STAT2004

Analytics for Observational Data

Semester 2, 2022

Final Project Report

Perth House Price Analysis

Eric Kian Lik Lau, 19758301, Data Science & Finance, e.g. Bachelor of Science and Commerce

Maksim Dmitriev, 19454874, Astrophysics, Bachelor of Science

1. Declaration

The work presented in this report is my/our own work and all references are duly acknowledged.

This work has not been submitted, in whole or in part, in respect of any academic award at Curtin University or elsewhere.

Makism Dmitriev

25/10/2022

Eric Lau

25/10/2022

1. Introduction/Objective

This project aims to utilise in-built functions of R, coupled with several open-source packages to analyse a dataset containing information on the real estate of Western Australia.

The objective of this analysis is to implement dimensionality reduction techniques such as principal component analysis, and factor analysis. Dimensionality reduction techniques can help reduce the number of attributes to a manageable amount that can represent all the attributes sufficiently. This analysis aims to utilise principal component analysis to explore the most prominent variables that can explain the most variance in the dataset. Moreover, factor analysis was also implemented to explore the unobservable correlations between attributes.

Another objective of this analysis is to model the response variable (property sale price) with explanatory variables such as floor area, land size, number of rooms, and etc. Machine learning techniques such as Extreme Gradient Boosting (XGBoost) and Random Forest were implemented to achieve this task to model the property price. Lastly, the most important features that model the price of a property as determined by XGBoost and Random Forest were compared.

1.1 Perth House Prices Dataset

The dataset is called "Perth House Prices" and it was extracted from Kaggle (Zainal, 2019). This dataset contains 19 columns/attributes and 33,657 observations of properties. The 19 attributes recorded are: the address, suburb, price, number of bedrooms, garages and bathrooms, land area, floor area, build year, CBD distance, nearest station name, nearest station distance, date sold, postcode, longitude, latitude, nearest school name, distance and its rank. The 33,657 observed properties are houses sold within WA containing a variety such as town houses, farm lands, single-detached and etc.

2. Methods

2.1 Random Forest

Random Forest is a machine learning method that can be used for regression and classification tasks (Mantas et al., 2019). Random Forest is an exceptional machine learning technique as it is a meta classifier that uses the ensemble learning technique of bagging (Mantas et al, 2019). The basis of Random Forest is that it forms a collection of weak learners such as decision trees to converge its predictions into a singular strong prediction model (Breitman, 2001). The bagging algorithm bootstraps the training set into different smaller sets for independent decision trees to learn (Breitman, 2001). In each decision tree, the predictors and node splitting criteria are independently selected. Then the outputs of each independent decision tree are combined to yield a single output (Breitman 2001). The R package "randomForest" will be used in this part of the analysis.

2.2 XGBoost

XGBoost (XGB), also known as Extreme Gradient Boosting is also an ensemble machine learning method that also utilises decision trees (Chen, 2016). However, the difference between XGBoost is that it is a boosting ensemble machine learning algorithm instead of bagging. The decision trees in Random Forest learn concurrently and produce outputs that are combined at the end to produce a single output. Conversely, the decision trees in XGBoost are constructed sequentially and iteratively after one another. The newly generated decision tree is improved and boosted upon the previous decision tree as it focuses more on improving the previous prediction

error (Brownlee, 2021). Then, to produce a single output, XGBoost combines predictions by weighing the averages of the constructed decision trees (Brownlee, 2021). The R-CRAN library has a "XGBoost" package that will be implemented in this analysis.

In this analysis, as the property prices are a continuous numerical value, the machine learning method is set to regression instead of solving a classification problem. The ensemble learning technique such as XGBoost and Random Forest will be fitted to predict property prices based on attributes such as the number of bedrooms and bathrooms, distance to CBD, land area, and etc. The performance of these two methods will be compared and the most important attributes that determine a property price will be investigated

2.3 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique that summarises the variables that represent the multivariate data as a smaller set of principal components (PCs) (Beatti, 2021). These principal components can be used for further analysis and visualisation. PCs are linear combinations of the attributes. Most of the time, PCA reduces the amount of attributes into principal components that can define a model plane. The base R command "princomp" will be used for this analysis.

In this case, PCA was used as an exploratory technique and is applied in order to probe the data and extract information that could help to form a more refined model.

2.4 Factor Analysis

Factor Analysis (FA) is a more sophisticated version of PCA. This technique is used to reduce a large number of variables into fewer factors, underlying unobserved random variables (Goretzko, 2022). Such grouping is based on common variance (i.e. the variables within a factor are highly correlated among each other). The approach has provided an insight on variables and the hidden correlation between them. The base R command "factanal" will be implemented in this part of the analysis.

