# MGT 6203 Group Project Final Report

Real estate price prediction in Brisbane, Queensland Australia, and influence of spatial features

**TEAM INFORMATION**

**Team #: 93**

**Team Members: Daniel Scott**, dscott89@gatech.edu, **Adhitya Arif Wibowo**, awibowo6@gatech.edu, **Yawen Yang**, yyang997@gatech.edu, **Rui Yan**, ryan87@gatech.edu, **Anthony Chan**, achan84@gatech.edu

## Choice of Topic, Business Justification, and Problem Statement

We aim to predict real estate prices in Brisbane, Australia to identify under/overvalued properties. Using listings from Realestate.com.au, spatial data from QGIS and government data, we'll build models incorporating location and non-location factors for accurate price insights.

Real estate investing provides additional diversification of an investment portfolio from the usual stock and bonds allocation. This is especially attractive during periods of low or increasing interest rates as mortgage agreements can be entered into at a fixed rate. It is however important to enter real estate investing when the property being invested in is not overvalued. It is therefore important for the investor to know if a property is undervalued or overvalued as identification of undervalued properties can be advantageous while overvalued properties can allow additional flexibility in price negotiation. Having a tool for estimating and forecasting property prices can be advantageous for the investor to make decisions for the best financial outcome. This includes having an estimate of the property value or as an aide in decision making whether it is a suitable time to buy or invest. Field professionals such as real estate agents can also use such a tool as an aid in offering prices to the buyer or for providing selling price advice to the seller.

Recent papers show machine learning improves real estate price prediction. Wang et al. (2014) optimized an SVM, lowering error 13%. Xiong et al. (2019) combined XGBoost, lightGBM and regression, reducing error 5-8%. Zhao et al. (2019) applied deep learning to XGBoost, decreasing error up to 28%. Sikder & Züfle (2020) combined matrix factorization and kriging, outperforming either alone. Sarathis et al. (2021) proposed a graph neural network using property locations and network topology, with lowest error among ML baselines.

## Understanding of the data and data wrangling

### 1. Data scraping

We have 5 sources of data, realestate.com.au, Brisbane city council datasets, Open Street Maps, Open Topo Data, and QGIS datasets. For the main data source from realestate.com.au, we built JavaScript code to scrape the site and successfully got 128,821 house listings from the Brisbane region, with columns and data type listed as in Table 1.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| full_address<br>string | type<br>string | zip_code<br>string | bedroom<br>integer | bathroom<br>integer | parking<br>integer | size<br>integer | sold_date<br>date |
| detail_link<br>string | description<br>string | AgentDetails<br>string | address_url<br>string | latitude<br>float | longitude<br>float | sold_price<br>integer | |

Table 1: property data from realestate.com.au

The process of scraping takes around 2 weeks to complete with the JavaScript code to download the properties information chunk by chunk then these chunks are compiled into 1 csv file using R script which then loaded to SQLite database for further analysis of data completeness. As of now, we have 878 records with missing latitude and longitude data. These 128,821 records required further clean-up and augmentation / transformation process depending on modelling approach used.

From Brisbane city council datasets, we downloaded Brisbane parks, bus stations, libraries, schools, and more, all with latitude and longitude. From Open Street Map data, we found a smaller and, due to the citizen data cataloguing involved with the service, possibly less reliable set of spatial data for points of interest like supermarkets, department stores and coffee shops. All these spatial data points were used to calculate additional features with QGIS.

The Open Topo data was used to extract the approximate elevation of each property (z spatial dimension) to add another possibly unique dimension to the data for exploration.

We then checked for missing or incorrect values, duplicates, outliers, and other issues that can affect analysis, formatting values consistently, handling encoding issues, replacing, or removing missing values, and resolving duplicates.

During our review, several real outliers were discovered. One outlier has encouraged us to filter out data points for housing properties with more than 6 bedrooms. This is due to multiple observed occurrences of a high number of bedrooms in relatively small houses. We think that this could be data entry error or real properties with non-standard uses, e.g., property for dormitories, apartment complexes, or multiple houses under one listing. These outliers make it difficult for us to determine the validity of data and the high number of bedrooms also makes it likely to be a dormitory or an apartment complex which is outside the scope of this study for predicting prices for single residential homes. Figure 1 is an example of a floor plan of a property with a high number of bedrooms.

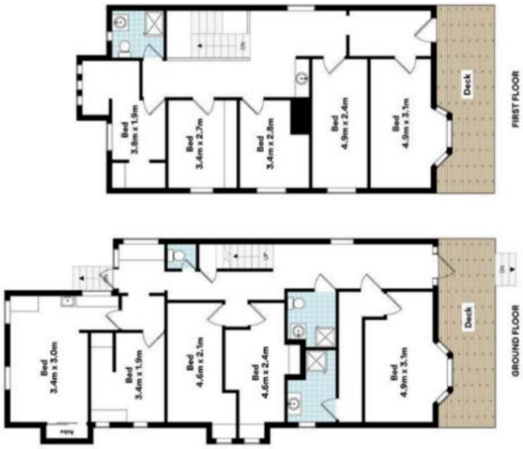

Figure 1. Property with high number of bedrooms

Data points with parking spaces that are greater than 10 will be filtered out, as these are likely to be apartment complexes or dormitories. The size feature of the property will be modified to reflect the correct units. Sizes are commonly expressed in square meters. We have assumed that a house must be at least 10m$^2$ (informed by reviewing a subset of properties). Properties with sizes less than 10m$^2$ will be converted to hectares which seems to be a valid transformation based on a review of scraped listings.

