The Impact of COVID-19 on Perth House Price:
A Machine Learning Perspective

Yue Hu [1, *]

[1]*Faculty of Mathematics, University of Waterloo, Waterloo, ON N2L 3H2 Canada*
**Email: y343hu@uwaterloo.ca*

ABSTRACT

After analyzing suburb information, school-area concentration, subway stations distribution, and 20 other parameters, a machine learning project using Catboost regression is utilized to predict house prices in Perth. To improve the prediction accuracy, a few innovative variables are created, like "Average_area" which indicates the average floor area per bedroom. Simultaneously, several regression algorithms were selected to apply to the model, and Catboost is selected based on a newly designed evaluation method in this research. With the project, using an alternative approach of the Difference-in-Difference (DID) method allows the product to keep the control group as the pre-COVID terms to analyze the impact made by the pandemic to house price market in Perth. The results show that there is no evidence indicating the house prices have been impacted by the pandemic. However, it could be noticed that this result only works for a city like Perth which does not have many cases and a statement of general lockdown does not have an impact on the real estate market. A few evaluation methods are utilized to make the judgment and further cater industry's needs and better filter out the models suitable for business. These methods are expected to apply in the industry for trade-in evaluation and future market forecasting.

Keywords: *Catboost, Evaluation Method, Machine Learning, Risk Analysis*

1. INTRODUCTION

Living through the "Great Lockdown" induced by the COVID-19 pandemic, although it is not over yet, the impacts it has on all aspects have attracted the attention of media and investors. The financial markets have seen dramatic movement on an unprecedented scale [1, 2]. The housing market in the US experienced the initial shutdowns on transactions and a surging housing market afterward, according to research in Redfin Data Center [3]. As one of the most stable fields and markets, house price in Perth, the largest city in West Australia is a good start to research.

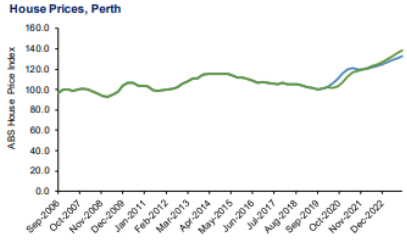


Figure 1 Predicted house price change due to COVID-19 from KPMG.

Notes: Impact of COVID on Australia's residential property market 2021. (Price index vs. timeline)

The new pandemic had a great influence on the global real estate market, started researching a middle-size city in a specific field could give a more typical idea to predict future impact for the next pandemic or next wave COVID-19 cases. Australia's population is estimated to be around 4 percent smaller (1.1 million fewer people) by 30 June 2031 than it would have been in the absence of COVID-19 [4]. This reduction in population growth will decrease the general real estate demand on a long-term basis. Simultaneously, during the "great lockdown period", people tend to be less interested in outdoor activity and relocate as a demand drop in the short term as well [5]. Apart from that, while the general supply won't change much, the transformation to online university and popular worldwide work from home since the global pandemic greatly decrease the rental demand in the area. Thus, a certain amount of houses is now empty and its owner would be more than happy to sell them [6]. As a consequence of the above, a steady decrease in Perth house prices was predicted and this research is settled to dig deeper into the impact of COVID-19 on the real estate market in Perth. There was

research regarding the general real estate market in Australia during COVID-19. Unfortunately, it is premature to make strong predictions about specific impacts from that research's perspective. Since COVID-19 is, in comparison, a newer topic that does not have much-correlated research focused on its impact on the real estate market, especially for Australia. The only research by Nygaard and Parkinson was done at the beginning stage of the pandemic while they are not comfortable giving strong evidence to have a statement [7]. Apart from that, their research was using traditional statistic techniques while this report is trying to find the impact using machine learning models.

However, with an idea come from the Difference-in-Difference method that we may remove all the data after COVID-19 breakout and have a control group data set. Applying this dataset into the machine learning project to predict the price for the house sold in the pandemic term. By comparing the actual price sold and predicted price sold, we may see two curves for the same period with COVID-19 and without COVID-19 hypothetically. To compare the trend for each of them and between them, we could analyze the pandemic impact on the Perth house market. Furthermore, multiple waves of COVID-19 cases could be compared to the trend above, and a curve presenting new COVID cases could be helpful. Finally, a relatively strong statement could be made which is COVID-19 didn't have an obvious impact on the housing market in Perth. After that, by keeping the same data amounts, a 10-year-trend of future house prices in Perth could be predicted. Applying a machine learning model to both all-term data and pre-COVID data would be efficient in describing a long-term trend.

