2024 COMP20008 Assignment 2

Group W09G4

Heather Maltby
1258489
hmaltby@student.unimelb.edu.au

Apurva Lekkala
1616156
alekkala@student.unimelb.edu.au

Melani Samarakoon
1615978
msamarakoon@student.unimelb.edu.au

Executive Summary

The purpose of this report is to summarise our findings in relation to the research question “To what extent do the features in the Victorian Communities Dataset predict housing prices?” We found that the level of education in a suburb was the biggest predictor of house price, and although they are not the strongest predictors of house price, better access to health services was associated with higher house price. This shows that equal access to quality education is extremely important for everyone having an equal chance to live in a nice suburb and have access to essential health services. Hence, we suggest that the government make it so that people with jobs that do not require a university education are paid more and therefore have similar access to quality housing and services those who went to university enjoy. The methods we used to predict house prices were a decision tree and linear regression. To predict house price, we used linear regression and a decision tree.

Introduction

The research question was “To what extent do the features in the Victorian Communities Dataset predict housing prices?” We chose this question due to the current context of the housing crisis in 2024. According to a recent statistic in the Guardian, only 13 percent of households on the median income can afford a home in the current market, down from 40 percent 3 years ago (Readfearn, 2023). This is a significant worsening of the accessibility of affordable housing for the majority of the population, and something that needs to be addressed. Of course, there will be many other factors that influence the cost of housing that are not mentioned in the Victorian Communities Dataset, however the factors in the Victorian Communities Dataset give an indication of which groups in the population are struggling the most, and therefore who the government should prioritise helping.

The data that we analysed was from 2013 so there will be some differences between 2013 and today, however we hope that it still provides useful insights that could guide future research with more recent data and inform what factors we would want to collect information about if we were to do surveys and data collection now. The data sources used were the Victorian Communities Dataset, which included information about demographic characteristics in different suburbs, such as education level, language spoken, occupation, and income level, amongst other things. There were 1080 rows 226 columns in the Victorian Communities dataset. The Houses By Suburb Dataset provided the average house price in each suburb from 2013 to 2023, however we only used 2013 as this matched with the Victorian Communities Dataset. There were 787 rows and 13 columns in the Houses By Suburbs dataset.

Methodology

The Python libraries Pandas, re, json and numpy were used to pre-process the data in order to fill NaN values, merge, add and remove columns, ensure consistent formatting for Victorian Communities dataset and Houses by Suburb dataset, remove rows and merge datasets. Numpy and matplotlib.pyplot were also used in exploratory data analysis in order to generate scatterplots of House Price vs every column in the pre-processed communities dataset, which were used to decide additional columns that could be removed if they had a number of NaN values and showed no apparent correlation with House Price based on the scatterplots. Some of the data was also linearised by using either logarithmic, squared or reciprocal transformations depending on what the data originally looked like on the scatterplots. Data was also discretised using sklearn.preprocessing in order to assist with the Decision Trees model later. Categorical variables were encoded using one-hot encoding for Linear Regression models. The library sklearn was further used to create a Decision Tree model and two versions of Linear Regression models, one with only numeric variables and another with both numeric and categorical variables.

Data Exploration and Preprocessing

Dealing with Null Values

A number of techniques were used to preprocess the data in order to make it ready for exploratory data analysis and modelling. NaN values were dealt in different ways depending on the columns they were in. The removal of NaN values was to prepare the data for modelling, since models use mathematical operations when deployed, and since NaN values are undefined, they lead to disruptions in the models, leading to errors. Similarly, the NaN values would make the use of mutual information (MI) impossible, since MI works based on probabilities, which cannot be calculated with the presence of NaN values.

Some columns ('Dwellings with no internet, %' (3 NaN values) and 'Equivalent household income <$600/week, %' (1 NaN value) for example) were imputed with values that were calculated by dividing the equivalent household income <$600/week column by the number of households in that suburb. This is because it was more useful to have the percentage of households rather than the actual number, since not all suburbs have the same population. Other columns were imputed with their median values since there was no way to calculate the missing values based on the available data.

The column 'holds degree or higher, %' (1 NaN value) was imputed with 0's since all the NaN values for that column were for communities that had very small populations, and imputing them with the median for the column might reduce the accuracy of the data due to the huge disparity in population densities.

Data Type Conversion

Another preprocessing technique used was data type conversion. The columns 'Primary school students', 'Secondary school students', 'TAFE students' and 'University students' were columns that were supposed to be integers, however were classified as objects due to the occurrence of the string '<5' 3 times, 2 times, 7 times and 5 times respectively. This was dealt with by replacing the '<5' with 5 and converting those columns to integer type. The column 'Number of families' however was classified as an object despite only having integers, so the data type of this column was simply changed to int. This was later useful when plotting graphs and modelling with the data as the numeric columns were able to be treated as numeric columns.

Altering and Removing Columns

Furthermore, the column 'Location' was split into two columns that represented 'Distance' and 'Direction' using regex, and Distance's data type was changed to int in order to numerically work with it, and the 'Location' column was deleted. This was to make it easier to properly analyse the distance and direction, which was impossible to do in the previous Location column due to numerical and categorical variables being combined.

