Developing a Machine Learning-Based Application to Predict Future Prices of HDB Resale Flats in Singapore

Joven Pua Zai Xiong (A0201693J), Lee Jing Han (A0206021A),
Lee Yu Xian Grant (A0183367W), Low Ee Ter (A0183691W),
Wong Yoke Ling, Noelle (A0205266J), Yang Zichang (A0206039L)
National University of Singapore

e0415502@u.nus.edu, e0425944@u.nus.edu, e0310162@u.nus.edu,
e0310486@u.nus.edu, e0425189@u.nus.edu, e0425962@u.nus.edu

Motivation

Housing Development Board (HDB) flats account for over 80 percent of Singapore's residents. As a potential buyer of a resale flat, researching flat prices can be a tedious task. After all, in land-scarce Singapore, many factors directly or indirectly influence the price.

An equal (if not greater) amount of research effort is required for a potential seller of a flat, so as to understand the reasonable range of prices to set for a listing. In fact, most buyers and sellers tend to prefer engaging a property agent in order to handle negotiations with regards to the pricing of the resale flats.

As for property agencies, their research consists of studying the statistical relationships between the aforementioned factors and the flat prices, as do real estate analysts and economists. Our proposed application attempts to take a machine learning approach to this task.

We aim to develop an application built upon a machine learning model which is trained to perform such price predictions. The significance of such a model is that it tends away from human bias and more towards statistical integrity. By employing regression techniques, we are likely to achieve a price prediction that more reliably and more objectively represents the market price.

With such an application, users can have convenient and reliable access to the current "market price" as well as the "future market price" of a particular resale flat configuration. Buyers, sellers as well as property agencies can then use this recommended price as a basis for negotiation of the sale price for a given resale flat. This puts all stakeholders on the same page and encourages honesty in the flat resale business.

Description

Our application is targeted at three main stakeholders in the flat resale business:

1. Potential buyers
2. Potential sellers
3. Property agencies

The hallmark feature of our application is to provide price predictions when given a certain configuration of a HDB flat. This will prove to be useful for potential buyers and sellers of HDB resale flats.

In addition, our application will also provide a graphical visualisation of the predicted price in the next 5 years. While this feature can be useful for buyers and sellers who are planning their flat purchase/sale, it is also a useful tool for property agencies to plan their business focus.

In its current stage, our machine learning model has been trained on historical flat resale data in the Jurong area. Hence, the application will deliver the most accurate price predictions for resale flats in the Jurong area. The reason for this choice is due to the Singapore government's focus on redeveloping the Jurong town into a second central business district.[1] This suggests that the resale flats in the Jurong town are likely to see a surge in interest.

In the near future, we intend to train our model on data from other towns in Singapore as well. Once we have done so, this enables us to introduce a few other features that can further enhance the functionality of our application.

Firstly, we can conduct a comparison analysis of future prices of similar-configuration flats in different towns. For example, real estate industry analysts could use this tool to compare the price movements between an up-and-coming town like Tengah and an older estate like Toa Payoh.

Secondly, we can conduct an ablation study to better understand the relationships between different flat features (such as storey range, geographical location) and their prices. By studying which factors have a greater effect on flat prices, we can generate a housing report that could be used as corroboration for consumer sentiment surveys.

The result of this study can also be used to help new potential resale flat buyers to set a more definite purchasing price for a new flat. This is important as HDB performs price validation after a deal has been struck on a house and any additional amount that is quoted on the deal cannot be paid using their Central Provident Fund (CPF) accounts. Therefore, the additional insights to housing prices provided by the study can help new buyers to get a quote that is similar to the estimated validation price. This in turn minimises out-of-pocket cost for these buyers.

Research methodology

Our research process consists of the following phases:

Literature review

We conducted a literature review in order to gain a greater understanding of:

1. Machine learning approaches for price prediction
2. The context of the local housing market
3. Market research of existing price prediction applications

In particular, we paid attention to the machine learning models that were used by other researchers in price prediction problems.

Data collection

The dataset we have used is sourced from the public Data.gov.sg database. Given that the source is a government agency, we are inclined to trust the integrity of the data. The dataset is organised into the following fields: *month of transaction, town, flat type, block number, street name, storey range, floor area, lease commencement date, remaining lease, resale price*.