Machine learning project is not common in the real estate market since it is experienced-based and a few risk analysis approach is discussed to have a chance to make machine learning more useful for the real estate market. KPMG has done similar research on the general impact of COVID-19 on Australia's residential property market and concentrating on a more specific area with a more detailed dataset would be beneficial for a more concrete conclusion. In addition, several innovative variables are created in this paper's data pre-processing step to improve the model like school ranks near the house and average floor area per bedroom. Besides, a few evaluations designed from the risk analysis method were set to select the best models according to a business application is introduced after the project, which significantly improves the efficiency of the machine learning process. At last, some possible applications for machine learning models on experienced-based market like the real-estate market was given to list a path making this kind of traditional market automotive and technologic. While developing progress, machine learning models could greatly help practitioners with data support in forecasting future trends. In future approaches, machine learning models could be developed to provide evenly accurate judgment with human beings or even better.

The remainder of this paper is as follows. Section 2 provides a general description of the data and variables about the research. Section 3 presents the methodology in the machine learning project. Section 4 discusses the empirical results. Section 5 concludes.

2. DATA AND VARIABLES

2.1. Data description

The data used in the research was scraped from house.speakingsame.com including data from 322 Perth suburbs, resulting in an average of about 100 rows per suburb. This data was updated to include all real-time house price data in Perth as of December 2020 to Kaggle 9 months ago by Muhammad S. Zainal. The main reason this dataset was selected was that the parameters chosen were innovative and crucial to the predictive analysis. Other than the common parameters in predicting house prices field, the uploader provides some new parameters as well, such as School Ranking data. This data was obtained from bettereducation.com.au to show each house's nearest school's rank defined by 'ATAR-applicable'. In the Australian secondary school education system, ATAR is a scoring system used to assess a student's cumulative academic results and is used for entry into Australian universities. And another fact needs to be noted is that under the "NEAREST_SCH_RANK" column, there are some missing rows as some schools are unranked according to these criteria by better education.

Another parameter that needs to be noted is called "Suburb" which is a special geography subregion in Australia that normally includes 3-4 blocks. Thus, this categorical factor provides a small and good partition for the house's location which comes to be a truly strong factor.

Table 1. The determinants of house price in Perth

Determinants	Explanation
Average_Area	Average floor area for each bedroom
Garden_Area	Garden area for the house
Date_Sold	The date of the selling record
Age	Age of the house
Nearest_school_rank	The rank of the nearest school on ATAR
Nearest_school_dist	Distance to the nearest 'ATAR-applicable' school
Nearest_stn	Name of the nearest subway station
Nearest_stn_dist	Distance to the nearest subway station
Floor_area	Floor area for the house

Garage	Garage number for the house
Bathrooms	Bedroom amount
Suburb	Suburb the house located in

2.2. Data Pre-processing

The original data was ambiguous, and some re-transform was needed for the data and future research, and this is one of our most important creative parts in the data analysis. Some preliminary analysis conducted showed a significant correlation between each of these columns and the response variable (i.e., price). A few new and original way was tried and acted in the research – The bathroom number factor was removed since we did fine for the majority of the houses, bathroom amounts meet the requirement and seem not to have much impact. "CBD_Dist" which describes the distance from the house to the Perth CBD area was also removed. This was a pretty adventurous attempt since this factor had always seemed essential in the house price prediction model, however, it was dropped since suburbs as a careful location parameter contained similar information, and compared to condominiums, houses related less and did not need fully accurate data on this parameter. For a similar reason, "POSTCODE" and "ADDRESS" parameters were also removed.