## 2. Feature engineering

From scraping data from Realestate.com.au, we get a column of descriptions, which are detailed natural language summary of house features written by agents. We thought this could be our independent variable source and built a python notebook (project_data_feature_extraction.ipynb) to extract features from this column, by cleaning with stop words, tagging, synonyms standardization and counting. An example of house features extracted from these descriptions are shown in Figure1, where word size is indicative of word frequency.

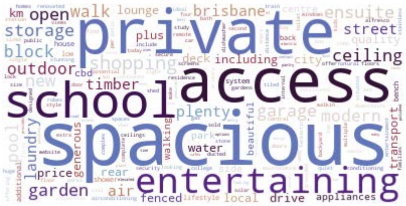

Figure 2: House Feature Word Cloud Extracted from Description on Realestate.com.au

Based on the top word list and our understanding of the house market, we selected 18 features and converted the description column into 18 factor variables for inclusion in our dataset.

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **Description Features** | spacious | private | access | school | open | storage | new | entertaining | modern |
|  | Pool | garden | shopping | outdoor | timber | fenced | park | laundry | auction |

Table 2: Binary features of interest extracted from the property listing 'description'

Spatial features that have been calculated are the distance from the property record to the nearest point of interest (e.g., bus, library, etc.), and the compass bearing between these two points. We have also calculated the number of points of interest within 100m, 500m, and 1000m of the property as 3 additional spatial features.

### 3. Data Finalization and Explanatory Analysis

The finalized version of our dataset has 125471 rows and 147 columns including 14 columns of non-spatial features, such as number of parking lots, bedrooms, 114 columns showing distance between points of interest and the properties, and 17 factor variables we extracted from descriptions.

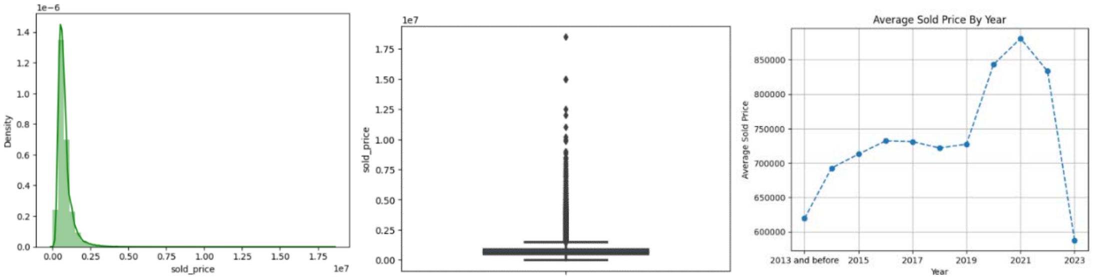

Figure 3. Distribution of Sold_Price and its relationship with sold_year

The distribution of 'sold_price' approximates a normal distribution with a long tail for farm/luxury houses. The relationship between 'year' and 'price' suggests that inflation and the pandemic could be accounted for by normalizing 'sold_price'. We obtained historical inflation rates from the Australian government website and adjusted 'sold_price' accordingly, effectively removing the inflation trend. We also created a binary variable indicating whether a house was sold before or after 2020 to account for the pandemic. A correlation check of all variables revealed promising linear relationships for several factors, while other variables require further exploration for potential non-linear relationships.

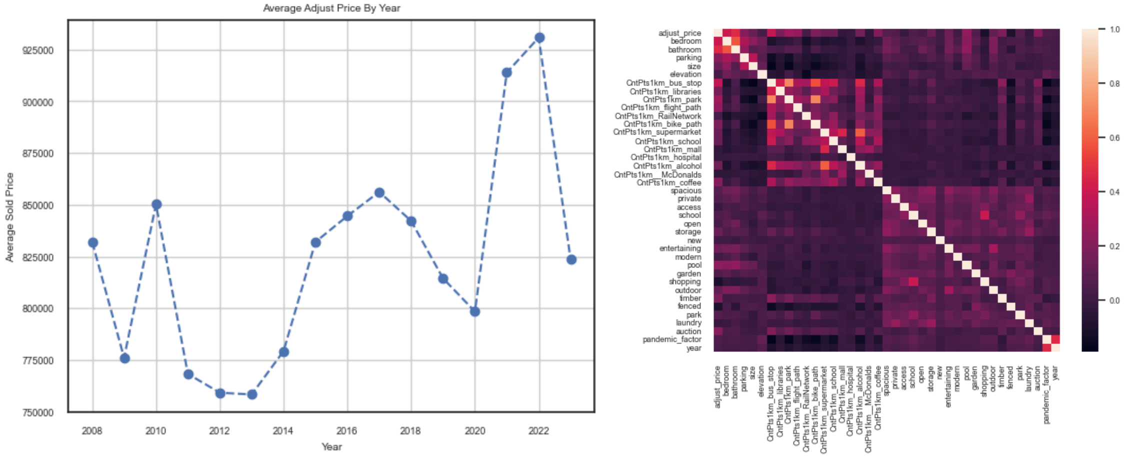

Figure 4. Adjust Price with Inflation and Correlation Plot of All Factors

**Modeling**

Our 128,821 datapoints are divided into two parts, data before 2023 (about 90%) for training and cross validation, and data in 2023 (about 10%) for model comparison. On this data, we use linear regression, random forest, PCA regression, Support Vector Regression, XGBoost, and KNN. RMSE performance on hold out data is used for model comparison.

