Resale price prediction of HDB flats in Singapore Team: Intelligent Realtors

Dolly Agarwal
National University of Singapore

Niranjana Anand Unnithan
National University of Singapore

Nishtha Malhotra
National University of Singapore

***Abstract*—Singapore has been chosen as one of the most expensive countries by The Economist Intelligence Unit for the fifth year in a row. As one of the most expensive cities in the world, accommodation constitutes a very big part of the expenses. In order to provide affordable housing for the general population in Singapore, the government has been offering some public housing options, more generally known as the HDB Flat. For this project, we have taken the HDB resale transaction data and trained a machine learning model to predict the flat resale prices of Singapore. We highlight the importance of various features like region, proximity to various facilities like mrt, malls etc to predict the intrinsic value of a HDB flat. By using a stacked Ensemble model using Catboost and XG Boost algorithm, we could achieve a RMSE of 15946.03 on Kaggle. We also do error analysis of our trained model to understand where it works well and in which scenarios it does not perform well.**

***Index Terms*—regression, resale price prediction, EDA, feature engineering**

I. Introduction

Buying an HDB flat is likely to be one of the biggest financial commitments for a young adult in Singapore and the resale market of HDB flats is a big business. It is very important to understand the market and available statistics on HDB resale prices, transactions etc to make an informed decision. This provides us with a very interesting data analytics question: how much is the real intrinsic value of a house based on objective factors such as the flat type, accessibility of the unit, as well as other amenities and estate information. Either as a buyer or a seller, we need to understand whether the flat is valued at a good price, for which we need to know out of all factors(e.g., size, rooms, type, model, location), what features/amenities are truly important to determine its good value. However, it is not obvious which attributes are indeed most important in a quantified sense. With this motivation, we formulate the following objectives that we tried to accomplish in this project:

1) To understand the key factors affecting HDB resale prices
2) To build a comprehensive pricing model to determine the intrinsic price of an HDB resale flat
3) Do error analysis and discuss limitations and potential extensions

II. Dataset

The core dataset of past HDB resale transaction is publicly available on Data.gov.sg.[1] However, for this project, we were provided with a customized dataset with the following files:

- Train.csv with 431732 previous resale transactions and 17 columns (month, town, flat_type, block, street_name, storey_range, floor_area_sqm, flat_model, eco_category, lease_commence_date, latitude, longitude, elevation, sub-zone, planning_area, region, resale_price).
- Test.csv with 107934 rows and 16 columns (all the columns same as in train.csv minus resale_price which is the label to predict).
- Auxiliary data containing 7 files with supplementary information such as the population demographics and location of MRT stations, malls, commercial centers, government hawker centers and primary/secondary schools.

III. Exploratory Data Analysis

It is important to visualize and understand our data before diving into the regression task. A good EDA can direct us towards required data cleaning/pre-processing steps. It can also give us an intuition on important features and a direction towards required feature engineering.

We started looking at different columns in the provided data and their relationship (if any) with each other and with resale price. Firstly, we looked at the overall statistics of all the columns in train.csv and found the following:

- There are no missing values in any of our columns.
- eco_category and elevation have the same values ('uncategorized' and '0' respectively) for all rows. This means these columns are irrelevant and can be removed in the data cleaning step.
- We have a few categorical features like flat_type, flat_model, region which we need to convert to transform/encode to ordinal columns
- There are 625 duplicate rows. We have 1174 duplicate rows in the test dataset as well.
- flat_type has values like 4 room and 4-room which can be standardized to one format (4 room)
- storey range in the current format is ambiguous with different overlapping range brackets, for example, 2-5, 3-7

[1] https://data.gov.sg/dataset/resale-flat-prices

- Resale month and year are in one column as an object datatype which can be separated into month and year
- There are 13 records in the training data where the resale price year was before the lease_commencement_year. We found two such records in the test data as well, hence we decided not to remove these records.
- We have slightly more data from the west region but overall we have fairly equal distribution of data from various regions. (Fig 1)