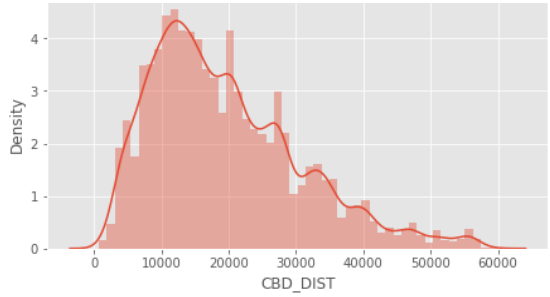
Simultaneously, a few new parameters were defined to help improve the model. A new parameter called "Average_area" was defined while it was rarely seen in another research. This parameter was calculated by Floor area to divide the number of bedrooms to give an average area per bedroom. This is a creative idea as the main inspiration is most people are looking for houses with bedroom amounts that meet their needs. While in this process, richer people tend to be more willing to pay more to have a larger house. On the other hand, higher-end houses are always designed to be larger with the same bedrooms which may not look obvious in a simple "Floor_area" parameter. Therefore, this new and original parameter was added to the model and worked pretty fine later on.

Table 2. Pre and Post data processing

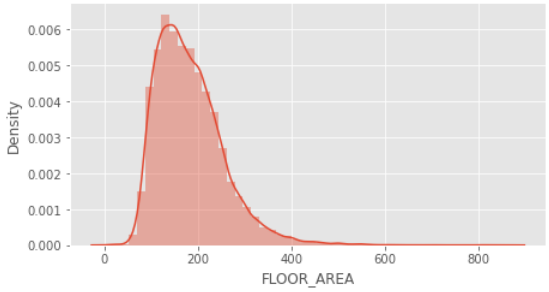
Before pre-processing			After pre-processing		
Variable	Missing value	percent	Variable	Missing value	percent
EAREST_SCH_RANK	9517	33.11%	SUBURB	0	0%
BUILD_YEAR	2729	9.49%	PRICE	0	0%
GARAGE	1787	6.22%	BATHROOMS	0	0%
NEAREST_STN	0	0%	GARAGE	0	0%
PRICE	0	0%	FLOOR_AREA	0	0%
BEDROOMS	0	0%	CBD_DIST	0	0%
BATHROOMS	0	0%	NEAREST_STN	0	0%
LAND_AREA	0	0%	NEAREST_STN_DIST	0	0%
FLOOR_AREA	0	0%	DATE_SOLD	0	0%
SUBURB	0	0%	LATITUDE	0	0%
DATE_SOLD	0	0%	LONGITUDE	0	0%
ADDRESS	0	0%	NEAREST_SCH	0	0%
POSTCODE	0	0%	NEAREST_SCH_DIST	0	0%
LATITUDE	0	0%	NEAREST_SCH_RANK	0	0%
LONGITUDE	0	0%	AGE	0	0%
NEAREST_SCH	0	0%	SOLD_TIME	0	0%
NEAREST_SCH_DIST	0	0%	GARDEN_AREA	0	0%
NEAREST_STN_DIST	0	0%	AVERAGE_AREA	0	0%
CBD_DIST	0	0%			

2.3. Descriptive statistics

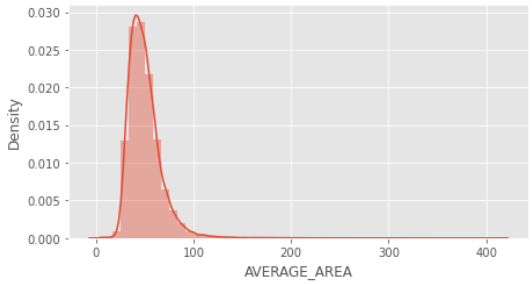
To better learn the data distribution for numeric variables, a few histograms were produced to show its distribution which might be helpful for further manual parameter adjustment. It is essential to look into the histogram graph for variable CBD_DIST which provides the idea of a flat distribution. This gives an idea that the houses in Perth locate averagely in the urban and the rural. On top of that, a comparison of Price and Floor_area shows a similar trend, while the price graph has a fatter tail in a high price which indicates that average price by area does have a larger range in large houses. This also keeps consistency with common sense. Besides, a less deviated average_area graph compared to floor_area gives the idea that most houses in Perth are designed for the middle class and the average area per bedroom is consistent for the majority of houses.



(a) CBD_DIST



(b) FLOOR_AREA



(c) AVERAGE_AREA

Figure 2 Histogram graph for numeric variables.