**Regression**

Before applying linear regression for housing price prediction, we assessed the data for potential issues. Fitted vs residuals plots from regression indicated heteroskedasticity and outliers in our data, as shown in Figure 3 (left). To address this, we applied a log-linear transformation and removed outliers. The effects of these transformations, evident in Figure 3 (center and right), improved the model's accuracy, as confirmed in Table 3. Model performance metrics like R-squared and RMSE were obtained by testing the model against the 2023 sales data set aside for this purpose.

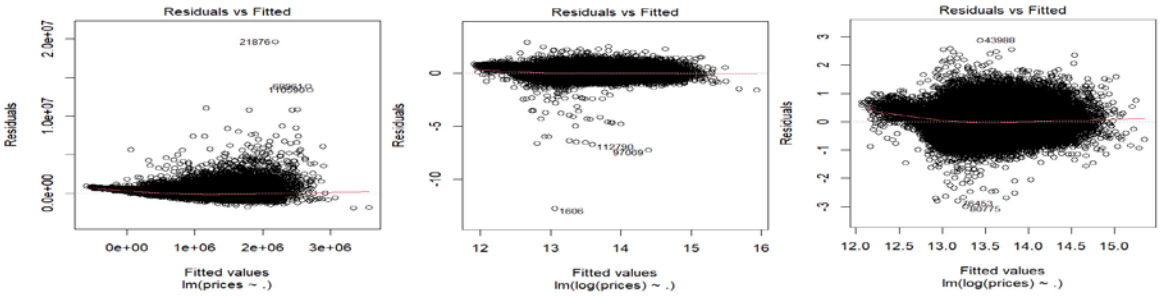

Figure 5. (left)linear-linear residuals vs fitted, (center) log-linear transformed residuals vs fitted, (right)log-linear transformed with outliers removed residuals vs fitted.

| **Model** | **Linear-linear RMSE** | **Linear-linear R$^2$** | **Log-linear RMSE** | **Log-linear R$^2$** | **Log-linear w/o outlier RMSE** | **Log-linear w/o outlier R$^2$** |
|---|---|---|---|---|---|---|
| **Linear regression** | $392,661.29 | 0.3451 | $375,703.97 | 0.4340 | $375,655.78 | 0.4341 |

Table 3. RMSE and R-squared for different data transformations.

We applied regular linear regression, lasso, and ridge regression to predict housing prices, but all three methods yielded comparable results. Table 4 presents the Root Mean Squared Error (RMSE) and R$^2$ values for linear regression models: one without spatial features, another with spatial features, and a third incorporating spatial features and select interaction terms, including distances to schools x libraries, alcohol outlets x McDonald's. Interaction terms brought only marginal improvements to R$^2$. Overall, spatial features accounted for an additional 20% variance in the sold price.

| **Model** | **Log-linear w/o outlier RMSE** | **Log-linear w/o outlier R$^2$** |
|---|---|---|
| **Linear regression (no spatial features)** | $375,655.78 | 0.4341 |
| **Linear regression (w/ spatial features)** | $322,975.28 | 0.6306 |
| **Linear regression (w/ spatial + interaction terms)** | $322,420.471 | 0.6310 |

Table 4. Regression RMSE and R-squared- spatial and interaction features influence

We divided our housing price dataset into two periods to investigate changes in the relationship between spatial features and housing prices during the COVID-19 pandemic (2020-2022). Linear regression analysis revealed that, during the pandemic, housing prices decreased with proximity to parks, flight paths, rails, and hospitals, while prices increased with proximity to amenities like bus stops, supermarkets, malls, and shops like McDonald's or coffee shops. We also observed a reversal in the price relationship with proximity to alcohol stores. With 92% confidence, houses closer to such alcohol shops sold for higher prices during the pandemic, compared with their lower prices pre-2020.

### Random Forest and Gradient Boosted Regression Trees

To find the optimal hyperparameters for the Random Forrest model, we employed grid search, an exhaustive method that trains the model on a predefined grid of parameter values, then evaluates performance via cross-validation.

Our predefined grid covered three key parameters: 'n_estimators' (number of trees in the forest) with values [50, 100, 150], 'max_depth' (maximum depth of each tree) with values [None, 8, 16], and 'min_samples_split' (minimum samples to split a node) with values [25, 50, 100]. This resulted in 27 total combinations.

