summary occurred as below:

Price = -0.064978 +
1.769 x BuildingArea_log +
-1.095627 x YearBuilt +
-0.858397 x Distance +
-0.326834 x Typeu +
0.285045 x Bathroom +
0.247642 x Car +
0.237815 x Typeh +
0.075239 x Typet +
0.022552 x Rooms

According to the model, the most impact on target provided from building area. Also, the age of the property is also crucial for the clients. Moreover, distance also have a negative impact on price. The model metrics can be seen below:

TABLE XII. TEST SET METRICS FOR LOGISTIC REGRESSION MODEL

Test Set Metrics for Logistic Regression	
Accuracy	0.8060
Precision	0.7984
Recall	0.8188
F1 Score	0.8085
Kappa	0.6120

3. *Artificial Neural Network*

Before conducting the model, tuning method was applied with using Optuna, in Python environment. The model indicated strong prediction performance on the test set, accessing the accuracy of 82.53%, the model truly classified an important majority of instances. The precision of 0.8261 indicates that the model predicts the class (cheap or expensive) true almost %83 of the time. In parallel, the recall value of 0.8241 indicates that it successfully classifies nearly 82% of all expensive houses. The F1 score, which enables the balance between precision and recall, is 0.8251, shows consistent performance across both classes. The Kappa value, 0.6506, indicates that the model catches meaningful patterns between predictions and actual values. Shortly, the model was successful.

TABLE XIII. TEST SET METRICS FOR ANN MODEL

Test Set Metrics for ANN Model	
Accuracy	0.8253
Precision	0.8261
Recall	0.8241
F1 Score	0.8251
Kappa	0.6506

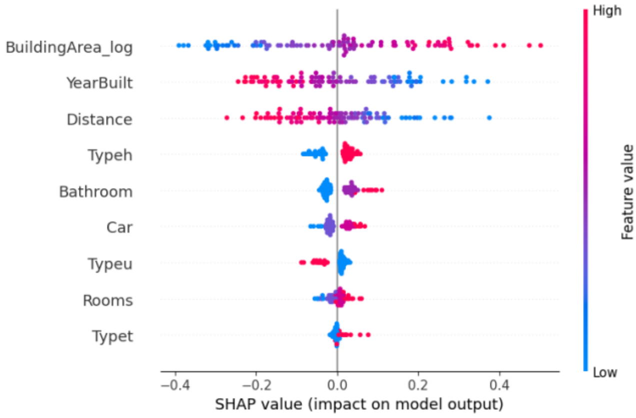


Fig 12. SHAP Plot

The SHAP Plot demonstrates the most impactful features on target in the ANN Model. Building Area, Year Built and Distance have the most impact according to the plot. Higher building area and newer property years are associated with a greater likelihood of a property classified as expensive. On the other hand, greater distance from the city center leads the price labeled as cheap.

4. *Support Vector Machines (SVM)*

This model has been tuned via GridSearchCV in the environment of Python with the best possible parameters. The SVM model was also successful according to the metrics. The accuracy of 82.25% was obtained, the model accurately classified the prices in this rate. The precision of 0.8157 means 81.6% of the houses predicted as expensive. The F1 Score of 0.8245 confirms that the model has consistent classification performance. The Kappa value of 0.645 indicates substantial agreement between predicted and true values, beyond what would be expected by chance.

TABLE XIV. TEST SET METRICS FOR SVM MODEL

Test Set Metrics for SVM Model	
Accuracy	0.8225
Precision	0.8157

Recall	0.8334
F1 Score	0.8245
Kappa	0.6450

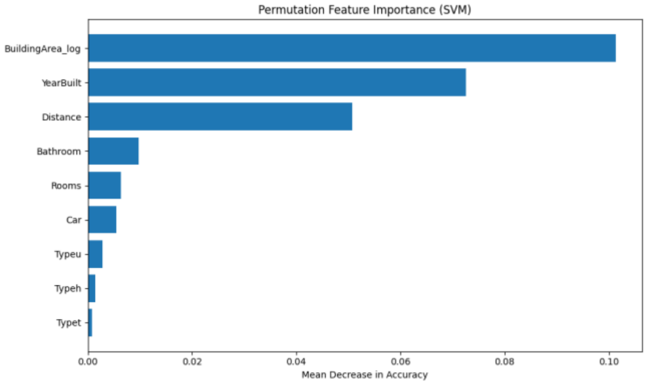


Fig. 13. Bar Plot

According to the feature importance plot, building area has the most impact on the target. Moreover, the age of the property is also effective. Distance is also one of the most important features in pricing.