On top of that, scatter plots are essential in descriptive statistics steps which could provide information about how related a variable is to the label – house price in this project, From the comparison of two, scatter plots below, it is obvious floor area is much more related to the house price comparing to nearest school distance. A fairly strong positive relation could be summarized from Figure 3 while a few outliers exist. A certain number of outliers on the top of the graph show that some extremely high-value houses may only have a small floor area, those cases may happen in the central heart area of CBD. While much fewer outliers may happen in the right-down corner of the plot indicate cases for a low trading history of large houses which may happen a long time ago or the house is located far away from the urban area.

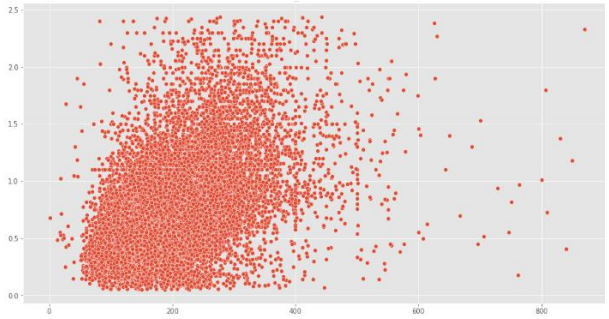


Figure 3 Scatter plot of Floor_area vs price.

Although less clue could be revealed from Figure 4 since most houses are located near a school, it is still obvious that houses won't be valued high when far away from a school (5 km), this may be the consequence of the importance of education, or education is developed highly in Perth that houses far away from school are all isolated in the city.

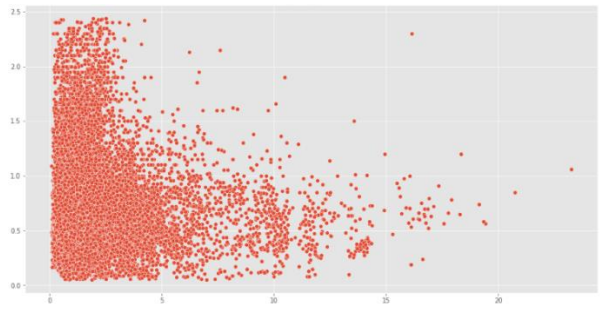


Figure 4 Scatter plot of Nearest_sch_dist vs price.

3. METHODOLOGY IN MACHINE LEARNING

3.1. Catboost data analysis and regression

Catboost was introduced and broadly used recently as a brilliant technique to analyze categorical data, and it allows them to apply to the machine learning model later on together with the rest of numeric data.

Afterward, several algorithms were tried in the research process, and we have fitting scores to be the initial evaluation step to choose from the numerous algorithms. While in this step, a new algorithm outstands from all others- Catboost algorithm which shares the same package with the categorical parameter analysis. As a

new regression algorithm, Catboost provides extremely excellent performance when modeling regression with categorical parameters and has been decided to practice at the end.

3.2. Final model evaluation

To compare the model and optimize the future research, a few traditional statistical analysis methods were used, and one new evaluation method was introduced to better meet the business requirement.

As one of the most common statistical methods, RMSE (the root of mean square error) was settled first to evaluate the whole image of the prediction. After that, another statistical number was mentioned in the research model evaluation step: Maximum of the difference, it is pretty important here that not only provide information for a predicted largest error but also helps to check any evidence for wrong raw data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}} \quad (1)$$

$$(N = sample\ size, x_i = sample\ value, \hat{x}_i = predicted\ value)$$

Simultaneously, a difference rate was introduced for each data and 95% percentile of difference rate, 90% percentile of difference rate, 80% percentile of difference rate were calculated to help the machine learning model have a greater business field application. As a dealer or even an online second-hand house transaction website, this model could be used to evaluate an approximate price for customers. Furthermore, for actual trade-in, it could also help dealers as a reference. Thus, the percentile of difference rate comes to be pretty important that dealers could know that, for example, for the final model now, it has an 80% confidence that its prediction of the house price will be 25% or lower deviation of the actual house price sold. This may help the dealer to decrease their predicted house price by 25% to have an 80% guaranteed riskless train-in value for any house located in Perth.

$$difference\ rate = \frac{\hat{x}_i - x_i}{x_i} \quad (2)$$

To better help the business application, a new parameter was also created to cater to the needs. It is called “Within 1/30 rate”, for any prediction done in the model, if the prediction price falls between 29/30 to 31/30 of the actual house price, i.e, within 1/30 of the deviation, then it is called a “somehow correct” prediction and almost 15% of prediction in the current model do give such “correct” prediction and it may keep increasing. This allows online house transaction website to provide customers and suggestion value for the house, and 15% of its customer would find its prediction concise and may further recommend the website to their friends.