We applied GridSearchCV, setting the base estimator and parameter grid, and designating 5-fold cross-validation. The model was then trained using each of the 27 hyperparameter combinations on the training data (Xtrain, ytrain). For each combination, the model's performance was evaluated through 5-fold cross-validation on the training set, utilizing the neg_mean_squared_error scoring metric.  The selection of the best model was based on the highest score (in this case, lowest negative MSE) achieved across the cross-validation folds. The optimal hyperparameters were found to be 'max_depth': 16, 'min_samples_split': 50, 'n_estimators': 150. The model was then trained with these parameters, used to make predictions on the testing dataset, and the RMSE calculated.

While we did explore Gradient Boosted Regression Trees (learning rate of 0.1), its RMSE did not perform as well as the Random Forest despite producing similar feature importance assessments.

### KNN

Initially, all the columns in the dataset were normalized, except categorical variables, year, and the target variable. After normalization, the stepwise method was employed for variable selection.  Variables with a p-value greater than 0.05, like 'bike path' were removed. In total, 34 variables were chosen to build the KNN model, consisting of 10 non-spatial variables and 24 spatial variables. We then proceeded with outlier removal as the next preprocessing step. Houses sold within the price range of 50,000 to 10 million dollars were retained. A significant preparatory step for modelling was choosing the optimal K separately for both groups (spatial vs. nonspatial) using cross-validation. We used 5-fold cross validation and tried k values from 2 to 50 to achieve the optimal model.  From this process, the optimal k for non-spatial and spatial datasets were respectively 34 and 13. From here, we individually trained KNN models to determine if the model with spatial features provides better housing price predictions than the non-spatial feature model.

The non-spatial feature model in this instance shows lower RMSE for housing prices than the spatial model.

| | RMSE | $R^2$ | # Properties sold OVERvalued | # Properties sold UNDERvalued |
|---|---|---|---|---|
| **Non-spatial features** | $368,738.32 | 0.42363 | 8735 | 3825 |
| **Spatial features** | $393,231.94 | 0.34451 | 6809 | 5753 |

Table 5. KNN Performance comparison between spatial and non-spatial data

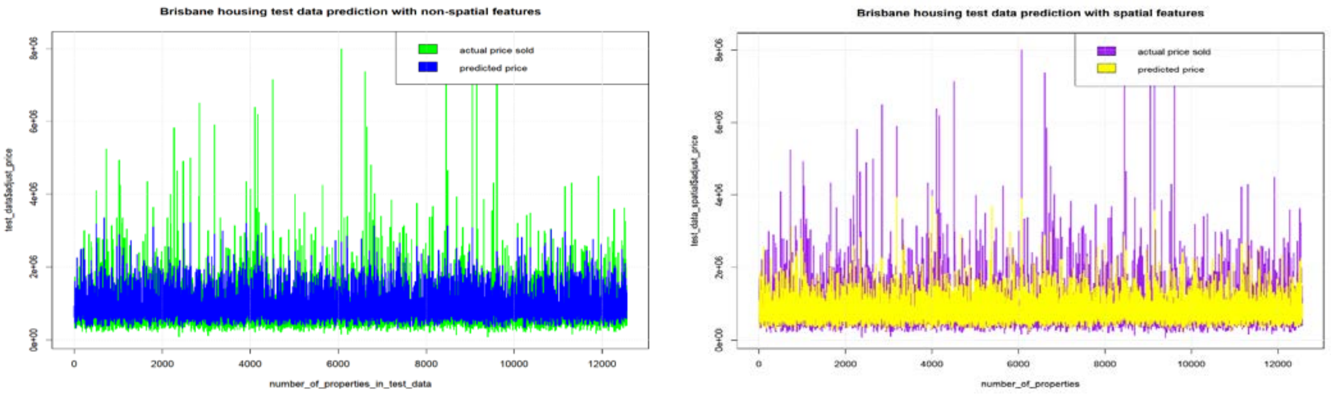

Figure 6. Predicted vs. actual sold price for the test data; non-spatial features (LEFT) and spatial features (RIGHT)

### PCA and Regression

Principal component analysis (PCA) was performed on the standardized training data for regression modelling, and the same standardization was applied to the test data set. Considering the time-series nature of real estate transactions, an expanding time window cross-validation with five folds was implemented, using all features.

PCA's hyperparameters selection was challenging due to PCA's feature combination process. Each principal component describes orthogonal variance in the data, with the composition possibly changing significantly as new data records or underlying data generating distributions change. Meaning, the components most predictive in the initial fold of cross-validation may not maintain that status in subsequent folds.

A Bayesian optimization process was used to find the optimal hyperparameters for each model. Consistently predictive principal components across all five folds were selected as the optimal hyperparameters for the PCA + regression model. Interestingly, model performance improved until the last fold. This seems to align with the period of successive mortgage interest rate increases in Australia starting August 2022.