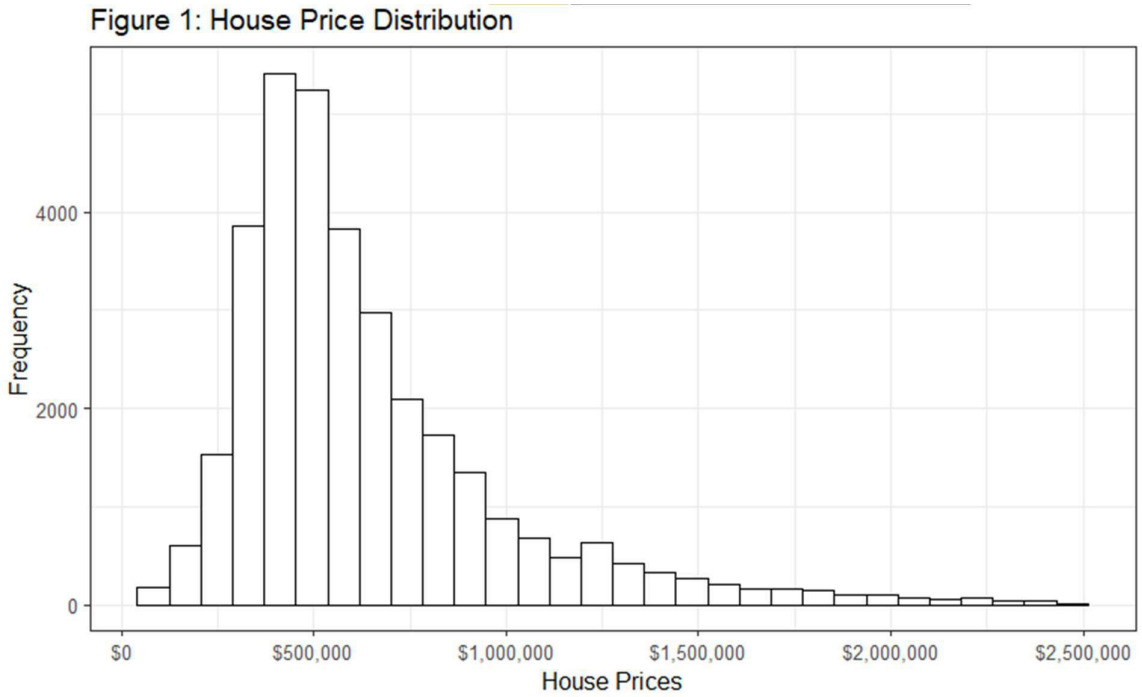
3. Results

3.1 Exploratory Data Analysis

Firstly, it is important to visualise the dataset with exploratory data analysis to plot important variables and identify interesting trends and characteristics of the dataset.

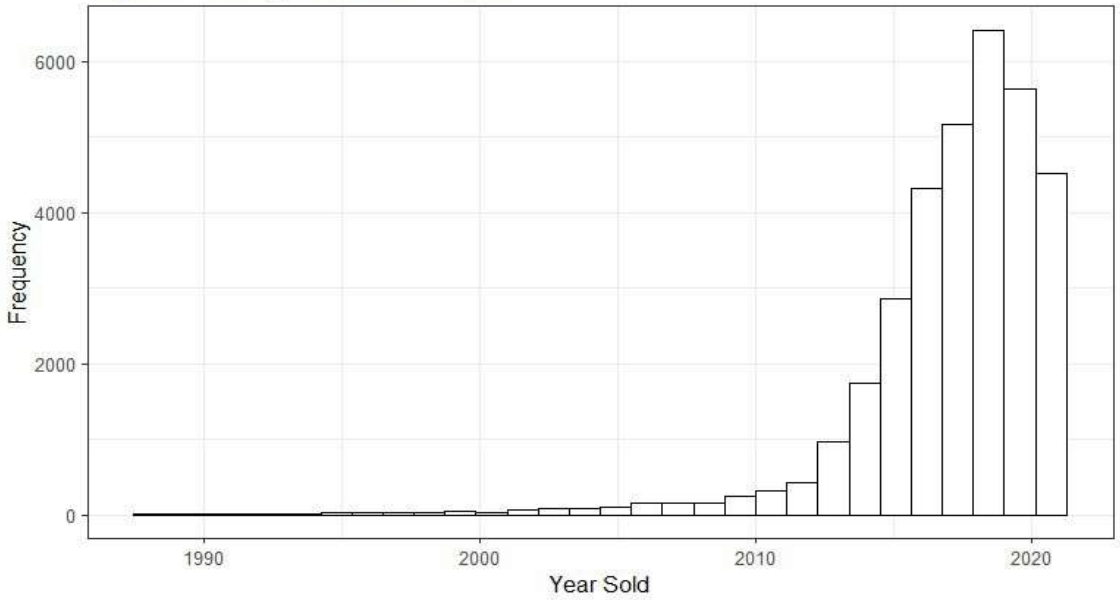
The response variable of this analysis is **Price**. Therefore, it is important to visualise the distribution of the response variable to determine what model is appropriate to fit this response variable. Below is the histogram plot of the property prices in the dataset:

Figure 1: House Price Distribution



As can be seen from the histogram above, the property prices are heavily skewed to the right. This skewness is caused by observations of luxurious properties that are in the price range over $2,000,000. However, the majority of the properties in the dataset is around the ranges of $300,000 to $750,000. This histogram is unimodal with a mode price of approximately $500,000. The histogram does not appear to be normally distributed. Hence, statistical regression models such as multiple linear regression will not be viable without normalisation. Nevertheless, machine learning methods such as XGBoost and Random Forest are robust to this type of skewness, hence can be applied in this analysis.

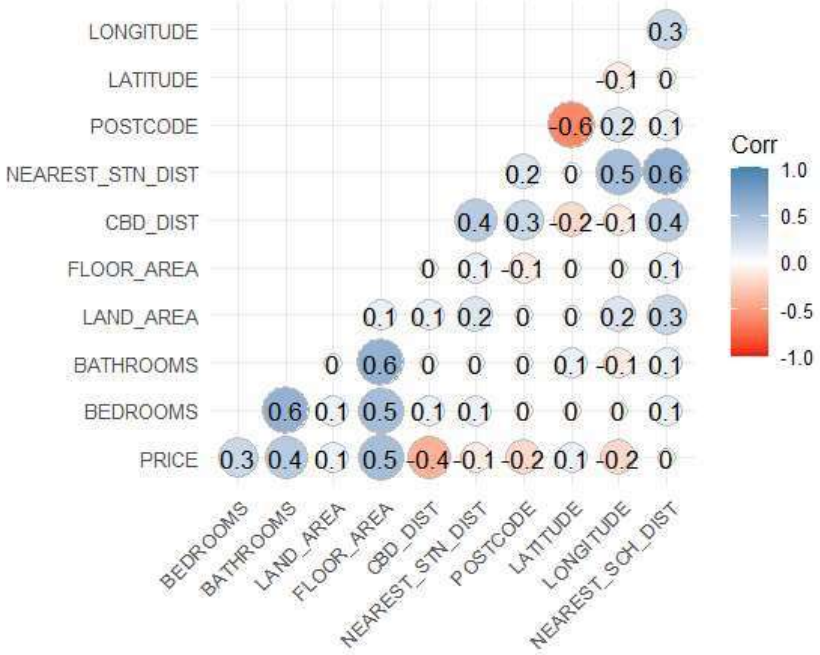
Figure 2: Histogram of Years Sold



The histogram above plots the years the properties were sold. The histogram is skewed to the left with some properties sold in the 1980s and majority of the properties being sold around the periods of 2010 to 2020. This histogram suggests to us that the dataset is up to date as most property sales are recent. This also means that the price of the properties would account some portion of inflation and not outdated.