4. EMPIRICAL RESULTS AND DISCUSSION

4.1. Two datasets applied in the model

To further focus on the research of the impact of COVID-19 on Perth house price, an experimental control group is needed based on the original 32200 house history data. After the data clean-up step stated above, all data that sold after February of 2020 was removed to have a control group compared with the experimental group. To keep the control research consistent in most of the experimental condition, around 3000 rows of data was randomly removed from the experimental group to keep both of the group in a similar amount of data.

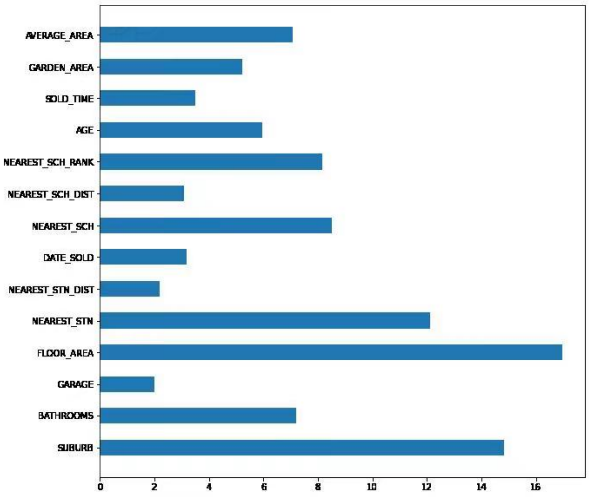


Figure 5 Categorical parameter analysis from Catboost package.

Further than that, a learn & test group division need to be done before all the machine learning step for both of the groups. 70% of data are classified as train datasets to provide a reference for the machine learning method for algorithms or regression. For the rest 30% of all data, they are set up to be the test data to give an intuitive appearance for the prediction model, which could also allow the new evaluation method to test if COVID-19 does have any impact on the house price in Perth.

Catboost was used to provide a deeper analysis of the categorical variables and help algorithms, later on, to handle the categorical variables and numeric variables together. The concrete method was label encoding which transformed the categorical parameters to be numeric and made the model understandable.

After the Catboost parameter analysis, a few popular models were attempted briefly and selected through fitting score built-in intuitively. Light GBM stood out from all well-known models like Xgboost, Random Forest, Bayesian Ridge, Lasso, Linear GAM, and Elastic Net. This model was primarily set up as the optimized model, however, a new model which was not usually seen in regression surprised us – Catboost, the regression

algorithm programmed with the label encoding process. This corresponding algorithm was performed out of our imagination and provided a much better effect on the regression process with a higher fitting score.

4.2. Machine learning results

All regression algorithms were prompted in the notebook, while some of them had a fitting score over 80% were selected to have an output table to illustrate their prediction value for the test data. Then, method evaluation methods are indispensable to judging the correct algorithms. A few traditional statistic techniques are stated above, and two new defined methods took part to evaluate each model.

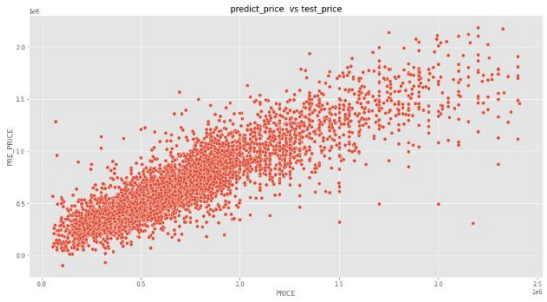


Figure 6 Scatter plots of predict price vs real price.

4.3. Compare future house price trends with/without COVID-19

After the model selection step for both data groups, two optimized prediction data were produced to provide a general idea of the prediction. Using both fitting scores and the evaluation method used before, two groups provided similar prediction results for 2020 cases. In this case, we could broadly outline that COVID-19 hasn't applied an obvious impact on the house price in Perth.