To decide which variables to focus on, we produced a scatterplot for every numeric column versus the House Price column. This allowed us to visualise the relationship between all numeric columns and House Price, helping us decide which columns had some correlation with House Price based on the scatterplots. After examining the scatterplots, columns which had no apparent correlation with House Price and consisted of multiple NaN values (more than 10) were removed from the data frame. These scatterplots also helped us decide which columns could and should be linearised in our modelling section. Table 1 showcases the scatterplots produced for the columns with the highest normalized absolute mutual information scores.

Column filtering was another technique used to preprocess the data. A number of columns were removed from the original communities dataset by manual inspection; the columns that seemed likely to influence house prices were included, while the rest were excluded. This reduced the total number of columns in the communities dataset from 226 to 68. More column filtering was then applied to merge other columns, such as the populations according to ages, and number of different health care centres, which further decreased the number of columns to 56. This made the data more readable and easier to work with during exploratory data analysis. Rows were also filtered out to only include suburbs from the communities dataset as the house prices dataset only included suburbs.

Combining Datasets

We made a function to match the suburbs from the Houses By Suburbs dataset with the Communities dataset. This function was successful at matching the majority of the suburbs (426 out of 452). The way that this worked was by using pattern matching to check if the lower case version of each suburb in the

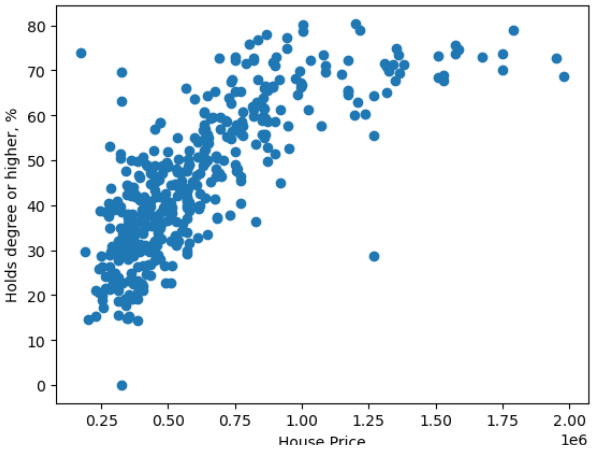
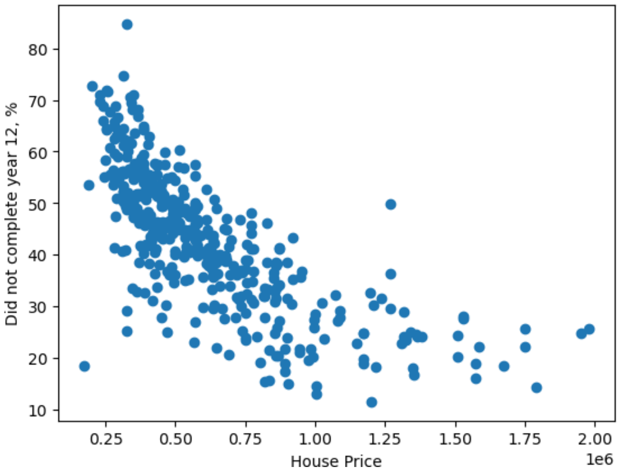
Houses By Suburbs dataset matched with a suburb in the communities dataset. The function went through each row of the communities dataset to check if there was a match, and stopped when it found one or put ‘Nan’ if there was no match. There were some difficulties with this due to missing brackets in the Communities dataset which made it more difficult to match suburbs, and also some different naming conventions. The function could be improved by having a lower standard for a match (for example a less similar name, such as one that has the same name but has part of the name in brackets). This would result in more matches and therefore more data.

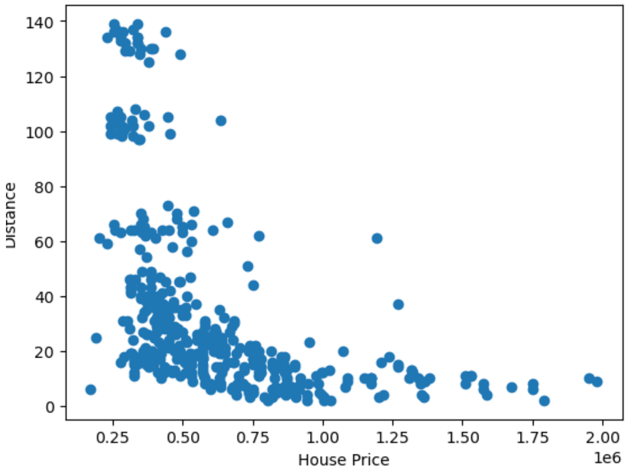
House prices from 2013 were then merged with the preprocessed communities data frame based on community names, using the matching suburbs function, and rows that had NaN values due to a community name not being in both the communities dataset and the house prices dataset were removed, since the suburb name and house price could not be predicted by using any other factors in either of the datasets.

Correlation Coefficients and Mutual Information

To determine which factors had the biggest influence on house price, and therefore which ones should be used in the model, we analysed the correlation between each factor and house price, and also calculated the mutual information between each factor and house price. The strongest correlations were that had a high percentage of people who did not complete year 12 had a lower house price (-0.73 Pearson correlation), and suburbs that had a high percentage of people who completed a university degree had a higher house price. (0.76 Pearson correlation). These values can be seen in Table 1 below, along with the scatterplots of these columns against House Price.

Table 1: Top 3 Correlation Coefficients (Absolute).