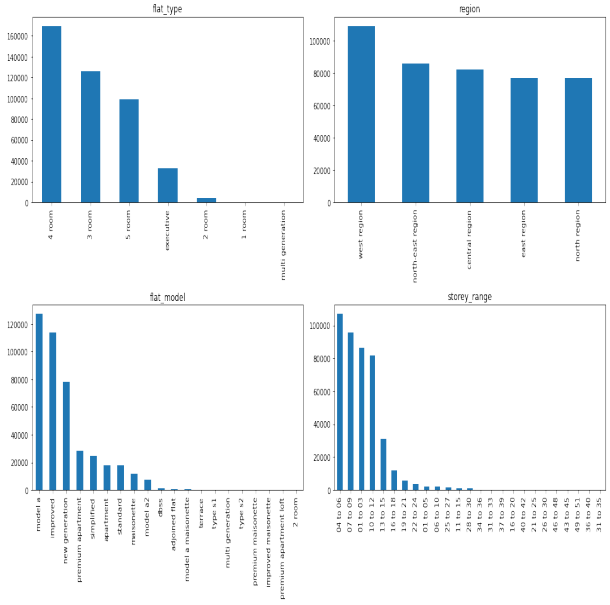

Fig. 1. Graph showing distribution of various features

We constructed a geographical price heat map (Fig 2 to understand the locations and price distributions of resal flats across Singapore. We observed that the flats that ar orange/yellow (meaning higher resale price) tend to lie nea the central area. It can also be observed that flats in the north-eastern region are more lightly colored.

Next, we plotted a graph to visualize resale price distribution (Fig 3). The histogram for HDB resale prices shows a right-skew distribution, with a mean of $3,01,820 and a median of $2,83,950. The resale prices ranged between $29,700 to $11,23,200. So, to handle this skewness we can transform it to log/sqrt which might help during our regression task.

Then, we tried looking at the resale price trend over the years by plotting a line graph (Fig 4) of average resale price over sale year. Interestingly, we found that the HDB resale price started to increase starting from 2006 and reached a peak in 2013. Then the resale price dropped and overall price remained a bit stable from year 2014 until 2020. This showcases that Singapore HDB price might have been affected by government policies and overall economic trend.

After looking at the overall resale price trend, we wanted to look at price trends based on different regions. So, we plotted a line chart (Fig 5) to see the trend of average resale price over the years region wise. We found that the average HDB