This next section explores the correlation between variables and the response variable.

Figure 3: House Prices Correlation Plot

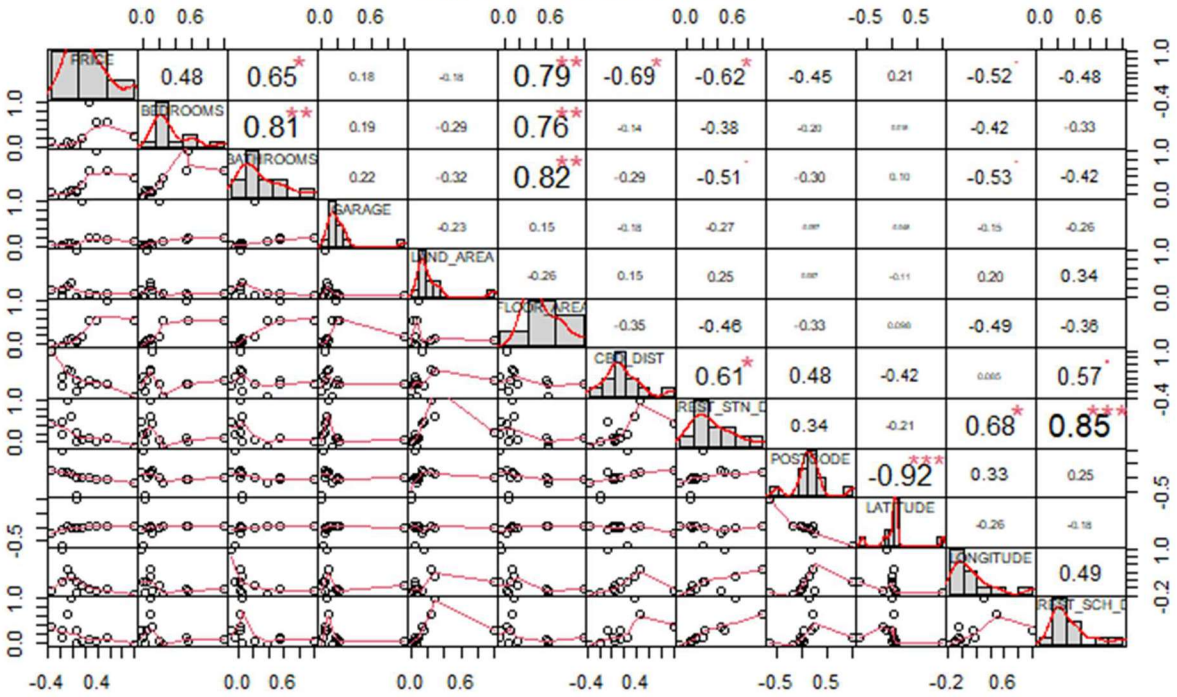


The last row of the correlation plot is the most important row to analyse as it shows the correlation between explanatory variables and the response variable. As can be seen from the correlation plot, the attributes such as **Number of bedrooms** and **Number of bathrooms**, **Land area**, and **Floor area** have a positive correlation with **Price**. This means that as the number of bedrooms and bathrooms and the size of the land and floor of the property increase, the price also increases. This correlation is as expected as the bigger the property, the more it will cost. The most positively correlated variable with **Price** is **Floor Area** with a Pearson correlation coefficient of 0.5, which might suggest an important variable in the model. Some variables are negatively correlated with **Price,** such as **Distance to CBD**, and **Nearest station distance** with a coefficient of -0.4 and -0.1 respectively. This means that as the property is closer to the CBD or a train station, the price of the property is also lower. This might be the case as the closer to the train station or CBD, the noise pollution is higher which devalues a property.

3.2 Variable Correlations

Principal Component Analysis is the next obvious step in learning more about the data and obtaining a better understanding of the parameters selected to represent the house pricing.

Figure 4: Correlation Chart



The two methods of visualisation shown above (Fig 3 & 4), demonstrate a strong correlation between a number of parameters:

- **Number of Bathrooms VS Number of Bedrooms:** a correlation noted is straightforward and is responsible for the house layout. Both variables shown to be strongly correlated with the **Floor Area**. The 3 variables could be said to be representative of **HOUSE SIZE**.
- **Distance to the CBD VS Distance to the Nearest Train Station**: once again, the causal explanation is easily obtainable. The further away a property is from the city centre, the larger block size / house is expected to be, which could be by a number of factors, such as subdivision rates, desirability and utility of the house package.
- **Latitude VS Postcode**: As mentioned earlier, a peculiar negative correlation is noted between geospatial location of the property and its areal postcode. Latitude is a set of spherical coordinate lines that traverse the globe parallel to the equatorial line, since Australia is located in the Southern Hemisphere, the Latitude is negative. Almost perfect negative correlation signalizing of the fact that Western Australian Postcodes are assigned from South to North, i.e. the larger the latitude the more North a property is.
- **Nearest Station Distance VS Longitude**: **Longitude**, on the other hand, is a vertical coordinate, the train lines in WA span along the coastline and the train connection into the country is not particularly well developed. Hence, the houses based in more in-land suburbs are expected to be on average further away from a nearest train station than houses from closer-to-coast.
- **Nearest Station Distance VS Nearest School Distance**: A very similar situation is observed for this pair of variables, less developed, more rural in-land suburbs are less industrialised on average and hence the infrastructure such as schools is going to be more sparse than in the more population dense areas.