We noted that these fields are closely aligned to the factors that buyers, sellers as well as property agencies often consider when pricing a resale

[1] Residential hotspots: Districts shaping up to be interesting investment bets. (April 1, 2021). *The Business Times*.

flat. We also noted that these fields are mostly well-defined, as opposed to more subjective factors (mentioned in our first assumption).

Assumptions

In order to build our current model, we recognise the need to set certain limitations with regards to the data that is available for us. As such, as far as our current research is concerned, we have taken on the following assumptions:

1. Certain often-subjective factors (e.g. *fengshui*, window view, interior design) are not used to train the model.
2. Patterns deriving from changes in time-based market/economic conditions have been factored into the price, given that our dataset consists of time series data.
3. Specific block numbers are removed from our dataset so that blocks in the same street are not differentiated. This is done to reduce overfitting of the data by the model.
4. The distance thresholds corresponding to the new columns added (see "Data preprocessing" section) are reasonable.

Literature review

Price prediction techniques

There is a modest amount of research into machine learning-based predictions of housing prices. Although most of the research has been done with reference to other cities, we can consider some of the housing features that other researchers have considered and apply them to our own research.

Gao et al. (2019)[2] took a location-based approach that considers not only geographical but also non-geographical features. Geographical features include facilities like proximity to transportation hubs and school districts. Gao et al. noted that support vector regression (SVR) and random forest (RF) were the best performing models for their research.

Truong et al. (2020)[3] suggested numerous data preprocessing steps as well as performance evaluation via root-mean-squared logarithmic error. Besides RF, Truong et al. suggested Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Hybrid Regression as well as Stacked Generalization Regression as machine learning models to consider.

Singapore's resale flat market

According to Lee (2019)[4], key factors that affect the price of a HDB flat are:

1. Accessibility (time taken to walk to the nearest MRT station)
2. Proximity to facilities such as schools or malls
3. Flat type (storey, remaining lease, floor area, flat model)
4. Estate demographic (i.e. proportion of working adults, young people, old people)

Redbrick Mortgage Advisory (2018)[5] further suggests the following factors:

1. Condition (e.g. cracks on walls)
2. Renovation and interior design
3. Reputation

According to De Silva (2017)[6], the following factors also come into play:

1. Infrastructure plans (e.g. the originally-planned High Speed Rail terminal in Jurong East as well as the development plans in the Jurong Lake District have contributed to a rise in resale flat prices in the area)
2. View/orientation (highly subjective)
3. Layout

Existing price prediction applications

There currently exists several HDB resale flat price prediction applications. Their implementations are discussed as follows.

First, HDB's Resale Price Index[7] tracks quarterly price movements of HDB resale transactions across towns, flat types, and models. This is more useful for macro-analysis of the resale flat market.

SRX's X-Value[8] performs machine learning-based Comparable Market Analysis using quantitative (e.g. size, storey) and geospatial data (e.g. proximity to MRT stations) from close to 40 property databases. The data is adjusted to current market conditions and represented in four dimensions (property attributes, time series, location elements, outputs such as rental values). The performance of their model is impressive, with 92.9% of actual 2009-2019 transaction prices falling within 10% of the predicted X-Value price.[9] Note that the price prediction is only available for the current time (i.e. the time of query).

EdgeProp's Edge Fair Value application[10] is similar to SRX's X-Value, but differs in that it collects data directly from real estate agents and professionals as well. Its performance is equally impressive, with 90% of actual 2009-2014 transaction prices falling within 10% of the predicted Edge Fair Value price. However, note that price prediction is also only available for the current time (i.e. the time of query).

Data preparation

Data cleaning

As previously mentioned, for purposes of demonstration, we decided to train our model only on data originating from flat resales in the Jurong area. This was done by simply filtering the dataset using the street_name variable.

Preprocessing techniques

The dataset consisted of both continuous data and categorical data. In order to fit both types of data into our machine learning models, we encoded them as follows:

Columns	Preprocessing method
• month	The month values are given in the format 'YYYY-MM'. This was converted into an integer value representing months from December 1989. For example, '1990-02' is converted to 2, and '2021-03' is converted to 375.