Variable	Pearson Correlation Coefficient	Scatterplot
Holds degree or higher, %	0.765	
Did not complete year 12, %	-0.738	

Distance	-0.498	

The normalised mutual information function we wrote added to the information from analysing the correlation because it would compare categorical variables as well. The binning strategy we used for this was equal-width bins because this means that all house prices across the range of house prices are covered well in the bins. The highest mutual information for numerical variables was the percentage who finished year 12, percentage who have a university degree, and distance from the city centre, which is the same top ones as were found in correlation. This can be seen below in Table 2.

Table 2: Top 5 Mutual Information (MI) Values

Variable	**Normalised mutual information score**
LGA	0.3624
Holds degree or higher, %	0.2784
Did not complete year 12, %	0.2701
Distance	0.2237
Top language spoken	0.1935

The highest mutual information for categorical variables was for LGA (0.3624), and top language spoken (0.1935). LGA being highly correlated indicates that suburbs near each other have similar house prices, which can allow us to see distinct trends as to where the house prices are higher and where they are lower. Top language spoken also having a moderate MI indicates that house prices differ based on the main ethnic community in an area. One issue this raises is that it means that there are some ethnic communities that are getting worse access to services than others because some areas are much better serviced and much more well off than others. Similar to what was suggested earlier in the report, this suggests that to ensure all migrants and people in Australia have equality and equal opportunities, the government needs to equally invest in services across all communities in Victoria, and also work towards making everyone have equal access to education.

We also found that almost all health services were positively correlated with a higher house price, even if it was only a small correlation. Public hospitals (0,13), Pharmacies (0.18), GP and Dental (0,20), Other health services (0.22), Mental health (0.23), and Private hospitals (0.27) were all positively correlated with a higher house price by between 0.1 and 0.3 correlation. This indicates that people who are wealthier and are able to afford a house in a more expensive suburb are more able to access these sorts of services, many of which are health services. It is concerning that people in poorer communities are further away from these services and therefore are more likely to have difficulty accessing them. Based on our findings we would recommend that the government invest in health services in communities where house prices are lower so that everyone has equal access to these important services.

It is notable that education level (not finishing year 12 versus having a university degree) has such a strong correlation with house price. This implies that occupations that require less education are being paid significantly less and are being undervalued by society, even though many of these jobs are extremely essential. Train drivers, supermarket workers, maintenance workers and many other jobs do not require a university education but society would not function without them, and yet these people are living in less nice housing with access to fewer health services.

Outliers

Outliers were also identified in a number of columns, with 9 columns having more than 10% of their data be an outlier. The column with the most number of outliers was Community Health Centres, with 20.18% of its data being outliers. This is likely due to the location of most of the communities that were outliers being close to CBD, which has more resources such as community health centres. Since the outliers were the result of the skew in the allocation of health resources and not an error, we decided to not alter them in any way. Another column with a high number of outliers was Rural (%), with 19.01% of its data being outliers. This was due to the inclusion of rural suburbs in the dataset, which naturally had very high percentages for the amount of rural area in those communities, leading them to be outliers. We decided to not alter the data as we were not specifically focusing on metropolitan areas, and such a skew was natural given the inclusion of rural suburbs.

Modelling

We used supervised learning models (linear regression and decision trees) to model our data. This is because we wanted to see which factors would be the most useful in **predicting** house prices, and predictions are not possible with unsupervised learning models.

Linear Regression

A linear regression model is suitable at modelling relationships between various numeric variables and their impact on a continuous numeric variable. Thus it was highly suitable for analysis, as 87% of our data was numeric, and we sought to predict a continuous numeric variable - house price. Furthermore, linear regression offers an ability to analyse feature importance, thus offering an opportunity to analyse how each feature influences house prices in relation to our research question.

Multiple linear regression models were made in an attempt to reduce the mean squared error (MSE) as much as possible. In fitting each model, we used cross validation to prevent overfitting and assess the model's ability to generalise itself especially given the limited amount of data we had. This process can be seen in Tables 3 and 4 where the mean squared error and R-squared score for each fold are documented. The figures demonstrate the importance of cross validation as the model's performance varies depending on what test training split was used. This was achieved through dividing the dataset into training (80%) and validation (20%) for a total of 5 folds, thus enabling us to more accurately evaluate the model's performance in general rather than on a specific subset of the data.

Table 3: Mean squared error and R-squared scores for each fold with only numeric variables

Fold number	Mean Square Error	R-Squared Score
1	168162.7678	0.6589
2	221018.5052	0.5668
3	190599.6673	0.6620
4	209168.0946	0.6673
5	175498.6701	0.5610
Mean	192889.5410	0.6232

Table 4: Mean squared error and R-squared scores for each fold after feature selection

Fold number	Mean Square Error	R-Squared Score
1	132632.6973	0.7878
2	202582.2487	0.6360
3	160617.4828	0.7600