From the relationship between **Postcode** and **geo-spatial coordinates**, the dimensionality of the data could be reduced even further. Since, the interest lies in predicting prices for the houses/townhouses/units and apartments, it

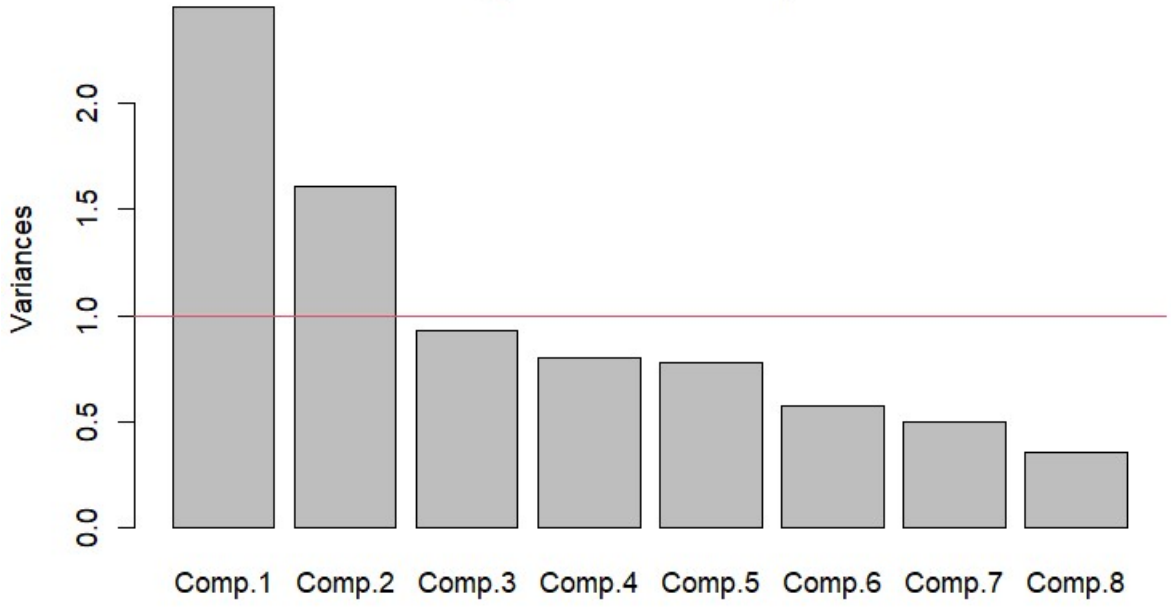
only makes sense to remove the farmland properties that skews the distribution. More of this will be discussed in the machine learning data pre-processing section.

3.3 Principal Component Analysis

The covariance matrix is not particularly useful since the attributes in this dataset have completely different units and of dissimilar scale. For instance, bedrooms and bathrooms are discrete counts, and on the other hand, floor and land area are continuous measurements. This means that the covariance matrix will not be suitable to use for dimensional reduction methods such as principal component analysis and factor analysis as attributes with small variances will be unfairly represented. Therefore, the correlation matrix would be more appropriate to use for dimensional reduction methods.

Since the variance is large and units are different, PCA performed on the covariance matrix would be nonsensical. Therefore, **scaling** or **standardisation** is required, alternatively PCA could be performed using correlation matrix. Based on the output, it could be estimated how many components are appropriate to use. For this step, the target variable (**Price**) could be removed to allow for inferential analysis of Principal Components. Moreover, it is important to note that PCA could only be performed on numerical values so all categorical variables such as **Postcode** and **Suburb** are removed. It was decided to further reduce the dimension of the data by two and remove the Latitudinal and Longitudinal coordinates, as it does meaningless to include these variables.

Figure 5: PCA Screeplot



```
Call:
princomp(x = HousePricesNumeric_reduced, cor = TRUE)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
1.5651428 1.2672377 0.9641268 0.8949858 0.8839620 0.7603045 0.7040741 0.5989359

 8  variables and  22493 observations.
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"

Loadings:
              Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
BEDROOMS       0.501         0.114  0.229  0.107         0.760  0.301
BATHROOMS      0.443        -0.218  0.275  0.594 -0.145 -0.225 -0.498
GARAGE         0.260        -0.848        -0.448
LAND_AREA      0.441         0.367 -0.248 -0.526  0.121        -0.557
FLOOR_AREA     0.500  0.156  0.196 -0.229                -0.540  0.580
CBD_DIST              -0.519  0.161  0.715 -0.262  0.203 -0.268
NEAREST_STN_DIST  0.153 -0.595        -0.244        -0.746
NEAREST_SCH_DIST       -0.575 -0.147 -0.433  0.285  0.602

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     1.565143 1.2672377 0.9641268 0.8949858 0.88396204 0.76030452 0.70407414 0.59893595
Proportion of Variance 0.306209 0.2007364 0.1161926 0.1001250 0.09767361 0.07225787 0.06196505 0.04484053
Cumulative Proportion  0.306209 0.5069454 0.6231380 0.7232629 0.82093655 0.89319442 0.95515947 1.00000000
```

The table above is the PCA output for the dataset. As can be seen form the table above, dimensions of this data can be represented by the first six PCAs (with 89.31% of total variance) or by the first seven PCAs (with 95.51% of total variance).

Figure 5 shows no clear domination of PCs, and only the first two PC has significant standard deviation (more than 1), meaning that two PC is sufficient in representing majority of the variance (50.69%). To reinforce this point, referring to the screeplot at figure 5, only the first two bars representing PC1 and PC2 exceed the significant line of variance more than 1. However, explaining only 50.69% of variance is insufficient. Therefore, more 2 more PCs are added to explain 72.33% of total variance.

The major variables of the four main principal component could be outlined as such:

- PC1 appears to be representative of "**House Size"** as it is dominated by attributes such as **Bedrooms, Bathrooms, Land Area** and **Floor Area.**
- Principle Component two (PC2) is ruled by **CBD Distance, Nearest Station Distance** and **Nearest School Distance,** and hence could be said to be representative of **"Convenience".**
- Component three (PC3) mostly consists of **Garage** which can be representative of **"Car Storage Space"**
- Component four (PC4) has a strong contribution of **CBD Distance** which might indicate a PC of **"Remoteness"**.

The remainder of the PC's from PC five to eight account for less than 10% of variance. Therefore, it was decided to be insignificant and disregarded.

3.4 Factor Analysis

Factor Analysis (FA) involves more assumptions to be made compared to the PCA. FA provides a dimensional reduction technique used to investigate strong correlation between variables and reduce it down to a single underlying factor. Such grouping of correlated variables into factors will reduce the dimensions of the dataset.

The next step would be to obtain a solution based on the reduced data frame of house price-related-variables. The maximum possible number of CMs is half the number of parameters, so a choice of **m = 5** factors is reasonable. From the result for 5 factors, under the null hypothesis that 5 factors are sufficient, the chi-square test statistics 73.61 on 1 degree of freedom. The **p-value** is 9.51e-18 .Hence, the null hypothesis is rejected as the p-value is extremely small. The factors cannot be extended any further. Hence, a different approach is required.