5. *Random Forest (RF)*

This model is also tuned with GridSearchCV to construct the model with the best parameters possible. The RF model was also successful with the accuracy of 83.56%, over than 83% of the predictions were accurate. The precision of 0.8340 tells us that the model predicted a house as expensive, it was correct 83.4% of the time. The recall of 0.8381 means the model truly identifies 83.8% of all correctly expensive properties. The F1 score of 0.8360 confirms the model is consistent for distinguishing between price classes. The Kappa score of 0.6712 shows substantial agreement beyond chance between predicted as actual price class.

TABLE XV. TEST SET METRICS FOR RF MODEL

Test Set Metrics for RF Model	
Accuracy	0.8356
Precision	0.8340
Recall	0.8381
F1 Score	0.8381
Kappa	0.6712

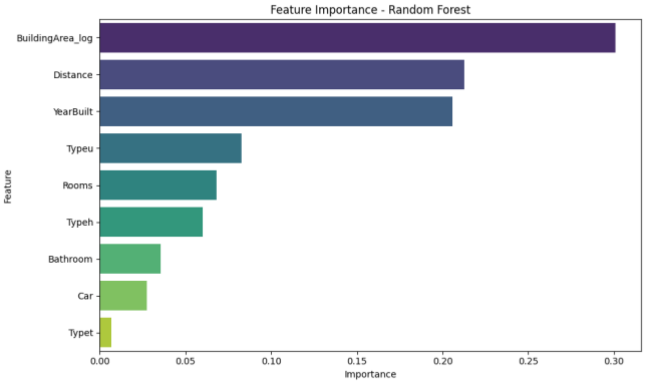


Fig. 14. Bar Plot

According to the feature importance plot, the most impactful features are building area, distance and year built (age of the property) respectively.

6. *XGBoost*

The model is tuned with GridSearchCV to build the model with the optimum parameters. This model shows the greatest performance among other machine learning models. The accuracy of 85.3% tells us the model classifies a large majority of instances correctly. The precision of 84.9% indicates that when the model predicts the class "expensive", most of the times, it is true. The recall value indicates that the model successfully identifies most actual "expensive" classes. The F1 score is the proof of the model performs well in both minimizing false positives and false negatives. The Kappa score of 0.7059 shows substantial agreement beyond chance between predicted as actual price class.

TABLE XVI. TEST SET METRICS FOR XGBOOST MODEL

Test Set Metrics for XGBoost Model	
Accuracy	0.8530
Precision	0.8487
Recall	0.8590
F1 Score	0.8539
Kappa	0.7059

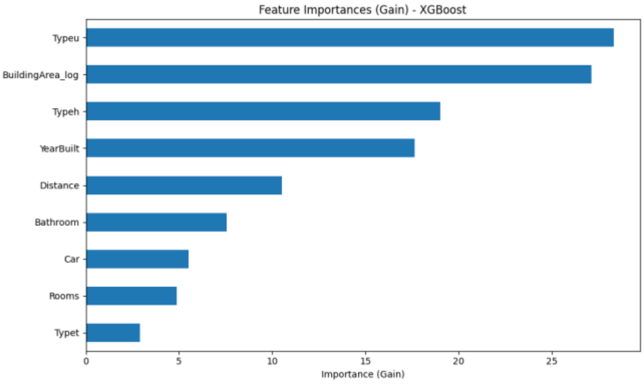


Fig. 15. Bar Plot

According to the feature importance plot of XGBoost, the most impactful variables are unit type, building area, and house type respectively.