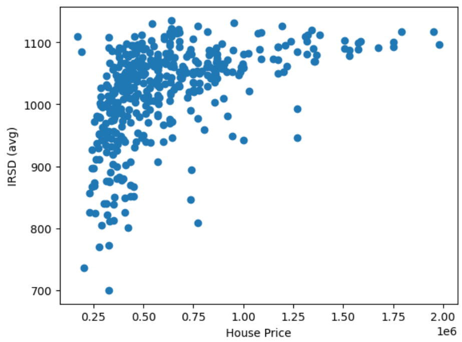
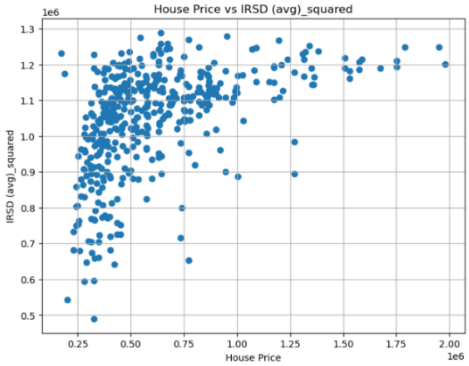
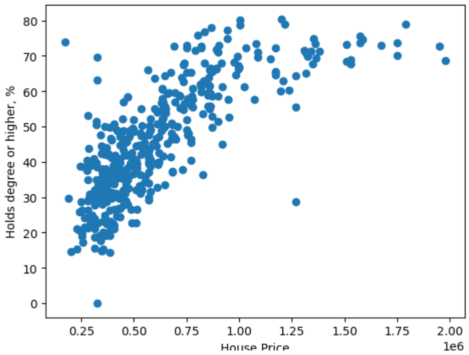
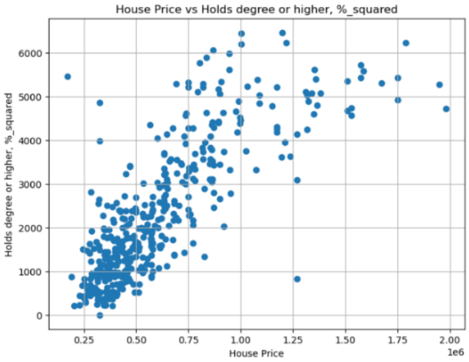
4	152128.6461	0.8240
5	141392.8729	0.7150
Mean	157870.790	0.7446

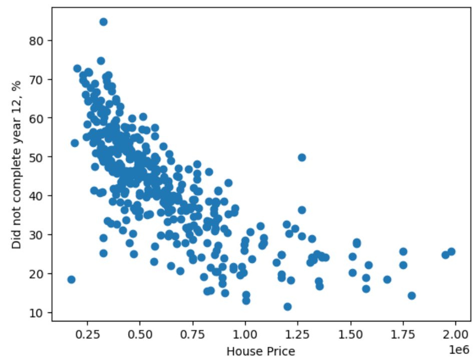
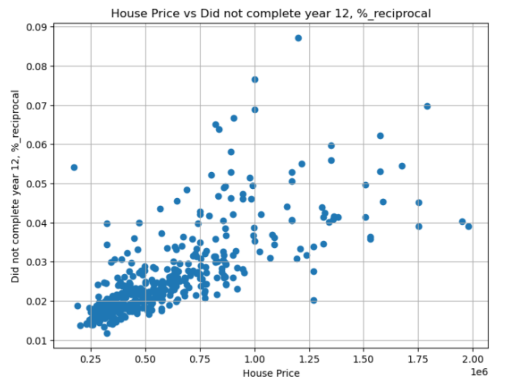
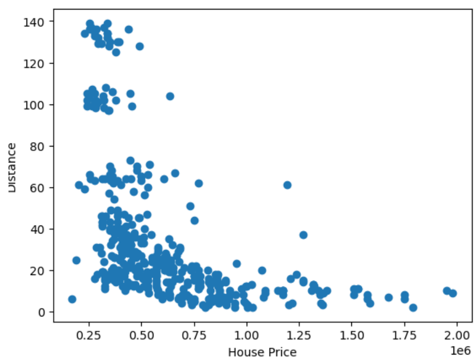
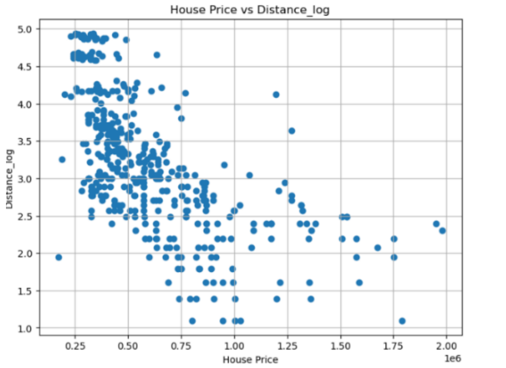
The initial model was made with only the numeric variables in the dataset, and had an average root mean square error (RMSE) of 192889.54. We decided to only use numeric values in this model because we wanted to make a simple initial model that could later be improved using encoding for categorical variables.

When investigating the columns that the model used, it was observed that columns which had a very low MI had a high feature importance. For example, ‘Number of Families’ had a normalised MI score of 0.0996, however it had a feature importance of -795446.54, meaning that the model was relying on features that had limited capacity to predict house price. This issue was tackled by implementing feature selection in the model so that the model uses more important features, hence improving the accuracy of its predictions.

Another linear regression model was then implemented, however this time, feature selection and variable transformations were implemented to improve it. Some variables had transformations applied in order to improve the MSE as linear regression models work best when the data is as linear as possible. Squared transformations were applied to ‘IRSD’, ‘Holds degree or higher%’, while a reciprocal transformation was applied to ‘Did not complete year 12%’ and a log transformation was applied to ‘Distance’. The effect of these transformations can be seen in Table 5.