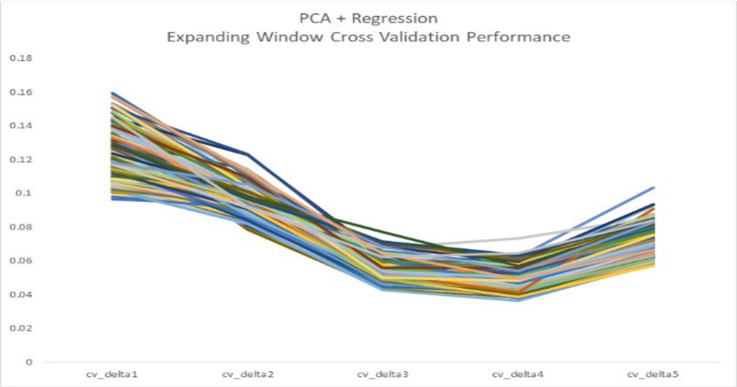

Figure 7. Cross-Validation fold performance of PCA + Linear Regression (absolute %difference from actual sold price)

The optimal hyperparameter model was trained on all data (minus hold-out) and tested on the hold-out dataset. It yielded a root mean squared error of $312,960.

**Support Vector Regression**

The structure for SVR's hyperparameter selection mirrored that of the PCA cross-validation analysis. However, we used only a randomly sampled subset of the training data for each fold due to the excessive time required to train an SVR on the full dataset through five folds at each step of the Bayesian optimization process. The 4 hyperparameters that were experimented with were the kernel type, C, gamma, and epsilon.

SVR is computationally expensive to run, so the kernel was fixed to be linear rather than experiment with other non-linear models that might be more performant but otherwise more expensive to explore. As for the remaining hyperparameters, “C” controls the level of overfitting / sensitivity to errors, “Epsilon” controls the region of the solution space through which approximate predictions will not be penalized. e.g., C is the size of the penalty for points that fall outside of the “window” that epsilon specifies. “Gamma” determines how influential any single point is in the non-linear kernels (e.g., smoothness of model), and so was not relevant once a linear kernel was chosen.

Support Vector Regression with a linear kernel achieved a mean squared error of $289,400 on the hold out data set. Like the PCA, the performance of the support vector regression also declined in the last fold of cross-validation.

**XGBoost and Deep Learning Alternative**

Like Random Forest and Gradient Boosted Trees, for XGBoost we use random search with 10-fold cross validation and 100 candidates resulting in 1000 combinations to find the best hyperparameters from below parameter distribution:

| **Parameter** | **Value** | **Remarks** |
|---|---|---|
| learning_rate | np.arange( 0.01, 0.1, 0.01 ) | step size shrinkage used in update to prevents overfitting |
| min_child_weight | [ 1, 5, 10 ] | minimum sum of instance weight (hessian) needed in a child |
| colsample_bytree | np.arange( 0.1, 0.5, 0.1 ) | subsample ratio of columns when constructing each tree |
| colsample_bylevel | np.arange( 0.1, 0.5, 0.1) | subsample ratio of columns for each level |
| max_depth | range( 4, 10 ) | maximum depth of a tree |
| n_estimator | range( 100, 1000 ) | how many trees are built within the ensemble model |

Table 6. Distribution of parameters in random search cross validation

We also use different dataset (all numerical) with summary below:

| Dataset | RMSE | Remarks |
|---|---|---|
| Full Spatial Data | $209,121 | 103 predictors |
| Subset Spatial Data | $206,675 | 45 predictors; remove bearing and keep only count of place of interest within radius of 1 kilometer |
| No Spatial Data | $390,478 | 5 predictors; remove all spatial related predictors |
| Feature Engineered Data | $244,530 | 37 predictors; include inflation adjusted price, pandemic indicator, and one hot encoding words of interest indicator |

Table 7. XGBoost performance summary with different datasets

We conducted further experiments using Low-Resource Text Classification [7] as a deep learning alternative in tandem with the XGBoost model applied to a subset of spatial data. For each test record, we used the predicted price from the XGBoost model to filter records from the training data within a range of [predicted price - lower bin width, predicted price + upper bin width]. We then calculated similarity distances according to the approach outlined in [7]. From this filtered result, we performed a KNN grid search with k = [29, 59, 89] and (lower_binwidth, upper_binwidth) = [(100,000, 900,000), (850,000, 1,150,000)]. This process yielded improved results with k=29, and (lower_binwidth, upper_binwidth) = (100,000, 900,000), further reducing the RMSE to $163,402. However, due to the high computational demands, this approach was only tested on 100 random sample testing datasets without cross-validation in the test dataset. Consequently, we've excluded these results from the model comparison. Nevertheless, this approach shows promise and merits further investigation.

**Conclusions:**

**Research Question 1: What type of model or algorithm performs the best on our chosen metrics**

In our project, we used more than 10 different models and found that, on the matrix of RMSE, XGBoost performs the best among others, then Random Forest and Gradient Boosted Regression Trees, and then linear regression and PCA regression. KNN may not be a good type of model for our project.