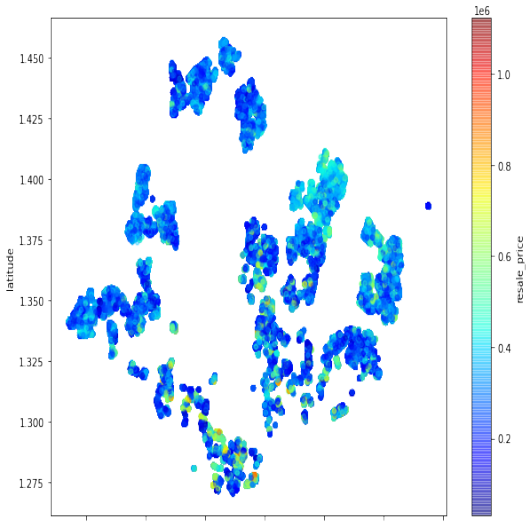

Fig. 2. Geographical price heat map of HDB flats

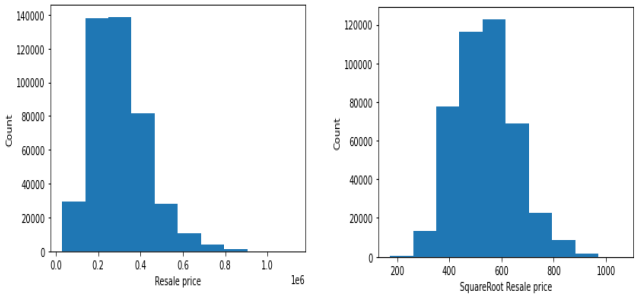

Fig. 3. Graph showing distribution of resale price

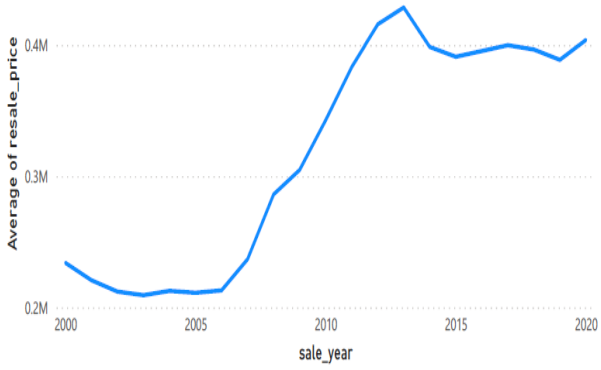

Fig. 4. Graph showing average resale price trend over the years

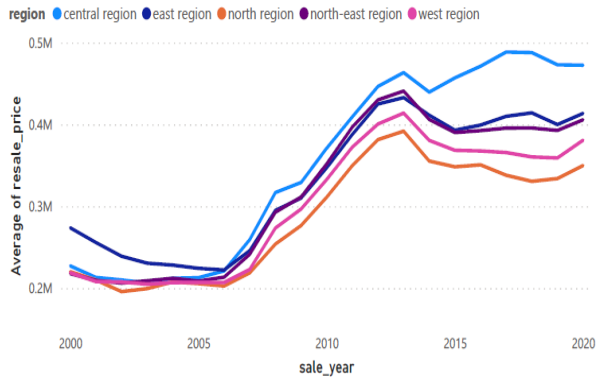

Fig. 5. Graph showing average resale price trend over the years by region

price of a flat in the central area goes up abruptly in around 2013 with an even higher slope than previous years. By this we can infer that, today (2021) it will cost us more to buy a flat in the central region of Singapore than other regions.

To explore distribution of resale price based on various features, we plotted Box plots and Bar charts. In the context of our project, this can help us understand the distribution of the flat price among different flat types, models and storey range. By looking at the charts (Fig 6-8), we can infer the following:

- 5-room, executive and multi-generation flat types have higher average resale price compared to 1-4 room flat types.
- Resale prices seem to be higher for flat at the higher storey.
- Resale prices of flat models like type s1, type s2, premium apartment, dbss are quite high compared to other flat models

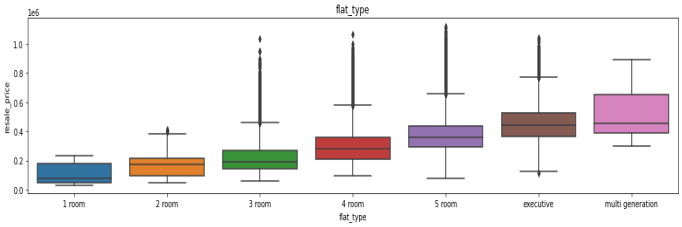

Fig. 6. Box Plot to show distribution of resale price across flat types

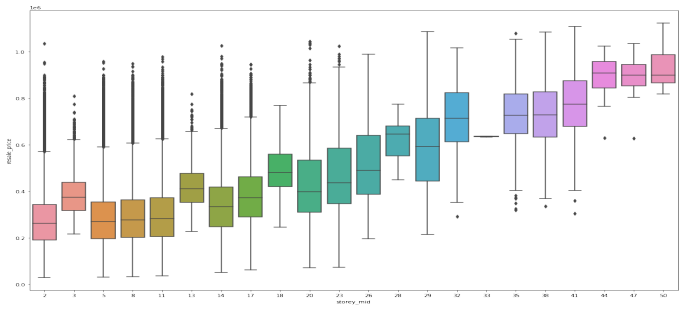

Fig. 7. Box plot to show distribution of resale price across storey

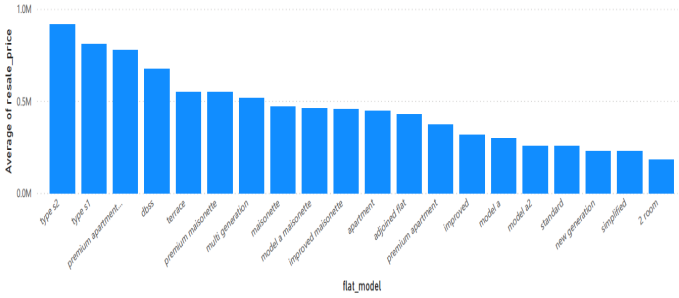

Fig. 8. Bar chart to show distribution of resale price across flat model

Finally, we did our EDA on auxilliary data (Fig 9-12). We tried looking for trends in resale price based on it's proximity to MRT, primary school, government Hawkers, commercial centers and CBD.