Table 5: Table of scatterplots before and after transformation and difference in MI scores after transformation

Variable	Scatterplot before transformation	Scatterplot after transformation
IRSD (squared transformation) Increase in MI = 0.0027		
Holds degree or higher% (squared transformation) Increase in MI = 0.0232		

Did not complete year 12% (reciprocal transformation) Increase in MI = 0.0232		
Distance (log) (log transformation) Increase in MI = 0.0187		

Feature selection was applied by considering the highest MI scores, where the final features for the model included LGA (0.3624), Holds degree or higher% (0.3016), Did not complete year 12% (0.2973), Distance (0.2424) and Top occupation (0.1689), Population Density(0.1612) and IRSD (0.1596). We used high mutual information features in the linear regression model because they capture strong relationships with the target variable, improving the model's predictive power and accuracy. Since linear regression requires all numeric variables, the categorical variables (LGA and Top occupation) were encoded using one-hot encoding before applying the model.

All these changes decreased the new model’s RMSE to 157870.78 which is a decrease of around 33512.17, making our model more accurate. The R-squared scores also increased for the new model in comparison to the initial model, with the new model’s R-squared scores being 0.7445 and the initial model’s being 0.5583, which is a huge improvement. These improvements can also be seen in Tables 3 and 4 above.

R-squared scores were used as a measure of assessing the improvement in the model as they quantify the variability in house price that is explained by the model, therefore allowing us to see how accurate the model is in predicting the required values. We used mean squared error because it is an easy way to have one number that summarises the average error for all of the predicted values.

Holds degree or higher, %_squared had the highest feature importance of 139105.34 in the new model, and also had a very high pearson correlation coefficient value of 0.7645, indicating that it is a major factor in influencing house prices. It can be seen in Table 5 that house prices tend to increase with the percentage of people who hold a degree or higher, leading to the conclusion that people who have a degree or higher tend to earn more and/or are more easily able to get house loans, enabling them to buy more expensive houses. This is further supported by the feature importance of Did not complete year 12, %_reciprocal (-8410.13), which shows that the house prices decrease and the percentage of those who did not complete year 12 increases, since the number of people holding degrees or higher will be lower when the percentage of those who did not complete year 12 is higher.