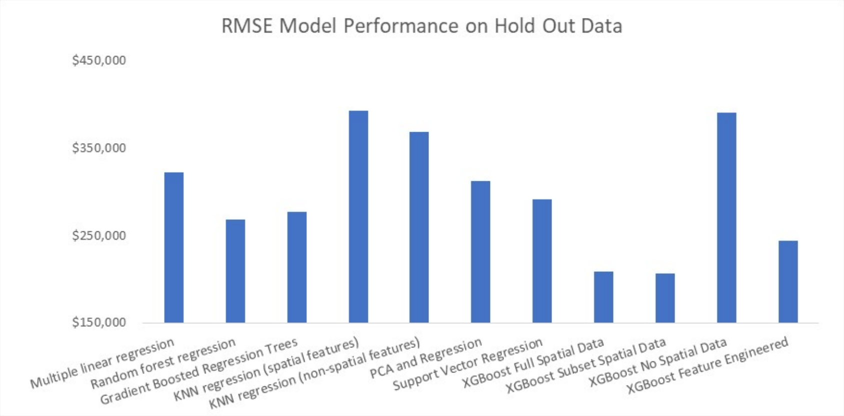

Figure 8. Model Comparison of root mean-squared errors on hold out data.

**Research Question 2: How much additional predictive power do spatial variables provide relative to a traditional hedonic model of real estate prices in Brisbane?**

Yes, spatial variables provide significant additional predictive power relative to a traditional hedonic model of real estate prices in Brisbane. Our research into key aspects of Brisbane's residential property values yielded significant insights. We found that spatial features play a crucial role in determining residential property values, being statistically significant in most predictive models we evaluated. The only exception was the K Nearest Neighbors model. However, these spatial features accounted for over 20% of the variance in residential property values, compared to non-spatial linear regression models.

**Research Question 3: How significant are spatial features for determining residential property values in Brisbane?**

To investigate the feature importance to our model, we rank and plot the importance of features and find: **<u>bathroom, bus_stop, size, bedroom, elevation, flight, pool, bike, railnetwork, park, parking, pandemic, type, supermarket, school, alcohol, coffee, shopping ,auction and timber</u>** are the top 20 most important features for house transaction price. So, **spatial features such as <u>bus_stop, flight, bike, railnetwork, park, supermarket, school, alcohol, coffee, shopping are all significant.</u>**

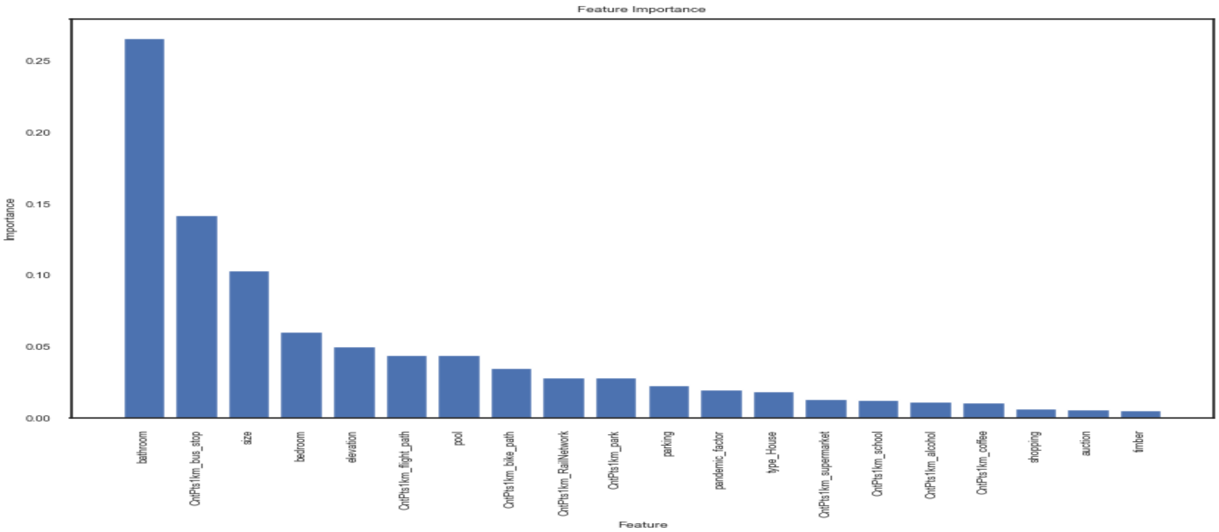

Figure 9. Feature Importance from Random Forest

**Research Question 4: Do any spatial features have nonlinear effects where they positively affect price up to a certain distance and then become a negative influence? E.g., proximity to schools, hospitals, public transportation, bodies of water, etc.**

Yes, some spatial features have nonlinear effects. We use partial dependence plots to show the marginal effect of features on the predicted outcome of a model. They illustrate how the predicted value changes as a function of the selected features, marginalizing over the values of all other features. For instance, the number of nearby bus stops matters. The line is so steep between 0 and 1 shows it really matters if there is one bus stop nearby. The level of elevation is split on 35, almost V-shape. Whether the house is timber made, has a pool, was sold after pandemic, near school, was sold by auction, etc., all these are quite positive factors on sold_price.