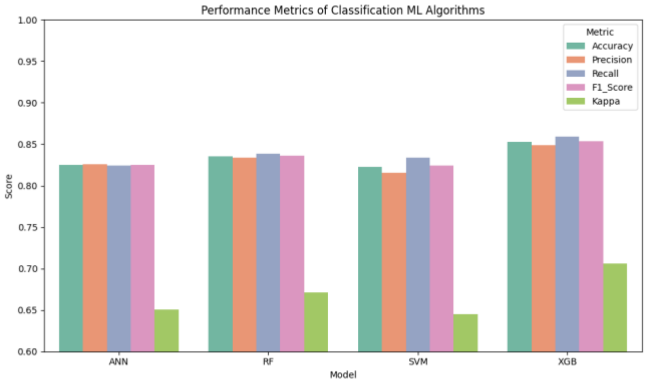


Fig. 16. Performance Metrics with Bar Plot on Test Set

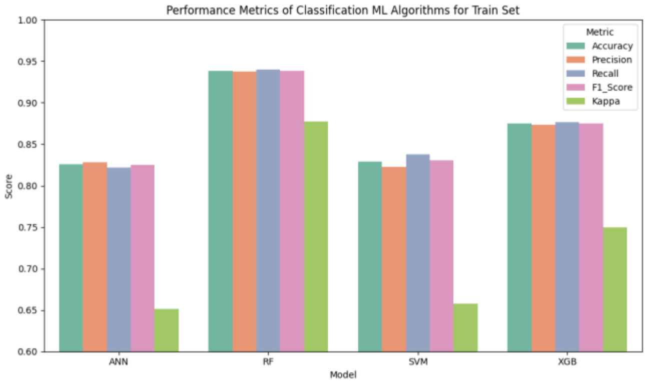


Fig. 16. Performance Metrics with Bar Plot on Train Set

From the plots, it can be concluded that all models perform better on the train set. It is expected because of the model fitting. In details, the models are optimized for learning patterns from the train sets, which can sometimes can be reason to memorization rather than generalization. For test set, XGBoost show the greatest performance compared to other models. Rest of the models are also showing great performance.

IV. Results

The study's dataset includes comprehensive details on Melbourne's housing market's structural, geographic, and demographic features. Significant patterns were found across variables using descriptive statistics. For instance, many of the properties were many decades old, as the average year of construction was 1965.
The majority of the dwellings were small, but several were noticeably larger, according to the right-skewed distribution of the BuildingSpace_log variable. The variety in the housing stock was also reflected in the variable distributions of other variables, such as the number of rooms, bathrooms, and parking spots.

Important trends were found using EDA. More expensive homes often had more rooms, greater built-up areas, more parking places, and were situated closer to the city center. Properties categorized as "Expensive" were situated in more accessible regions and were more frequently connected with Typeh housing. There were additional correlations between property price categories and categorical data, such as dwelling type and sales technique.

The continuous Price_log variable is classified using the median classification in order to facilitate binary operations. This creates a Price_Class variable that is classed as "cheap" and "expensive." The data's balance with regard to this goal variable was confirmed. The smallest and most likely MCAR missing data were imputed using the proper imputation techniques. Principal Component Analysis (PCA) was used to reduce the dimensionality, and the sum of the top payoffs accounted for almost 85% of the variance.

Binary Logistic Regression, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), and XGBoost (XGB) were among the many models that were constructed and assessed. XGBoost achieved 85.3% accuracy, 84.87% precision, 85.9% recall, and 83.60% F1 score, demonstrating the strongest performance on the test set. With accuracy scores ranging from 82 to 83% and strong precision-recall trade-offs, SVM and RF models showed demonstrated dependability.

V. Conclusion

Out of all the evaluated models, the XGBoost model had the best accuracy, recall, accuracy, and F1-score and Kappa score, making it the most successful algorithm for predicting home price classes. Dec. RF has closely followed it, exhibiting continuously strong and dependable Decent performance across all parameters. Even while ANN performs well, there may be over-compliance given its seeming drop from education to test metrics. The SVM model produced consistent and dependable results despite the slight performance lag. These results emphasize how crucial it is to choose suitable machine learning models according to assessment standards in order to produce reliable and accurate predictions for home price classification jobs. In addition, factors such as building area, year of construction, distance to the city center and number of rooms have been determined as the most effective factors in determining the price class.

References

[1] Zhang, Q., Lyu, Y., & Yu, F. (2023). Machine learning-based house price prediction using ensemble methods and SHAP for interpretability. Journal of Big Data Analytics in Real Estate, 5(2), 101–115. https://doi.org/10.1016/j.jbdae.2023.05.003

[2] Ritu, P. (2023). *Review on house price prediction models using machine learning techniques*. Materials Today: Proceedings, 74(3), 2202–2208. https://doi.org/10.1016/j.matpr.2023.03.062

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.