[2] Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, T., & Wu, Z. (2019). Location-centered house price prediction: A multi-task learning approach. *arXiv preprint arXiv:1901.01774*.

[3] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science, 174*, 433-442.

[4] Lee, M. (2019, January 15). *Data-Driven Approach to Understanding HDB Resale Prices in Singapore.* Towards Data Science. Retrieved from https://towardsdatascience.com/data-driven-approach-to-understanding-hdb-resale-prices-in-singapore-31c3beecfd97

[5] Redbrick Mortgage Advisory. (2018, November 7). *4 factors that will affect your flat's resale value.* iProperty.com.sg. https://www.iproperty.com.sg/news/4-factors-that-will-affect-your-flats-resale-value/

[6] De Silva, A. (2017, November 28). *Your Property Agent Says: Eight Factors That Determine A Property's Value.* EdgeProp. https://www.edgeprop.sg/property-news/your-property-agent-says-eight-factors-determine-property%E2%80%99s-value

[7] Housing Development Board. (n.d.). *Resale Statistics.* https://www.hdb.gov.sg/residential/buying-a-flat/resale/getting-started/resale-statistics

[8] SRX. (2015, February 18). *What is X-Value?* https://www.srx.com.sg/ask-home-prof/5522/what-is-x-value

[9] SRX. (n.d.). *X-Value Performance.* https://www.srx.com.sg/XValue-performance

[10] Lin, Z. (2015, August 29). *Psychological pitfall in pricing properties.* https://www.edgeprop.sg/property-news/psychological-pitfall-pricing-properties

• town • flat_type • flat_model • street_name	One-hot encoding was used to relabel the different towns within the data.
• lease_commencement_date • resale_price • floor_area_sqm	These are float values, hence they will be used as-is.
• storey_range	Data in this column is given in the form of "10 TO 12", which cannot be fitted into any model. As such, we took the average of the 2 numbers. The new value is stored in a new column named storey_simplified.
• block_number • remaining_lease	Both columns are removed since remaining _lease is directly correlated with lease_commencement_date. We also assumed that block_number does not directly affect price (i.e. blocks in the same street are not differentiated) in order to reduce overfitting.

After preprocessing the data, we wanted to further improve our model by introducing more geospatial/location-based features that could affect resale flat prices. This was in line with what we had learnt from our literature review.

New columns (mrt_station, bus_stop, hawker_centre, clinic) were added into the dataset. As the column names imply, the boolean values are used to signify whether said facilities are present in proximity to the street on which the resale flat is located. (1 is used to represent the presence of the facility near to the street and 0 is used otherwise.) The metric that we used to determine the presence or the lack of facility within the vicinity is as follows:

- mrt_station (800 metres)
- bus_stop (800 metres)
- hawker_centre (1500 metres)
- clinic (800 metres)

The distances within the parentheses refer to the walking distances from the flats. For example, for an arbitrary flat in Jurong East Avenue 1, if we can find an MRT station within 800 meters, mrt_station is labelled with value 1. This labelling was manually done via retrieving the walking distance calculated by Google Maps. In the future, we can improve on this by making use of Google Maps' Distance Matrix API instead.

Last but not least, we added new columns representing the latitude and longitude of the resale flats, as an explicit representation of the location of the flats.

Model training

Machine learning models

Given that we were interested in a machine learning-based approach, the four models we chose to explore are:

1. Support vector regressor (SVR)
2. Random forest (RF) regressor
3. Neural network (NN)
4. K-nearest neighbours (KNN) regressor

The reasons for using the respective machine learning models are as follows:

Support vector regressor: We considered training using SVR as it is known to work reasonably well with high-dimensional spaces, given that we had performed one-hot encoding earlier. Additionally, SVR is robust against outliers, which may be present in the dataset.

Random forest regressor: RF was to be the basis of comparison for the other machine learning models. We understood that generally the RF regressor does well in such regression tasks and as such decided to consider it as one of the machine learning models. We also noted that **RF regression does not work for extrapolation problems**. However, we decided to still run this model to use as a baseline for comparison of RMSE (root-mean-square error) values.

Neural network: We wanted to use a NN to predict house prices as we thought that NNs would be able to bring about predictions that are accurate since it is able to capture complex features in data and classify non-linear data very well. Initially we had wanted to experiment with CNN (convolutional neural networks). However we later discovered that these networks are usually used in the context of deep learning where images are involved and it is prohibitively time consuming on our machines with little return in the form of model performance.