Decision Tree

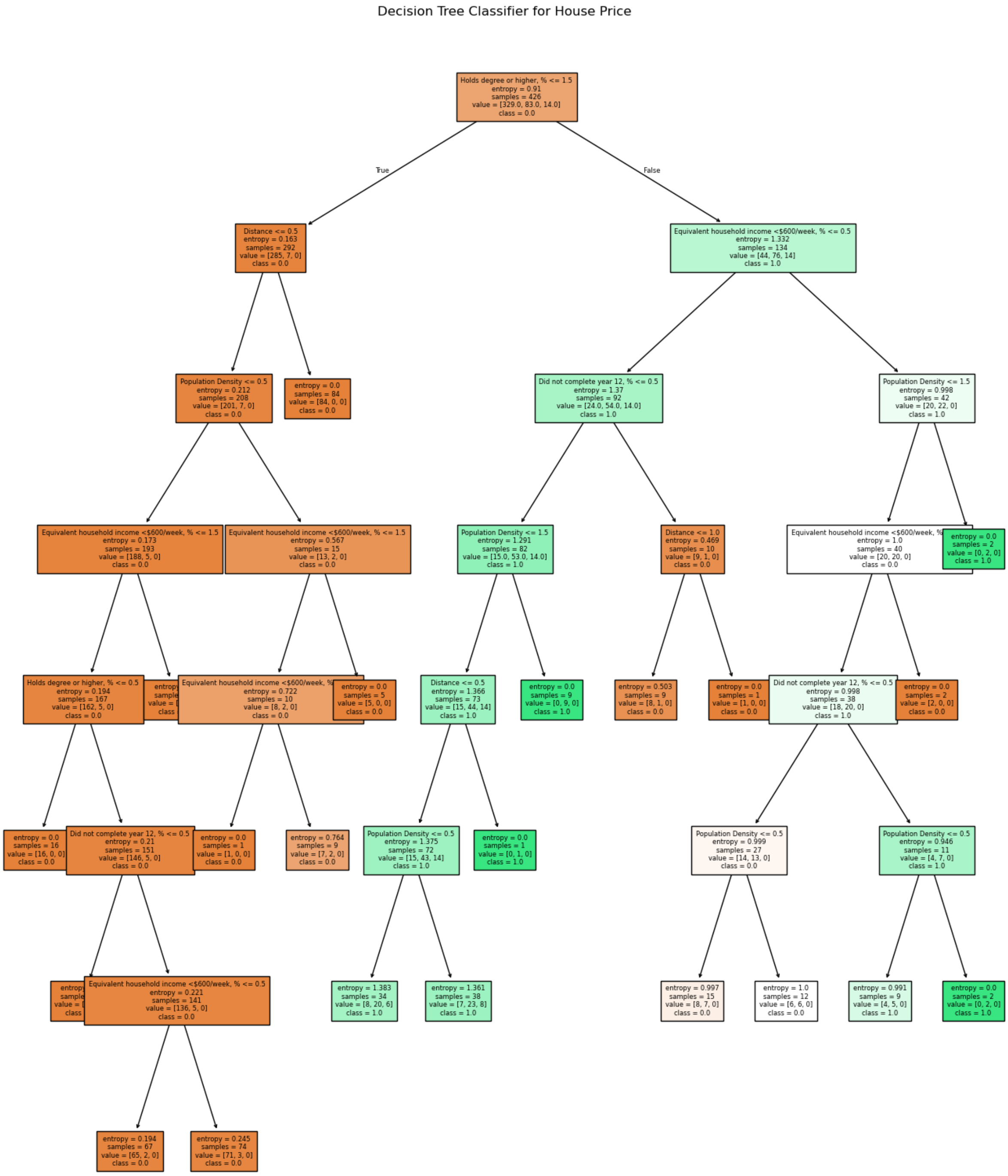

Decision Tree Classifier for House Price
Holds degree or higher, % <= 1.5 entropy = 0.91 samples = 426 value = [329.0, 83.0, 14.0] class = 0.0
True
False
Distance <= 0.5 entropy = 0.163 samples = 292 value = [285, 7, 0] class = 0.0
Equivalent household income <$600/week, % <= 0.5 entropy = 1.332 samples = 134 value = [44, 76, 14] class = 1.0
Population Density <= 0.5 entropy = 0.212 samples = 208 value = [201, 7, 0] class = 0.0
entropy = 0.0 samples = 84 value = [84, 0, 0] class = 0.0
Did not complete year 12, % <= 0.5 entropy = 1.37 samples = 92 value = [24.0, 54.0, 14.0] class = 1.0
Population Density <= 1.5 entropy = 0.998 samples = 42 value = [20, 22, 0] class = 1.0
Equivalent household income <$600/week, % <= 1.5 entropy = 0.173 samples = 193 value = [188, 5, 0] class = 0.0
Equivalent household income <$600/week, % <= 1.5 entropy = 0.567 samples = 15 value = [13, 2, 0] class = 0.0
Population Density <= 1.5 entropy = 1.291 samples = 82 value = [15.0, 53.0, 14.0] class = 1.0
Distance <= 1.0 entropy = 0.469 samples = 10 value = [9, 1, 0] class = 0.0
Equivalent household income <$600/week, % entropy = 1.0 samples = 40 value = [20, 20, 0] class = 0.0
entropy = 0.0 samples = 2 value = [0, 2, 0] class = 1.0
Holds degree or higher, % <= 0.5 entropy = 0.194 samples = 167 value = [162, 5, 0] class = 0.0
entropy sample value = [class
Equivalent household income <$600/week, % entropy = 0.722 samples = 10 value = [8, 2, 0] class = 0.0
entropy = 0.0 samples = 5 value = [5, 0, 0] class = 0.0
Distance <= 0.5 entropy = 1.366 samples = 73 value = [15, 44, 14] class = 1.0
entropy = 0.0 samples = 9 value = [0, 9, 0] class = 1.0
entropy = 0.503 samples = 9 value = [8, 1, 0] class = 0.0
entropy = 0.0 samples = 1 value = [1, 0, 0] class = 0.0
Did not complete year 12, % <= 0.5 entropy = 0.998 samples = 38 value = [18, 20, 0] class = 1.0
entropy = 0.0 samples = 2 value = [2, 0, 0] class = 0.0
entropy = 0.0 samples = 16 value = [16, 0, 0] class = 0.0
Did not complete year 12, % <= 0.5 entropy = 0.21 samples = 151 value = [146, 5, 0] class = 0.0
entropy = 0.0 samples = 1 value = [1, 0, 0] class = 0.0
entropy = 0.764 samples = 9 value = [7, 2, 0] class = 0.0
Population Density <= 0.5 entropy = 1.375 samples = 72 value = [15, 43, 14] class = 1.0
entropy = 0.0 samples = 1 value = [0, 1, 0] class = 1.0
Population Density <= 0.5 entropy = 0.999 samples = 27 value = [14, 13, 0] class = 0.0
Population Density <= 0.5 entropy = 0.946 samples = 11 value = [4, 7, 0] class = 1.0
entropy sample value = [class
Equivalent household income <$600/week, % <= 0.5 entropy = 0.221 samples = 141 value = [136, 5, 0] class = 0.0
entropy = 1.383 samples = 34 value = [8, 20, 6] class = 1.0
entropy = 1.361 samples = 38 value = [7, 23, 8] class = 1.0
entropy = 0.997 samples = 15 value = [8, 7, 0] class = 0.0
entropy = 1.0 samples = 12 value = [6, 6, 0] class = 0.0
entropy = 0.991 samples = 9 value = [4, 5, 0] class = 1.0
entropy = 0.0 samples = 2 value = [0, 2, 0] class = 1.0
entropy = 0.194 samples = 67 value = [65, 2, 0] class = 0.0
entropy = 0.245 samples = 74 value = [71, 3, 0] class = 0.0

Figure 1: Decision Tree

When making a decision tree we tried various combinations of factors to try and get the highest accuracy. In the end, the best one was the one that only considered the factors “Did not complete year 12, %” and “Holds university degree or higher, %”. It turned out that having more factors did not necessarily make the result better. This is useful for real life applications, as the best one requires the smallest amount of data in this case.

The first attempt at a decision tree predicts the price of a house based on the following factors: Did not complete year 12, %, Equivalent household income <$600/week, %, Population Density, 'Holds degree or higher, %. We chose these factors because they had a high mutual information with house price, meaning that they would help the decision tree to predict accurately.