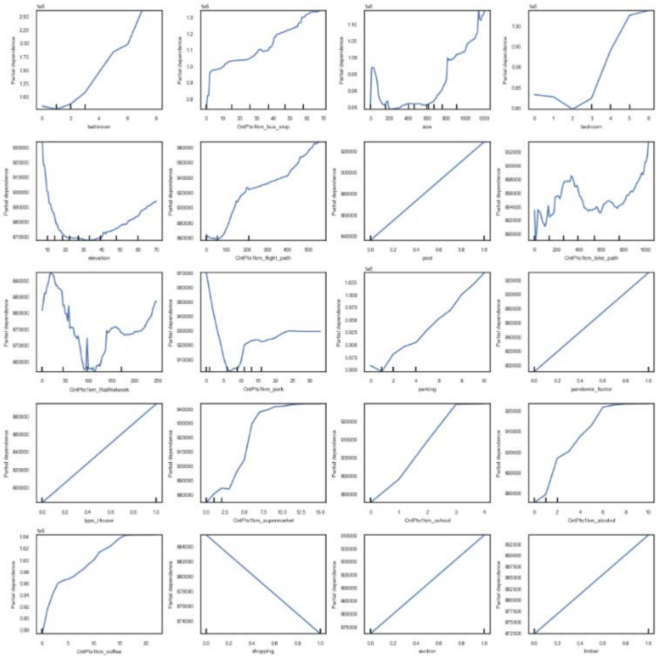

Figure 10. Feature Partial Influence from Random Forest

**Research Question 5: Is pandemic factor significant for price prediction? And how do pandemics influence spatial features?**

Yes, pandemic factor is significant for price prediction. It is significant in almost all our models, from linear regression to random forest, to KNN, to XGBoost . And yes, pandemic factor does influence buyer's spatial feature preferences for houses. This is supported by an EDA performed on pandemic factors.  As shown below, all property types (except for townhouses) see an increase in transaction price. For basic features, the number of bedrooms, bathrooms, parking, size and elevation as well as adjust_price increases.  It seems people are inclined towards buying bigger houses after the pandemic.

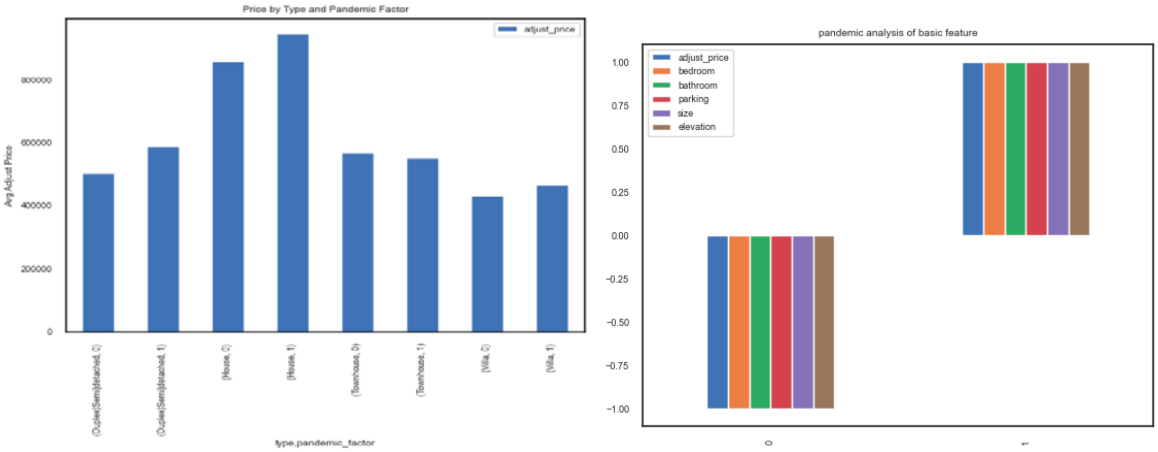

Figure 11. Pandemic analysis of basic features of house

As for spatial features, people seem to want to move far away or the number of supermarkets, schools, McDonalds etc. all decrease due to economic slowdown and lockdown policy. As for description features, people seem to want to buy more private houses, with more pools, storages and close to outdoors.

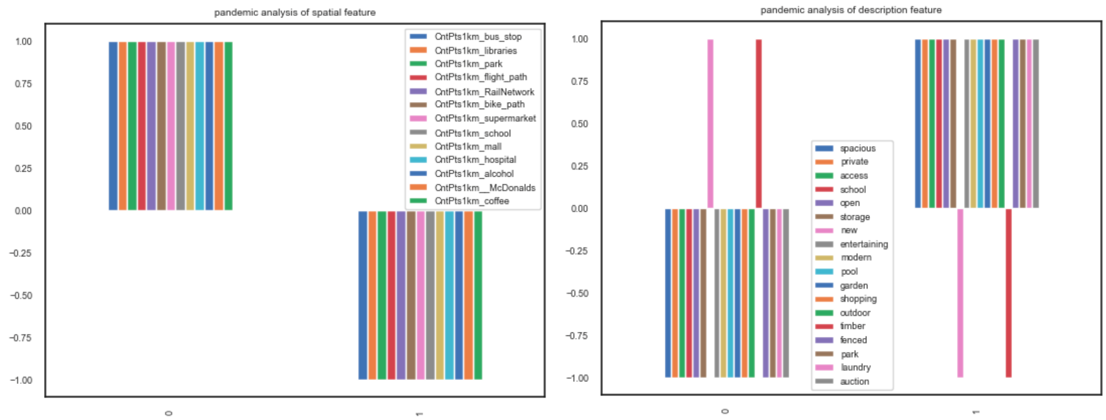

Figure 12. Pandemic analysis of spatial and description features of house

**Research Question 6: Are there any significant spatial interaction terms that stand out from the analysis?**

While some spatial interaction terms boosted the predictive power of our models, their overall effect on increasing $R^2$ was minor.  More investigation is needed.