Similar to the principal component analysis, the variables have different scales and will require standardisation. Therefore, the factor analysis will performed on the correlation matrix instead of the covariance matrix.

```
Call:
factanal(x = HousePricesNumeric_reduced, factors = 5, method = "mle")

Uniquenesses:
        BEDROOMS        BATHROOMS           GARAGE        LAND_AREA       FLOOR_AREA         CBD_DIST NEAREST_STN_DIST
           0.433            0.432            0.005            0.005            0.439            0.534            0.331
NEAREST_SCH_DIST
           0.375

Loadings:
                Factor1 Factor2 Factor3 Factor4 Factor5
BEDROOMS         0.741                           0.114
BATHROOMS        0.748
GARAGE           0.170                   0.982
LAND_AREA                0.185   0.978
FLOOR_AREA       0.731   0.115
CBD_DIST                 0.337                   0.589
NEAREST_STN_DIST         0.742                   0.328
POSTCODE                                         0.387
NEAREST_SCH_DIST         0.752   0.139           0.185

               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      1.688   1.288   0.986   0.980   0.665
Proportion Var   0.188   0.143   0.110   0.109   0.074
Cumulative Var   0.188   0.331   0.440   0.549   0.623

Test of the hypothesis that 5 factors are sufficient.
The chi square statistic is 73.61 on 1 degree of freedom.
The p-value is 9.51e-18
```

The table above is the output for factor analysis. Referring to table above, examining the estimated factor loadings, several conclusions could be drawn:

1. The first factor is dominated by **Number of Bedrooms (0.741), Number of Bathrooms (0.748) and Floor Area (0.731)**; perhaps this factor could be labelled as "**Property Size"** and could be used to explain how well a house was designed and how convenient it could be presented to the buyer.

2. Factor number two comprised of **CBD Distance (0.337), Nearest Station Distance (0.742), and Nearest School Distance (0.752)**, and hence presents the driving factor for young families. The factor could be said to be representative of "**Convenience**".

3. The third factor reflects utility for certain groups of population, the highest loadings are for **Land Area (0.978)** and **Nearest School Distance (0.139)**. This might be labelled as "**Investment Value"**, the factor could be important for the investors, builders and sub-dividers or potential landlords.

4. Factor four is only **Garage (0.982)** and could correspond to "**Storage Available".**

5. Factor number five is mostly **CBD Distance (0.589)** and **Postcode (0.387)** and could be representative of **"Remoteness"**.

From the above it could be concluded that the variables represented by 5 factors is insufficient as the p-value is negligible.

This should not be surprising, as different age and demographics groups use different reasoning for property selection. The variables already appear to be fully defined factors that affect the property cost in its own way.

3.5.1 Machine Learning Results - Data Pre-processing

The objective of machine learning is to fit a model predicting the response variable of property price using explanatory variables and find he most important variable that determines the sale price.

Firstly, the dataset is cleaned and pre-processed for training to optimise performance. The data was initially arranged alphabetically, so randomising the rows will remove any possible random bias and row dependencies.

Next, conducting a feature selection to select important attributes of the dataset that could be useful for training and remove any noisy attributes. For instance, categorical variables were removed in this analysis. The categorical variables in this dataset do not present any information useful for regression, e.g. the **Address** of the property, the name of the **Suburb**, **Nearest Train Station**, **Date of Sale**, **Nearest School** provide very little information as it serves as more of an identifier of the property. The generalised versions of the categorical attributes, such as the **Postcode, Build Year, Distance to the Nearest Train Station,** and **Distance to the Nearest School** present more useful information that could be used to predict house prices. Moreover, the data was cleaned by further reducing redundant attributes by removing attributes such as the **Postcode**, **Latitude**, **Longitude**, as these attributes are not characteristics of a property that can predict its price. The **Garage** attribute has missing values (NA). After comparing the data on several specific properties available online, it was found that properties in the dataset with missing values for garages refer to properties with no garages. Therefore, the missing values of garages were imputed with 0, assuming that all missing garage values refer to 0 garages.

Furthermore, the rank of the nearest school attribute was also removed as it has a there is a large number of missing rows (32.54% instances are missing). It was impossible to impute the missing school ranks as some schools are not in any school zones and some rural schools are not in the ranking system. Therefore, the decision was made to remove this attribute.

One issue found with the dataset is that it included farm properties. Farm properties are characterised as having significantly large land area and minute floor area at an abnormally cheap price. These farm properties are essentially outliers as the size of the land area does not correspond with its cheap price. Therefore, farm lands was excluded in this analysis because these properties account for a minor proportion of the dataset and the aim is to predict the prices of metropolitan family houses as they are more relevant. A map of WA postcodes was investigated and it could be seen that postcodes that does not start with 60 or 61 will correspond to a more remote area. The remote area properties are expected to have completely different price valuation which will not be investigated in this analysis. The criterion for excluding farm properties is to only include properties with the postcodes between 6000 and 6200 as these postcodes encompass most metropolitan suburbs and do not include rural farm suburbs. The second criteria to cull the farm properties is to remove any observations that have **Land Area** 10 times its **Floor Area**.

The attributes **Nearest Station** and **Nearest School Distance** and **CBD Distance** are of different scale. CBD distance is measured in kilometres and nearest station and school distance is measured in metres. By making the **CBD Distance**, and **Nearest School** and **Station distances** to be on the same scale of kilometres to keep the measurements consistent and at a smaller magnitude. Lastly, to reduce the complexity of the target variable of **Price**, the scale was reduced to thousands of AUD. The data pre-processing step reduced the original dataset of 33,656 observations and 18 variables to only 29,087 observations and 9 variables.

3.5.2 Machine Learning Results - Methodology

Fitting XGBoost and Random Forest require different parameters as they have different underlying algorithms. For instance, XGBoost requires parameters such as the depth of decision trees and number of iterations it runs for. Whereas Random Forest requires parameters such as the number of trees it generates. In this experiment, XGBoost is set to propagate for a maximum depth of 5 leaves and train for 100 iterations. On the other hand, Random Forest is set to also generate 100 trees. These parameters are decided as it ensures that both machine learning model generate the same amount of decision trees and train with approximately the same resources. Therefore, this allows for fair comparison of the models. For instance, the XGBoost running for 100 iterations will generate 100 decision trees similar to random Forest generating 100 decision trees. This means that XGBoost and Random Forest both have the same number of decision trees and we can compare the prediction performance without bias to one model.