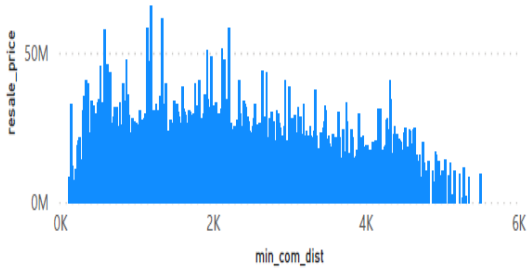

Fig. 9. Graph showing relationship between resale price and minimum distance to Commercial Centers

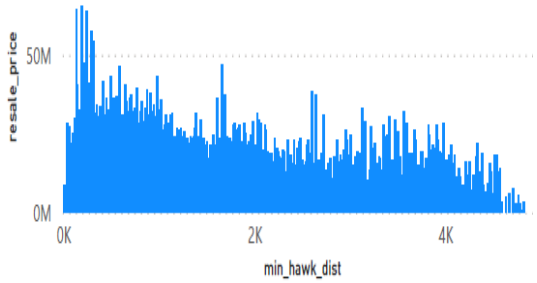

Fig. 10. Graph showing relationship between resale price and minimum distance to Hawkers

IV. Data Pre-processing and Feature Engineering

As informed by our EDA, we took various data cleaning and preprocessing steps mentioned in section A. We also worked on various new features from the existing features which are mentioned in section B

A. Data Preprocessing

As informed by our EDA we performed the following data cleaning steps:-

- Removed columns eco_category and elevation

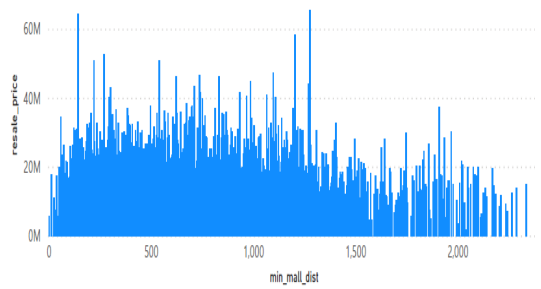

Fig. 11. Graph showing relationship between resale price and minimum distance to Malls

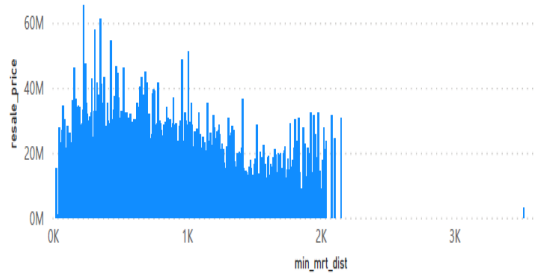

Fig. 12. Graph showing relationship between resale price and minimum distance to MRT

- Removed hyphen ('-') from flat_type values to standardize all the value
- Transformed resale price using sqrt transformation
- Converted categorical column storey range to numerical column storey_mid by calculating the average of the range per record. For example, if storey_range is 10 to 12, storey_mid is 11
- Converted month column to resale_month and resale_year
- Standardized all the numerical columns

B. Outlier Detection

To detect any potential outliers, we first looked at the box plot of resale price. In regression estimation problems, outliers can influence the algorithm significantly and distort the results and predictive power of the model. We plotted box plots to

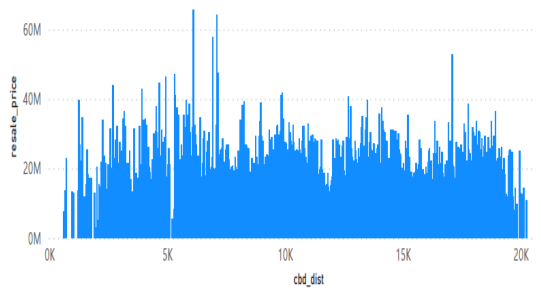

Fig. 13. Graph showing relationship between resale price and minimum distance to CBD

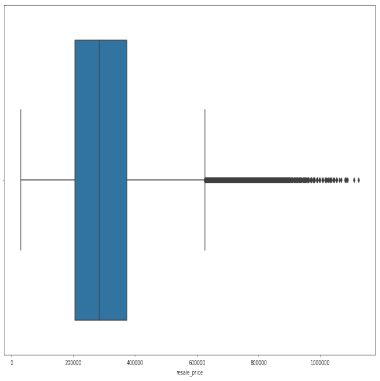

Fig. 14. Box plot showing resale price distribution

detect outliers. We dealt with them carefully and did not delete a lot of rows in the process as it may create biased model.