We used cross-validation to check the accuracy of the decision tree. We chose this method because it allows you to test the model on different data each time, so you have a better idea of how good it actually is than if there was only one test. We chose accuracy as the measure of how good the tree is rather than precision because accuracy says how good it is overall rather than just how good it is at predicting “low” house price, for example. This is more useful given that we want a decision tree that works well overall. The accuracy of each fold can be seen in Table 6.

Table 6: Accuracy scores for each fold for decision tree

Fold Number	Accuracy score
1	0.872093023255814
2	0.8705882352941177
3	0.8588235294117647
4	0.8117647058823529
5	0.8470588235294118
Mean accuracy	0.852065663474692

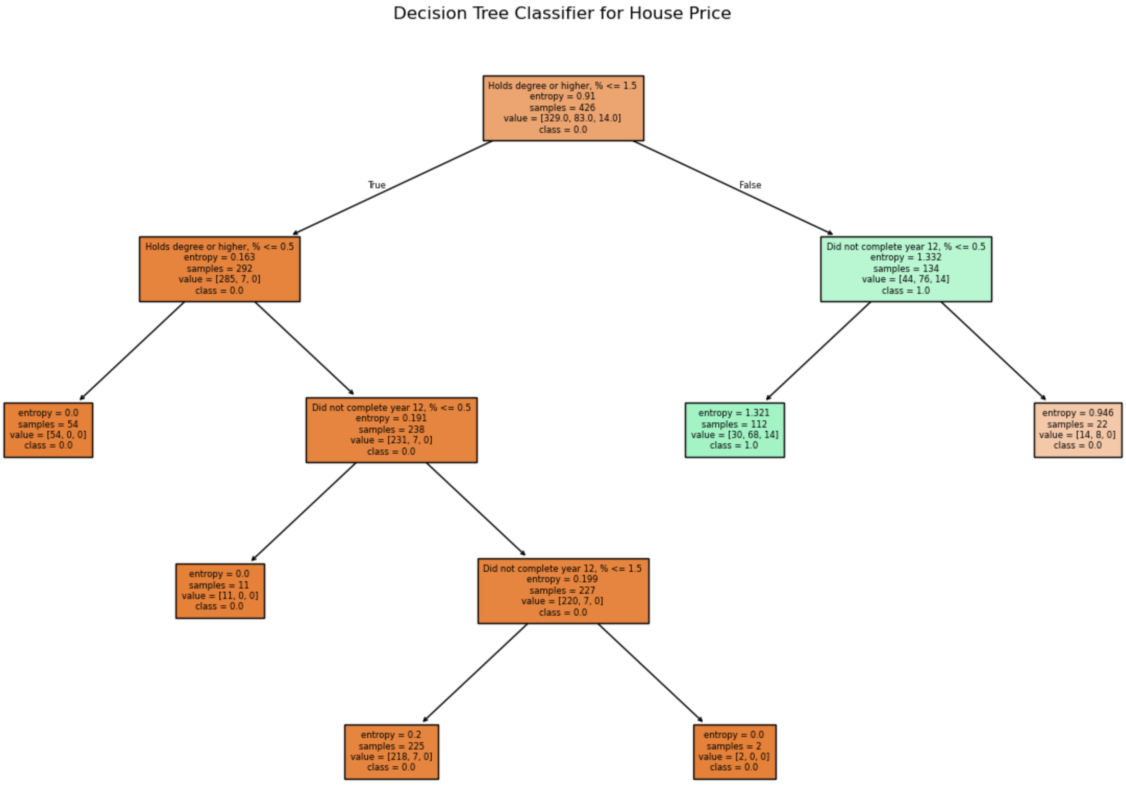


Figure 2: Decision Tree with only 2 factors

Figure 2 only uses 2 factors to predict house price, “Did not complete year 12, %” and “Holds university degree or higher, %”. The accuracy, using cross-fold validation can be seen in Table 7.

Table 7: Accuracy scores for each fold for decision tree with 2 factors

Fold Number	Accuracy score
1	0.9069767441860465
2	0.8470588235294118
3	0.8588235294117647
4	0.8117647058823529
5	0.8823529411764706

Mean accuracy	0.8613953488372094

This is a slight improvement on accuracy by just under 0.01. We believe that this tree gives a better accuracy than the other one because it only considers the factors that are best at predicting house price, rather than getting distracted by other less important factors.

The reason why we made a decision tree is because we wanted a model that would more easily include categorical variables and also that would show what is important clearly in a diagram. The decision tree is useful for visualising what factors are important, however the purpose of including more categorical variables turned out not to be useful. We also made a decision tree that included top language spoken, which was a categorical variable, but this did not improve the accuracy so we took it out. A limitation of the decision tree is that it does not predict the exact house price, instead it predicts a range that the house price will fall into. This is useful classifying house prices as low, medium or high but not as useful for specific values. One thing that is notable about this decision tree is that there are a lot of houses that are classified in low and a few medium but it misses out on the high values. This could be due to the factors chosen, however we tried it with other combinations of factors and this was the model with the highest accuracy.

Discussion and Interpretation

Our results showed that education influences the house price people can afford to a significant extent. They also showed that good health services in a suburb is associated with increased house price, but this was not nearly as significant as education.