K-nearest neighbors regressor: Seeing that we have quite a number of features available in the data, we thought of using a KNN regressor since it provides a reasonable and stable estimation of the unknown instances based on the similarity of features of the neighbours near it. The algorithm works in a manner such that extreme values are brought down, provided that the number of neighbours is not too small and not too large. This encourages our prediction to be more stable.

Prior to our research, we noted that RF regressors and classifiers tend to be powerful models that provide great results **for interpolation problems**. In this study, our goal was to find a model that could provide similar performance (as compared to the RF regressors) that also gave reasonable performance for our extrapolation task.

Approach

Given that we intend to predict prices into the future, our train-test split should not be random, but instead be based on time of transaction. Ideally the training set consists of data from an earlier time period than the data in the validation set. We can then measure performance of our model by comparing the predicted prices and the actual resale prices for the rows in the validation set.

The preprocessed dataset was hence split as such:

- Training set: Rows consisting of resale flat transactions from January 1990 to December 2015
- Validation set: Rows consisting of resale flat transactions from January 2016 to March 2021

Taking reference from prior experience as well as our literature review, we chose to evaluate our model using the RMSE of the prediction compared to the actual price. RMSE has a few properties that are desirable when it comes to being a metric for our models. It penalises larger errors as compared to smaller errors that are spread out in more points. This gives us a model that tries not to leave outliers unaddressed.

We are aware that the current state-of-the-art method to search for hyperparameters is in fact performing a grid-search for the suitable parameters for our model. However we wanted to demonstrate the performance of various parameters to justify various choices of hyperparameters. This is not only just proof that the hyperparameter chosen at the end is optimal, it also acted as a "sanity check" for us. Doing this explicitly decreases the amount of "black-box" operations behind the scene and allows us to see the training loss decreases followed by the overfitting as we increase the number of regressors and as the iterations of training increases. We repeat this over the various models, from RF, SVR and NN, we also experimented with ridge regression, however we discovered that there are very minimal returns of regularising the distributions since this problem does not seem to have multiple underlying linear regressions. Furthermore, we had noted that the individual variables in the problem statement seem to be rather independent from each other.

Results

Results for each trained model

The following results were achieved empirically from running the trained SVR model on the validation set:

Kernel (C=1)	Root-mean-square error
Linear	73337
Radial basis function	184784
Sigmoid	184864

Results of tuning and testing the SVR model.

The following results were achieved empirically from running the trained NN model on the validation set:

No. of layers	Final layer activation	Epochs	RMSE
3	relu	150	108076
3	none	200	103813
		250	103146
5	none	600	90795

Results of tuning and testing the NN model.

The following results were achieved empirically from running the trained RF regressor model on the validation set:

Maximum depth	Number of trees	RMSE
12	100	35318
13	100	35060
14	**110**	**34798**
15	120	35230
16	120	35428

Results of tuning and testing the RF model.

The following results were achieved empirically from running the trained KNN regressor model on the validation set:

No. of neighbours	RMSE
1	56973
3	54433
5	**53687**
7	53800
9	54012

Results of tuning and testing the KNN model.

In addition, we ran a few other regression models for comparison purposes. The following results were achieved from running trained ridge and linear regression models on the validation set:

Type of regression	Root-mean-square error
Ridge	73812
Linear	109137

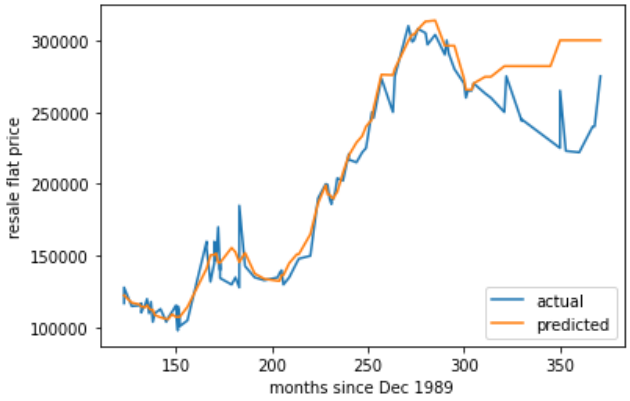
Results of tuning and testing Ridge and Linear regression models.