The dataset is split into **training** and **testing** parts. As a rule of thumb, the data is split into 85% training set and 15% testing set. Information about the target variable (**House Price**) was removed from the **testing** set to be used as ground-truth. Both Random Forest and XGBoost will train on the same training set and test on the same testing set for fair comparison of performance.

3.5.3 Machine Learning Model Performance - XGBoost

XGBoost is set to train for 100 iterations like mentioned. After 100 iterations, the training Root Mean Squared Error (RMSE) is **$158.72 thousand**.

Below is a plot depicting the exponential decrease in training RMSE at each iteration.

Figure 6. XGBoost Training Iterations

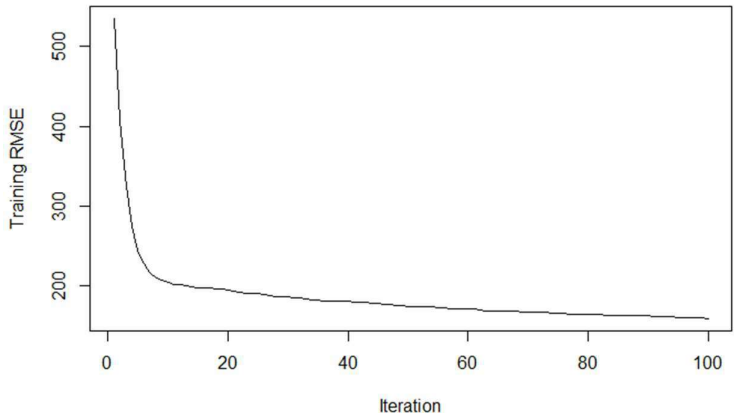
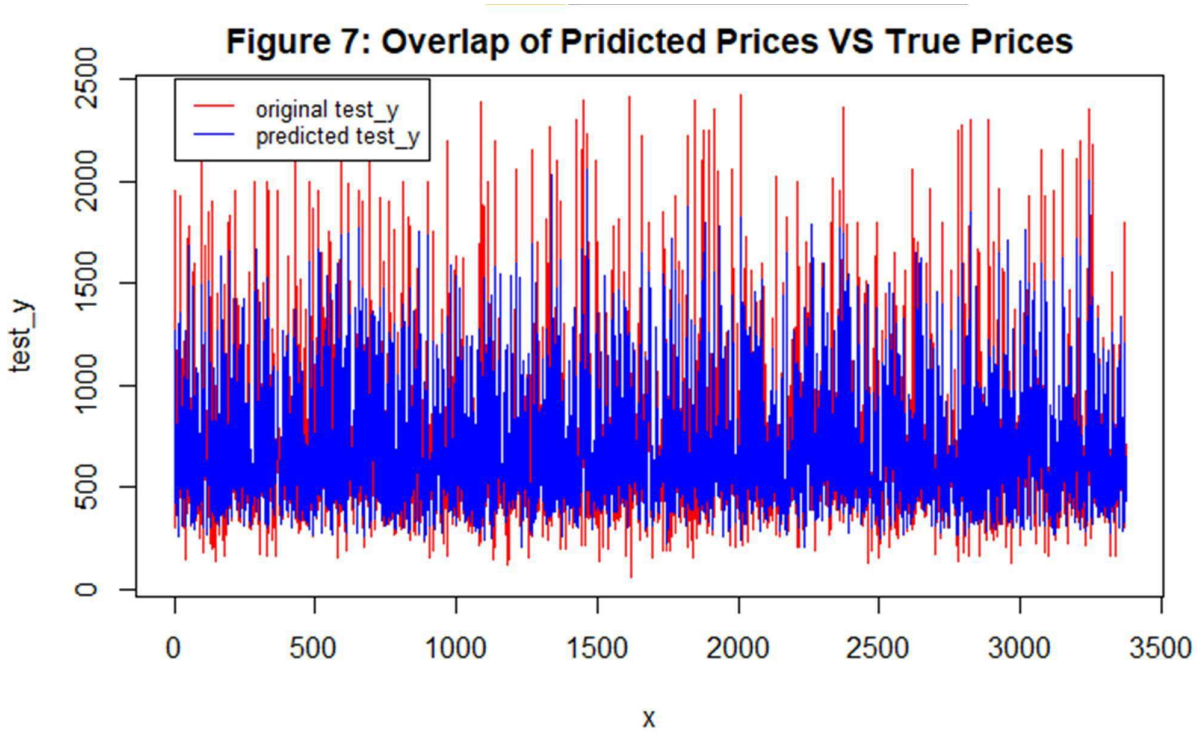
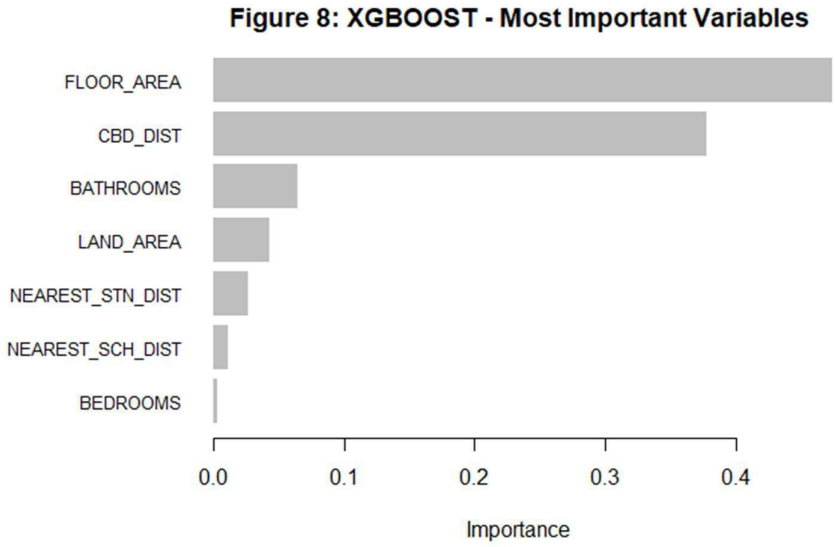


Figure 6 shows that the training RMSE of XGBoost decreases rapidly after each iterations. This can be seen with the exponential decrease in graph. This suggests that XGBoost is able to predict the property prices given the training data.