To remove as few records as possible, we squished extreme values closer to the rest of the data so that they get better chance to not be identified as an outlier in the first place. Hence, we log-transformed resale price before the outlier detection. Based on our graphs, we could see some records with a very high value of resale price. We tried extracting these records to have a closer look. For outlier detection, we extracted records which are three standard deviations away from the mean resale price (Fig 14). The standard deviation approach arises from the so-called empirical rule which states that given a normally distributed variable, approximately 99.7% of the data is within three standard deviations. We found 764 such records. We considered this number a bit high and felt that removing them as outliers could potentially lead to a biased sample. Out of these 764, we chose top 32. The experiment after removing top 32 outliers is mentioned in the result section.

C. Feature Engineering

Apart from cleaning and pre-processing existing data, we looked for ways to convert our categorical features like town, flat_type, flat_model to numerical features as most of the ML models work only with numerical features. This would also help capture relationships between the categories if any. Following are the new features we worked on along with the rationale behind our decision.

- Convert Town categorical column as binary (0/1) based on whether it is a mature town or non-mature town.
  Rationale: HDB resale transaction price is determined by the town in which the flat is located in (e.g.Bukit Timah Town). The more mature the town in which the flat is located in, the higher the resale price. We categorized every town into mature and non-mature towns. We found the list of mature towns from Property Guru and converted the town categorical column into a numerical column by flagging it 1 if its a mature town, 0 otherwise.

TABLE I
LIST OF MATURE AND NON MATURE TOWNS

Mature Towns	Non-Mature Towns
ang mo kio	bukit batok
bedok	bukit panjang
bishan	choa chu kang
bukit merah	hougang
bukit timah	jurong east
central area	jurong west
clementi	punggol
geylang	sembawang
kallang/whampoa	sengkang
marine parade	woodlands
pasir ris	yishun
queenstown	
serangoon	
tampines	
toa payoh	

- Converted flat_type ordinal values.
  Rationale: Based on the average resale price of each flat type (Fig 6), we created an ordinal feature to get more meaningful results. The ordinal values given to each flat type is as follows: '1 room':1, '2 room': 2, '3 room': 3, '4 room': 4, '5 room': 5, 'executive': 6, 'multi generation':7

- Converted flat_model into ordinal values.
  Rationale: Based on the average resale price of each flat model (Fig 8), we categorized them into 3 categories and used it as an ordinal feature to get more meaningful results. Category 1: Improved, model a2, model a, standard, new generation, simplified, 2 room Category 2: Maisonette, model a maisonette, improved maisonette, apartment, adjoined flat, premium apartment Category 3: typeS2, typeS1, Premium Apartment loft, DBSS, Terrace, Terrace, Premium maisonette, Multi generation

- Converted region into ordinal values.
  Rationale: Based on average resale price per region (Fig 5), we assigned an ordinal value to each region as follows: 'north region':1 , 'west region':2,'east region':3,'north-east region':4,'central region':5

- Created a new column: remaining_lease_year.
  Rationale: All HDB flats come with a 99-year lease from the government so the years left on lease might be an important factor for predicting the resale price. So, we calculated remaining lease year as follows:- remaining_lease_year = 99 - (resale_year - lease_commencement_year)

- Added Demographic composition to capture percentage of working adults (20-60), percentage of retired (60+), percentage of young people in the estate(0-19).
  Rationale: We divided into demographics 3 categories as the working adults would prefer to stay as close to mrt/comm centers for commute to work. Similarly proximity to schools especially primary schools would matter if there are a lot of young kids in the area. We applied grouping based on planning_area and aggregated counts of the each demographics.

  All the distances below are calculated using geopy library.
- Created columns to capture distance to Central Business District (CBD).
  Rationale: Central Business District (CBD) is Singapore's business and financial district, and home to leading international businesses and financial institutions. Our hypothesis is that flats in close proximity to CBD will be expensive.

- Created columns to capture minimum distance to nearest mrt, malls, hawkers, commercial centers and primary schools
  Rationale: Proximity of HDBs to facilities such as mrt, malls, hawkers, primary schools can be important factors in determining its resale value.