Table 3. Model Evaluation for data Include & Exclude pandemic term.

	RMSE	95% percentile	90% percentile	80% percentile	Within 1/30 rate
Include	169333.7	0.5390	0.3783	0.2596	0.1458
Exclude	152532.4	0.5054	0.3650	0.2572	0.1558

While comparing to an inconsistent conclusion from KPMG's report regarding the general real estate market in Australia, it can be concluded that as an isolated city located in Western Australia, which hasn't been influenced much by the pandemic, Perth did not receive a much impact for the house markets. First of all, KPMG did the research by comparing trade history in 2020 to predicted market which could not identify if the error comes from an impact from the pandemic or their prediction models. Second, KPMG researched for the whole of Australia which has over 100 thousand positive cases up to now [8]. However, as an isolated western city, Perth hasn't suffered much from it and as the consequence, maybe impacted smaller compared to the whole country, Third, KPMG's research was on the whole real estate market that houses solely do reveal a steadier market that most international investors would rather prefer to condominiums markets and could have an impact due to the "great lockdown". At last, but not least, although KPMG made a decrease impact comment on the overall market, its separate indication for Perth which cited at the beginning of the paper indicates they share a similar opinion that the Perth house market hasn't got much impact from the global pandemic. This may be a consequence of special city properties, and a few suspects could be provided. Perth is a city near the sea, which makes most of the houses in the city have a fixed value come with the great view of the Indian Ocean. Apart from that, the outbreak is not that serious in the area – The total positive case in Western Australia is only around one thousand during these two years, while only 9 of them are infected interstate. Perth itself is not an international city may be another main reason.

4.4. VaR and ES application for machine learning model

To further provide evidence to research the business application of the machine learning model, two typical risk analysis techniques are introduced here to the model evaluation step. The first one, which is also the most common definition in the risk analysis field, is called VaR (Value at Risk) which describes at a certain risk, (percentage) the predicted loss will not be exceeded.

For dealers, it's normal that they may provide a lower-than-market trade-in value to their sellers for a quicker and more convenient transaction. Thus, taking an example of a machine learning model with an 80% percentile difference rate of 0.15 (which method we introduced before in 2.4) does give an option for the dealers to provide a 15% lower trade-in value offer based on the model prediction price. And therefore, by the

definition of 80% percentile difference rate, the transaction would have an 80% possibility that won't make him a loss. Similarly, we could take the definition of VaR above to apply that this model does provide the dealer a VaR0.8 of 15% of the total house price.

VaR also allows ES (Expected Shortfalls) to be applied here in a similar way. ES is also a typical risk analysis parameter, while more practical in business applications. Expected shortfalls are the expected loss given that the loss falls in the worst 20% part of the loss distribution. In the traditional statistic and risk analysis field, an accurate integral is needed to calculate the ES value, and it also works here.

Simultaneously, since house prices vary and the percentage is a much better way to calculate to evaluate all the methods rather than exact difference rate, we could just simply calculate the average of the last 20% percentile difference rate to have a similar ES as we could have in the tradition risk analysis field. This ES would also provide an expected loss under the worst 20% part of the loss distribution, on average.

Both the VaR and ES are not only a good model evaluation method but also two very good side notes provided to the objective customers – real estate dealers. These two risk analysis data may be much more familiar for them since they are both pretty common in the finance field, while at the same time giving an idea of how the model appears. Also, if there is a trade-off of "correct rate" and these two risk parameters happen, dealers could also have a reference to judge their most comfortable model based on their risk-taken ability. For example, that, although the two models we produced share similar statistic errors, the one in Figure 6 is preferable for business prediction.

Table 4. ES application for Catboost Base & Adjusted parameter

	RMSE	90% percentile difference rate	80% percentile difference rate	90% Expected Shortfall	80% Expected Shortfall
base	169333.7	0.5390	0.3783	0.7065	0.5564
Adjusted	152532.4	0.5054	0.3650	0.6038	0.4988

4.5. Possible Model Application on the experienced-based business market

As is known to all, the real estate market is mainly experienced-based that dealers could have recent history sale records in the neighborhood which could be a much more precise prediction of the house price. Thus, machine learning algorithms always seem not to be a popular way among dealers. However, they could provide an alternative way for the dealers to evaluate real estates which are out of the ordinary or have much difference from the rest. There was research from 2018 that gave an idea that they aim at developing a machine learning application that identifies opportunities in the real estate market in real-time, i.e., houses that are listed with a price substantially below the market price which is a pretty good example [9].