Results analysis

After analysing our results, we noticed that our results corroborated well with what we had learnt in our literature review. In particular, our experience closely matched that of Gao et al. (2019),[11] since our results showed that RF was the most accurate predictor (smallest RMSE) and linear SVR being one of the runner-ups, right behind KNN regressor.

At first glance of the results, it seems that the RF regressor seems to be the best performing model. However, as previously mentioned, the RF regressor is not useful for extrapolation problems and was only used as a basis for comparison.

Going by RMSE score, the next best performing model was likely to be the KNN regressor. In order to further test our model's ability to extrapolate, we proceeded to identify a particular flat configuration for which there were similar data points (in our dataset) from 1990 to 2021. We then plotted a graph comparing the (extrapolated) predicted price against the actual price. Note that extrapolation occurs from *months = 312* onwards.



Graph demonstrating extrapolation using the KNN model.

The following observations can be made:

1. The model successfully captures the general upward trend of resale flat prices.
2. Although the lines deviate at the point of extrapolation, the decline in house prices (from *months* ≈ *270* onwards) was factored in and represented as a decrease in the gradient of the subsequent incline in predicted price.

Additionally, we have made the following technical insights about our model training and testing process:

1. Most of the models that we have chosen fit the requirements of the application and have certainly met, if not exceeded, the expectations that we had. We can see that there are certain models like RF and KNN that have done quite well in the regression, as seen from them having an "acceptable" RMSE value that is not too large.

[11] Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, T., & Wu, Z. (2019). Location-centered house price prediction: A multi-task learning approach. *arXiv preprint arXiv:1901.01774*.

2. One learning point that was rewarding was that initially, when we had used categorical features with more than 2 classes, some models like SVR had relatively high RMSE. It was after some research that we had realised that the preprocessing we had initially done was not ideal. Hence, we relabelled the data using binary factors via one-hot encoding. This led to the decrease of RMSE in most if not all of the models.
3. It is important to note that extrapolation using regression models can be risky as we can only cross-validate with observed data. With a regression model, we may instead prefer to assume continuity and smoothness of the response surface. This means that the further (from observed data) we attempt to predict prices, the less reliable the predictions tend to be.

Application functionality

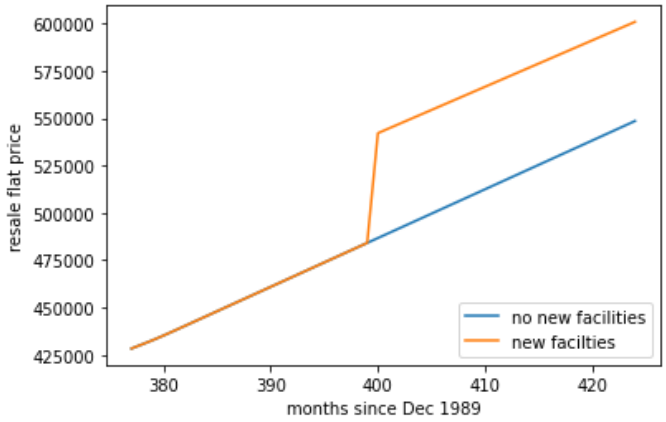
Price prediction

Given our best performing extrapolation model (KNN), we proceeded to showcase the functionalities of our application. As discussed in the "Description" section, the first feature provided by the application would be a direct price prediction, given some configuration of the flat as input. This has been demonstrated in the previous section.

Price visualization

The second feature of the application would be a graphical visualization of the predicted price in the next 5 years. In this use case, we can take on the perspective of a property agent attempting to understand how the predicted price of a resale flat may change, given that an MRT station, hawker centre, bus stop and clinic will be constructed in the vicinity within the next 5 years.

The plot below gives a visualisation of predicted price from April 2021 (*months = 376*) to April 2025 (*months = 424*). It can be seen that upon an addition of the four new facilities in April 2023 (*months = 400*), the predicted price increases.



Graph demonstrating price visualization functionality.