These results show that if people do not have equal access to quality education, they will not have an equal opportunity to live in a nice suburb and have good access to essential health services. Based on this, we recommend that the government ensure that people in all suburbs have equal access to good education. This would involve investing more in state schools in poorer communities, and making sure that the school is in a good condition, that teachers there are paid enough and are able to do their job, that teachers are not being overworked due to understaffing in public schools, and that children are supported in other ways that allow them to not be held back in school (for example good health care so that sickness does not get in the way).

Another finding of the data was that the presence of almost all health services in a suburb was positively correlated with higher house price. This indicates that people who are wealthier and are able to afford a house in a more expensive suburb are more able to access these sorts of services, many of which are health services. It is concerning that people in poorer communities are further away from these services and therefore are more likely to have difficulty accessing them. Based on our findings we would recommend that the government invest in health services in communities where house prices are lower so that everyone has equal access to these important services.

We recommend that more government funding is put into ensuring equal and good quality education for everyone, and also healthcare services. This money could come from various places, including:

- $360 billion spent on nuclear submarines in AUKUS deal could be used on education and healthcare
- $85 million PLC spent on a swimming pool recently
- Taxing biollionaires

The government could choose which funds they want to redirect based on their priorities, or use all of them to massively improve education, healthcare, and anything else that is contributing to inequality.

Limitations and Improvement Opportunities

One way that the results could be improved is by improving the matching suburbs function. Although this matches the majority of suburbs, there are still some that it misses. For example it does not match “Ascot” with “Ascot (Greater Bendigo), which are probably the same suburb. This could be done by making a matching function that ignores everything in brackets. Matching more suburbs would mean more data and therefore a more certain result.

A limitation of our entire analysis is that the data we used was from 2013. Housing affordability has gotten significantly worse since then, so these statistics do not necessarily represent the current situation. This being said, it is likely still the case that people with a better education have better housing, so our results are still relevant in that sense. It would be good to repeat the analysis we have done here with more recent data so that we can see if there are new factors that are influencing what housing people can afford and address those issues.

An assumption that was made in this report is that the data we were working with was accurate and representative of the Victorian population. We believe that this is a relatively safe assumption as it is from the government and they have a lot of resources to collect accurate data, and are less likely to have a sectional interest in it than a private company. This being said, the data could be more biased towards people who are more able to fill out the census, such as people who have access to a computer or live close to a post office.

For linear regression, we assumed that the variables are independent and that they have a linear relationship with the target variable, house price. This is a potential weakness in the model because it is plausible that many of the variables are not independent and thus may vary with each other. Furthermore, in our analysis, we used features with the highest MI to include within our model, however, a high MI can imply a strong nonlinear relationship which may be incompatible with the model. Thus, future feature selection could involve placing more emphasis on variables that have a high Pearson's correlation coefficient with house price, which better gives a measure of linearity. A limitation in the model is that we assumed that each feature has equal variance, which they almost definitely don't.

The decision tree is limited in its ability to predict an exact value for a house price. It can only predict a range, and we found that including more factors beyond the few most significant ones did not make much difference to the accuracy. The aim of the decision tree was to include categorical variables such as top language spoken more easily than the linear regression, however since numerical factors such as education level ended up influencing house price the most, it did not add as much. We believe that out of the models we made, linear regression is the most useful one for predicting house price.

Another limitation was the use of the MI function. As 87% of our data was continuous and numerical, we had to first discretise it into bins before being able to apply the normalised MI function. We chose to use 15 bins to discretise our data, however, this was a random value based on our data and may have not been an appropriate value to use, leading to results that may not be considered as the optimal results. We chose uniform binning but it would be good to try the regression again with equal frequency binning to see if this improves it.

Multiple columns were also discarded from the Victorian Communities dataset based on manual inspection at the beginning. We believe that this is also a limitation, as repeating this analysis with all the columns would likely lead to an interesting analysis, and our results and conclusions could potentially change due to those columns.

For future research, it would be better to have more detailed data on the average income for each occupation, and what jobs people with different levels of education are likely to do. It would also be useful to analyse data about current school funding and the number of private schools in an area versus the number of state schools. This would give more insights on how equal opportunity to good education could be improved, and where funds could be redirected.

Conclusion

To answer our research question "To what extent do the features in the Victorian Communities Dataset predict housing prices?", it can be concluded from our key findings in this analysis that the Victorian Communities dataset does indeed contain variables that can predict housing prices with at least 70% accuracy, with the prominent features being LGA, percentage of people who hold a degree or higher, percentage of people who did not finish year 12, and distance from the CBD. It can also be concluded from our Decision Tree model that there is no need to use huge datasets with more than 100 columns in order to somewhat accurately predict housing prices. The disparity in housing prices due to location is an area that the government can dedicate more research to in order to find the root cause, as it is an issue that persists in 2024, and it would also be suggestible for the government to cater more resources to improving the quality of life for people who do not receive the same services as others due to their educational background.

References (Times New Roman 12, Bold, Justified Left, Single Line, 0pt before 6pt after)

Capone, A. (2024, March 14). REIV: Melbourne house prices tend to drop about $30,000 for every 1km away from CBD - realestate.com.au. www.realestate.com.au. https://www.realestate.com.au/news/reiv-melbourne-house-prices-tend-to-drop-about-30000-for-every-1km-away-from-cbd/

Readfearn, G. (2023, September 2), Australian households on six-figure incomes can now only afford 13% of homes, The Guardian, https://www.theguardian.com/australia-news/2023/sep/02/australian-households-on-six-figure-incomes-can-now-only-afford-13-of-homes