Apart from that, compared to the traditional linear regression method, the machine learning algorithm provides a more comprehensive prediction on future trends [10]. By inputting recent sales records along with the historical records together, a preference for future trend prediction could be obtained. It could be used to advise on specific houses, for example, this house would be great to sell during July or August since it is a medium-size house near a famous school with many bedrooms which is comfortable for a typical family and this kind of house is easy to sell on a fair price in summer. On the other way, a numeric analysis for future trends based on data could also be a strong help to risk aversion.

5. CONCLUSION

In research for COVID-19 impact on house price in Perth, a machine learning project using the Catboost project is established. Within the same model, datasets with and without pandemic period data produced a similar result which indicates that the pandemic didn't implement a considerable impact on the Perth house market. While simultaneously, the similar output from both of the datasets also reveals that the pandemic is hopefully not going to have a long-term influence either, which is questionable from KPMG's traditional statistic method research.

A few innovations are designed in variables for the machine learning project, like school rank, removed CBD_dist, and newly created variable for average floor area per bedroom. In addition, new evaluation methods designed for similar machine learning models are introduced to better choose the correct algorithm and help with parameter adjustment. After the evaluation, it can be concluded that no obvious evidence shown to prove COVID-19 have an impact on house price in Perth. Finally, a few risks analysis approach in VaR and ES perspective were discussed to provide machine learning models prospect in the experienced-based market like real estate. This could be applied in the industry for trade-in evaluation and future market forecasting.

House price prediction is a very bold but solid topic, and this research paper also has a few deficiencies that could be improved in the future. First of all, the fitting score and the chosen model were unsatisfactory, and they could be improved later on to provide a better prediction. Although the model at this point is sufficient to support the results that COVID-19 haven't applied a much impact on the Perth house market and it is eligible to test the new evaluation method, an advanced model may provide more details when comparing Perth's house price with others and give more specific advice to improve the evaluation method. Other than that, a combination of machine learning model and evaluation method could be improved to further adapt to experience-based markets such as real estate.

REFERENCES

[1] P. K., Mishra, & S. K., Mishra, Corona Pandemic and Stock Market Behaviour: Empirical Insights from Selected Asian Countries. Millennial Asia, 2020, vol 11, pp. 341–365. https://doi.org/10.1177/0976399620952354

[2] A. Haldar &N. Sethi. THE NEWS EFFECT OF COVID-19 ON GLOBAL FINANCIAL MARKET VOLATILITY. Buletin Ekonomi Moneter Dan Perbankan, 2021, vol. 24, pp. 33-58. https://doi.org/10.21098/bemp.v24i0.1464

[3] B. Wang, How Does COVID-19 Affect House Prices? A Cross-City Analysis. Journal of Risk and Financial Management. 2021. https://doi.org/10.3390/jrfm14020047

[4] KPMG economics. (n.d.). Impact of COVID on Australia S residential property market. 2021.

[5] The effect of COVID-19 on the Perth Property Market. CR Settlements., 2021.

[6] Department of the Treasury. 2020 Population Statement. Centre for Population. 2020.

[7] C.A., Nygaard and S., Parkinson, Analysing the impact of COVID-19 on urban transitions and urban-regional dynamics in Australia. The Australian Journal of Agricultural and Resource Economics, 2021, vol. 65, pp. 878-899. https://doi.org/10.1111/1467-8489.12449

[8] Australian Government Department of Health. Coronavirus (COVID-19) case numbers and statistics. Australian Government Department of Health. 2021.

[9] A. Baldominos, I. Blanco, AJ. Moreno, R. Iturrarte, Ó. Bernárdez, C. Afonso. Identifying Real Estate Opportunities Using Machine Learning. Applied Sciences. 2018, vol. 8, pp. 2321. https://doi.org/10.3390/app8112321

[10] I. Baturynska, K. Martinsen, Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. Journal of Intelligent Manufacturing, 2021, vol. 32, pp. 179–200. https://doi.org/10.1007/s10845-020-01567-0