Discussion

Ethical impact

One of the concerns of this project is the inherent responsibility of a price prediction application. Assuming that our user base increases, we have a responsibility of ensuring that our model remains accurate. This is because users may regard the price predictions produced by our application as an "objective" basis for their resale flat purchase. Given that, it is our ethical responsibility as the developers of this application to ensure its accuracy, as well as to inform our users of its accuracy. We intend to do this in a manner that is similar to the way that SRX showcases the accuracy of its X-Value algorithm.[12] This means that we intend to provide timely updates to our users on the performance of our model.

[12] SRX. (n.d.). *X-Value Performance.* https://www.srx.com.sg/XValue-performance

In addition, we intend to conduct regular testing and updating of our price prediction model to ensure that it remains accurate alongside changes in the housing market. We intend to be transparent with our users, and strive to clearly and openly communicate to our users the findings of our regular tests, such as features for which our model's predictions are based on as well as features that are not.

Impact of our work on Singaporeans

With an accurate model, we can cater to the needs of certain subgroups of people who are looking to sell or buy a house here in Singapore. These subgroups are:

1. Homeowners who are looking to sell their HDB flat for various reasons, e.g. to move to another part of Singapore, or to upgrade to another one with more rooms, or due to a divorce or death.
2. Non-homeowners who are looking to buy a HDB flat for various reasons, e.g. moving from another part of Singapore, or as an upgrade from their previous one with fewer rooms, or due to a marriage or a personal choice.
3. Homeowners who have bought the HDB flat as an investment and now want to get returns on their investment
4. Non-homeowners who are looking to buy a HDB flat as an investment as they think prices of HDB flats in that area may rise in the future.
5. Financial advisors who advise clients on whether or not to sell their HDB flat, or on which flat to buy. This model may put them out of a job as people use the model themselves. However, the model may also give them more accurate predictions to work with, thus empowering them to provide their clients with better service.

To these people, we can offer a fairly accurate prediction of the buying/selling price of the HDB flat, given certain factors like the year of transaction, the type of flat, the size of the flat, whether or not there is a hawker centre nearby, and whether or not there is an MRT station nearby.

Our prediction could potentially aid buyers in their planning process, to decide whether or not they can afford the flat, or how much loan they have to take, and how much the flat will be worth in the future, although of course the prediction is most accurate when the difference in years is small. Adding a data pipeline to continuously update the data over the years would also be helpful here.

For homeowners looking to sell their HDB flat, our model could potentially help them to decide when would be the best time to sell, or what selling price to expect if they sell at a certain time. For example, an HDB flat owner who needs money to cover his expenses might decide to sell the flat sooner, even though he would have been able to get a slightly better price for the flat in a few months' time. In contrast, if the predicted price of the flat in a few months' time were to be a lot higher, he might decide to take a loan first and sell the flat in a few months.

Possible applications

In the "Description" section above, we have mentioned two possible features that are catered to industry analysts. As for potential buyers, another feature of our application in future iterations could be a recommendation system that provides users with informed recommendations of resale HDB flats (that are on sale), based on preferences that they select. These options can include their preferred town, budget, the name of the educational institution, etc.

Possible extensions

As mentioned previously, the values signifying proximity of facilities (mrt_station, bus_stop, hawker_centre, clinic) are currently hand-labelled. This can be algorithmically done by calling Google Maps' Distance Matrix API instead. In addition, we could also include other facilities such as community parks and marketplaces.

Another possible extension would be to use travel time via other modes of transport (besides walking distance) as a better measure of the presence of facilities. This can be logically explained in this analogy: Two different

streets are 1km away from the same MRT station, but the former has a longer travelling time to the station than the latter because the buses serving this street do not go straight for the station (stops at many other stops, resulting in increased travel time). Fortunately, this can also be implemented using Google Maps' Distance Matrix API. However, a limitation with using travel time is that live data is required for it, which is susceptible to factors like traffic, weather conditions, bus waiting times, etc. This can translate to inconsistencies and a potentially larger margin of error.

Besides the above mentioned geospatial features, we could also include factors that could negatively influence the prices. A commonly held view regarding property prices in Singapore is that the proximity of certain institutions such as nursing homes, columbariums and foreign worker dormitories might depress property prices. In order to do this, we can conduct further research into flat buyers' sentiments and seek out databases that provide such information.