Furthermore, referring to Figure 7 The red line represents the true house prices, whereas the blue line represents the predicted house prices. There is a significant overlap between the two lines which indicates that XGBoost provides quite a good fit to house prices and has utility in predicting property prices.



The most important variables XGBoost determines to predict house prices are plotted below.



Referring to the plot above, XGBoost concluded that **Floor Area** and **CBD Distance** to be the highest contributing variables to predicting house prices. This finding is expected and could be elucidated with common sense and a causal explanation: **Floor Area** is representative of the overall size of the household, while **CBD Distance** is highly related to the remoteness and location of the property, which are both very important in determining the demand and price of a property.

After the model is trained, the XGBoost model is used to predict the test set to assess its performance. The model had a Mean Squared Error (MSE) of **37,858**, Mean Absolute Error (MAE) of **120.69** and Root Mean Squared Error (RMSE) of **194.57.** The test RMSE is sligtly higher than the training RMSE which is expected. However, the testing RMSE is not significantly higher than the training RMSE which suggests that the model is not overfitted and can accurately generalise and predict property prices.

3.5.1 Machine Learning Model Performance - Random Forest

Random Forest training produced a training RMSE of **$194.72** thousand. Below is a graph plotting the decrease in error of each training iteration of Random Forest. Figure 9 demonstrates how absolute error converges to a value slightly below 40,000.

Figure 9: Random Forest Training

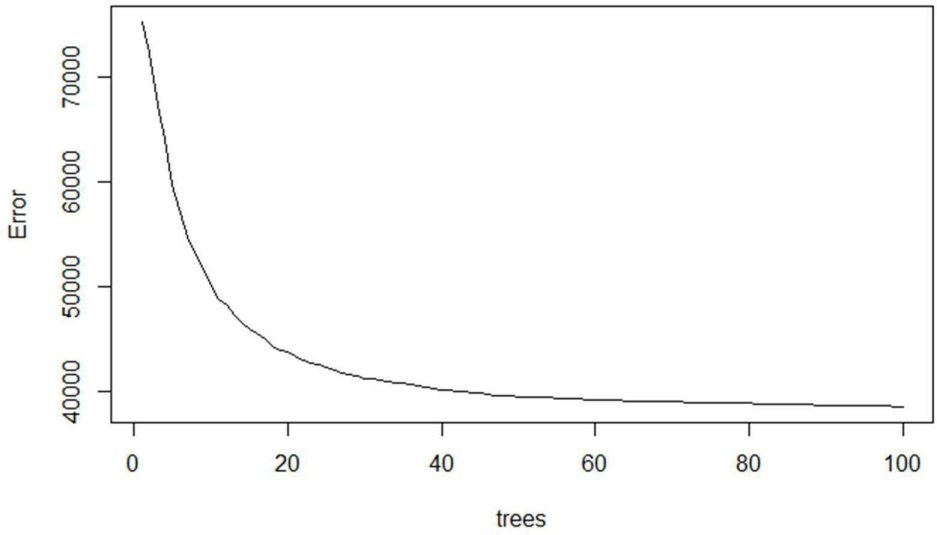
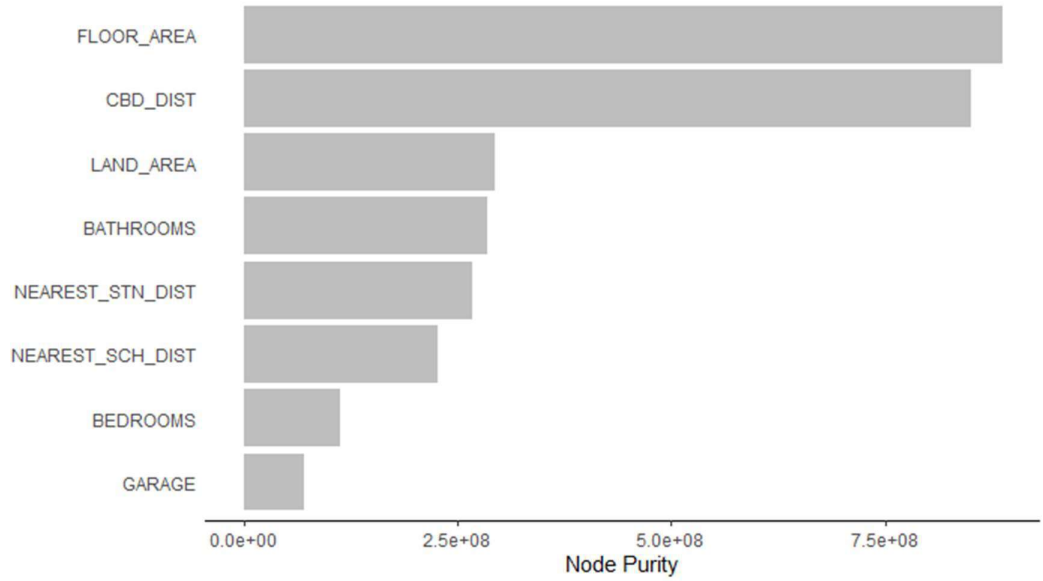


Figure 10: Random Forest Most Important Variables



The plot above shows the most important variables Random Forest determines to predict house prices. The highest ranked variables in Random Forest are **Floor Area** and **CBD Distance.** This result is highly similar to the outputs of XGBoost.

The testing result of Random Forest on a testing set is as follows. Random Forest had a Mean Squared Error (MSE) of **40,473.66**, Mean Absolute Error (MAE) of **124.46** and Root-Mean-Squared-Error (RMSE) of **201.18.**

4. Discussion

To summarise the findings, it is fair to say that the fitting was successful. Using two different methods, the most prominent variables in predicting house prices were identified to be **Floor area** and **CBD distance**. Random Forest algorithm produced a model with a test RMSE of **$201,180AUD**, while XGBoost resulted in **$194,570AUD**. It is difficult to choose a better performing model solely based on the test RMSE as the difference is not statistically significant.