However, the research did not have the opportunity to explore the potential nonlinear effects of certain spatial features, such as proximity to schools, hospitals, and bodies of water. Further analysis, potentially through a CART analysis, is required to identify any nonlinearities.

**Business justification: How to use our project to make business decisions, which is to check whether a property is overvalued or undervalued, and accordingly to buy or sell?**

While our analysis has not yet produced a model robust enough to confidently provide sales price recommendations, it offers valuable insights into how specific features influence property prices. These insights can guide homeowners when deciding on renovations or choosing sales methods. For instance, we now comprehend how adding a bedroom, installing a pool, or switching to timber floors can impact the sales price. Additionally, our analysis can inform whether an auction could yield better returns than a private sale.

To aid in property transaction decisions, our project offers ten distinct models to predict transaction prices. The ensemble approach allows these predictions to be used collectively, perhaps in a system where each model casts a 'buy' or 'sell' vote. If seven out of ten models predict a house is undervalued, it may be considered a potential purchase, considering all other factors. The ensemble model approach is less likely to overfit compared to individual models, as they counteract each other's inherent biases.

In conclusion, our findings provide valuable insights for homeowners considering renovations before a sale, while laying a strong groundwork for further research.

**Discussion and Next Steps:**

To transition from preliminary work to a viable business, we recommend further data acquisition, feature engineering, and model optimization to decrease prediction errors.

The team started to go through the process of extracting real estate image features (objects, picture brightness / color palette), but was not able to complete the work in the allotted time.

Late in the project, we introduced a new feature inspired by[4], namely the spatiotemporal inverse distance squared weighted average property price. This measure uses the sale prices of the nearest 30 properties sold 30 days prior to a given property's sale, weighted by their respective squared distance (in space and time). This feature correlates strongly ($R^2$ of 0.45) with actual home sales prices, suggesting potential to enhance our predictive models. However, its inclusion might diminish the effects of other spatial data (and estimate of importance), as it captures all space and time autocorrelation sources.

From a modelling standpoint, it would be worth exploring custom objective functions beyond MSE or RMSE during hyperparameter selection, based on the customer type. A buyer of property would be looking for information on the lowest bid they could reasonably make to be successful, while a seller or real estate agent would be looking for the expected value of a property and upper bound so they could apply appropriate markups / be prepared with minimal acceptable offer strategies.

Also, during expanding window cross validation for SVR and PCA, a notable decrease in model accuracy was detected in the last fold. This might indicate that models are using too much historical information in their predictions. Experimenting with limiting historic data or up sampling recent sales might improve performance.

Also, while we did explore some ensemble modelling approaches like Random Forrest, we have not explored ensembles of the various linear and nonlinear models we have created. This might improve predictive accuracy.

Lastly, it is worth noting that while current RMSE performance is not ideal for making investment decisions, this work could still be used to provide consultancy services to inform buyers of properties that have the right spatial features to suit their individual needs and to also provide high level guidance on homes that might be good “fixer uppers” (e.g. build a pool at this house and you could get $X return on investment) and get ad revenue by linking buyers to contractors that could do the work.

**Appendix:**

**Works Cited:**

- www.realestate.com.au
- www.openstreetmap.org
- www.opentopodata.org
- www.spatial-data.brisbane.qld.gov.au
- https://qgis.org/en/site/
- https://scikit-optimize.github.io/stable/modules/generated/skopt.gp_minimize.html
- https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

**Reference:**

1. Wang X, Wen J, Zhang Y, Wang Y (2014) Real estate price forecasting based on SVM optimized by PSO. Optik 125(3):1439–1443
2. Xiong S, Sun Q, Zhou A (2019) Improve the house price prediction accuracy with a stacked generalization ensemble model. In: International conference on internet of vehicles, pp 382–389. Springer
3. Zhao Y, Chetty G, Tran D (2019) Deep learning with XGBoost for real estate appraisal. In: 2019 IEEE symposium series on computational intelligence (SSCI), pp 1396–1401. IEEE
4. Sikder A, Züfle A (2020) Augmenting geostatistics with matrix factorization: a case study for house price estimation. ISPRS Int J Geo Inf 9(5):288
5. Sarathis, S., Zhang, Q., Wang, H., & Feng, Y. (2021). Boosting house price predictions using geo-spatial network embedding. arXiv preprint arXiv:2009.00254.
6. Rothstein, R. (2023, June 8). Housing Market Predictions For 2023: When Will Home Prices Be Affordable Again? Forbes Advisor. https://www.forbes.com/advisor/mortgages/real-estate/housing-market-predictions/
7. Jiang, Zhiying; Yang, Matthew; Tsirlin, Mikhail; Tang, Raphael; Dai, Yiqin; Lin, Jimmy; "Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors. https://aclanthology.org/2023.findings-acl.426/