We can also conduct further analysis to study the impact of the above mentioned features. For example, one might ask several questions about our application given Singapore's constant evolution of the MRT landscape, and a few are listed here:

1. Does an upcoming MRT station within the vicinity (800m) of a street count as the presence of an MRT station?
2. What if the given street already has an MRT station, and there is another upcoming/existing station in the vicinity?
3. What if the closest MRT station is well connected? (i.e. belongs to several different lines)
4. What if there is no nearby MRT station (within 800m), but there is a LRT station?

We think that these are important questions to consider. In particular, MRT stations take a significant amount of time to both plan and build, so a station slated to complete in 1 year might not have the same influence on the price of nearby resale HDB flats as compared to those slated to complete in 5 years. Moreover, a second station in the vicinity of the same street could affect the prices, and the extent of the effect could depend on where this other station improves accessibility to. Connectivity of the nearest station may or may not affect the prices significantly, given that many stations are a few stops away from the nearest MRT interchange with more MRT lines being built.

As we have mentioned previously, accessibility is one of the few key factors that affect the resale price of a HDB flat. In its current form, our model has yet to fully explore the many ways in which accessibility can directly or indirectly affect flat prices, and this will always remain as room for further improvement and enhancement of our model.

Source code

The source code for our application can be found here: https://deepnote.com/project/CS3244-Project-3GyHjAbySuO4qNkQyvkCMw/%2Fnotebook.ipynb

Member roles and reflections

Member	**Project roles**
Joven	Report writing
Jing Han	Implementation of Random Forest model Testing and tuning of Random Forest model Report writing
Grant	Literature review Implementation of SVR model Testing and tuning of SVR model Report writing
Ee Ter	Implementation of price visualisation Report writing
Noelle	Implementation of NN and KNN models Testing and tuning of NN and KNN models Report writing
Zichang	Report writing

Grant: I have had two main takeaways. Firstly, I have learnt the importance of learning from other researchers' lessons by conducting an extensive amount of literature review. This not only allows us to evaluate the novelty of our work, but also provides us with a basis on which our project can be built upon. Secondly, I have also learnt the importance of data preprocessing before feeding it to whichever machine learning models. Although this is not something that has been taught in class, having to implement practical preprocessing methods (such as one-hot encoding) for the project has enriched my understanding of the model training process.

Ee Ter: Through this project, I got to understand data preprocessing through seeing it in action, the different ways it can be done and the impact on the accuracy of the model. After doing the price visualisation, I also learnt the importance of visualising the data and gained insight into how the models operate. I could also apply what I learnt from the lectures to the models. For example, random forests use decision trees, which bin continuous features. As such, it is not as suitable for extrapolation applications, where it just gives a flat line on the graph. Having the opportunity to apply this knowledge on a real life application helped to reinforce the concepts I learnt.

Noelle: I learnt various technical and theoretical aspects of machine learning. Because of this project, I have coded my first neural network model, and learnt that it is not actually the best model to use in all situations. I have also learnt the importance of the human touch in machine learning, that simply cleaning and fitting models is not enough. When developing a model, we must also consider the needs of the target group, in order for the final model to be useful to them.

Michael (Zichang): The one thing I ended up learning in this course which I did not anticipate is how effective tree based regressors can be. Initially before the start of the course, I have known that Random Forest Regressors are well known for their performance as a classifier. However I did not expect this level of performance to flow over to the regressor side too, out performing some of the regression based models (SVR and Linear Models), this is very interesting to me. Perhaps there may be underlying parameters within these models that we have not optimised yet. I would like to look more into these parameters after the course has ended.

Jing Han: Throughout the course, I have come to understand that machine learning is not so easy as we cannot just take the data and ram it into the model. Rather, we have to pre-process the data into a certain format (i.e. categorical/numeric) before we can use it to train the model. Also, initially, I was also expecting the neural network to perform as well or only slightly worse than RandomForestRegressor when it comes to regression but it seems that from the results that it was very far from the truth. It could have possibly been due to lack of data or overfitting.

Joven: I have learnt some basics when it comes to practical machine learning, such as learning what a tensor is, what it is used for, and what are its basic operations. I also learnt a little on CNN and how it's useful in computer vision since it has useful properties such as translational invariance.