However, it is important to note that XGBoost trains significantly faster compared to Random Forest. This is due to the implementation of different underlying algorithm between Random Forest and XGBoost. XGboost's boosting algorithm adjusts its trees in an efficient manner when iteration progresses meaning that it takes less training time in total (Gupta, 2021). On the other hand, Random Forest's bagging algorithm generates all its decision trees at the same time which costs extra time and efficiency compared to XGBoost (Gupta, 2021). Therefore, in terms of training time, XGBoost significantly outperformed Random Forest.

Nevertheless, comparing the training plot of Random Forest and XGBoost, XGBoost seems to have a faster decrease in error in a smaller amount of iterations. XGBoost also has a smaller RMSE after 100 iterations of training compared to Random Forest. Like aforementioned, this is insufficient evidence to conclude that Random Forest is inferior to XGBoost as the difference is not statistically significant.

Some of the properties of the dataset is sold in the 1980s and the recorded sale price is outdated in the modern times as inflation and economic growth would have affected the price. This analysis suffers the limitation that it included the outdated observations and assumed that inflation is insignificant. A future work for this analysis is to account for inflation in prices to develop a better model.

In this analysis, **Nearest School Rank** was excluded from the model. However, it is known that **Nearest School Rank** plays a significant role in determining house prices as home buyers frequently look at school catchment areas as they desire their children to attend a high-ranking school. Inclusion of this variable could have significantly improved the model's prediction performance and could be an enhancement to this analysis.

5. Conclusions

In this analysis, dimensionality reduction techniques such as principal component analysis and factor analysis was implemented. These techniques found out very interesting characteristics about the dataset and uncovered hidden correlations between property characteristics.

Two machine learning models namely XGBoost and Random Forest were developed to predict house prices according to a property's characteristic such as the number of bedrooms, bathrooms, land area and etc. Both XGBoost and Random Forest - demonstrated satisfactory precision in this regression analysis and it is concluded that both models are suitable and not one model outperformed the other significantly. This conclusion is derived from the fact that both models have very similar MSE, MAE and RMSE. The development of predictive models is an important and complicated task. Building a robust and accurate model that can predict house prices could benefit everyone, from individuals to institutes.

6. References

Beattie, J. R., & Esmonde-White, F. (2021). Exploration of principal component analysis: Deriving principal component analysis visually using spectra. Applied Spectroscopy, 75(4), 361-375. doi:https://doi.org/10.1177/0003702820987847

Breiman, L. (2001). Random Forest. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2021). *A Gentle introduction to Ensemble Learning Algorithms.* Machine Learning Mastery. https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi:10.1145/2939672.2939785

Goretzko, D. (2022). Factor retention in exploratory factor analysis with missing data. Educational and Psychological Measurement, 82(3), 444-464. doi:https://doi.org/10.1177/00131644211022031

Gupta, A. (2021). XGBoost versus Random Forest: Geek Culture https://medium.com/geekculture/xgboost-versus-random-forest 898e42870f30#:~:text=One%20of%20the%20most%20important,hyperparameters%20to%20optimize%20the%20model.

Mantas, C. J., Castellano, J. G., Moral-García, S., & Abellán, J. (2019). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing, 23*(21), 10739-10754. https://doi.org/10.1007/s00500-018-3628-5

Zainal, M. (2009). Perth House Prices [Data set]. Kaggle. https://www.kaggle.com/datasets/syuzai/perth-house- prices

7. Appendices

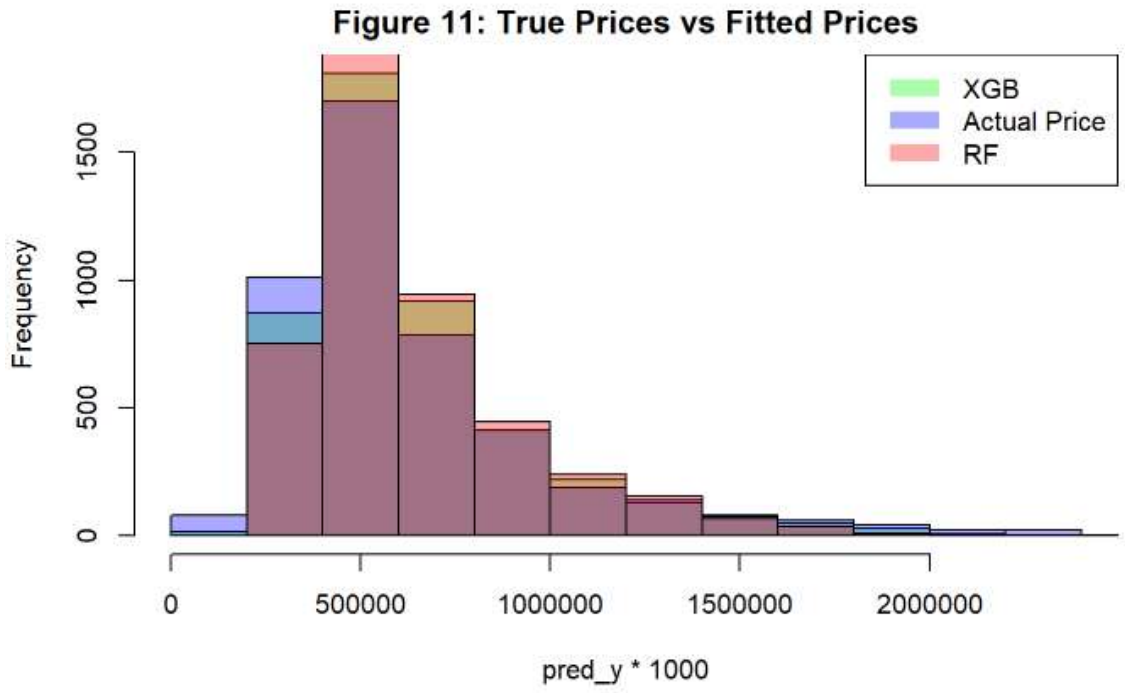


Fig 11: Comparison of different predictions by different machine learning model