- Number of nearest primary schools, MRT, malls and hawkers
  Rationale: Based on our EDA (Fig 9-12), it can be seen that if the minimum proximity to MRT is roughly 1.3 km the resale prices are generally higher. Using the same rationale we calculated n_neighboring facilities i.e number of mrts, malls, primary schools, hawkers, commercial centers within the distance radius which matters as seen from the graphs. Example: n_neighbouring_mrt within 1.25 km , n_neighbouring_malls within 1.3 km, n_neighbouring_pschools within 1.5 km, n_neighbouring_hawks, n_neighbouring_commcenters

V. EXPERIMENTS

After preprocessing and feature engineering, an initial experiment was conducted using different regression models to get an intuition of the best model. For this, we first did a study to select best features for our experiments. Then, we did our evaluation by splitting the training data into train-validation split as described in the sub-section B. After having some initial results on the model performance (sub-section C), we did K-Fold cross validation to do hyper parameter tuning of the best models. It is to be noted that along with the performance, we considered run time as an important factor while doing model selection.

A. Feature Selection

After feature engineering, we had around 26 predictors. So, we considered a test for multicollinearity an important step to get some insight on the relationship between features (if any). To do this, we created a correlation heatmap to find correlations not only amongst independent variables but also between the target (resale_price) and each possible feature.

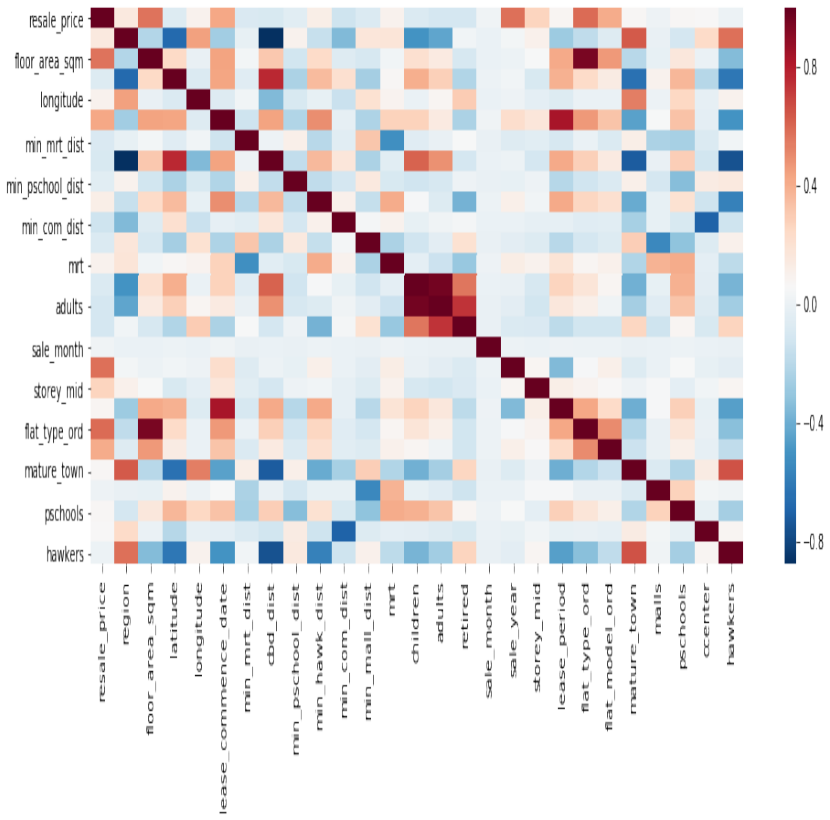

Fig. 15. Correlation Matrix

As can be seen in Fig. 15, features flat_type_ord, sale_year and floor_area_sqm are positively correlated with the resale price. Also, it can be seen that features flat_type_ord and floor_area square are highly correlated with each other. Lease commence date and lease period are also positively correlated. Having features with high correlation might increase the chances of errors. Therefore, it is important to choose the best features for the model. Initially, we experimented with features based on this correlation matrix information and our intuition. But it's not obvious to judge the usefulness of a feature in isolation. Also, we realized it is not possible to try so many combinations for the optimal result. Therefore, for this important task, upon our research we found a feature selection algorithm - Recursive Feature Elimination (RFE). RFE is easy to configure and use; it is also effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable. When using RFE, we need to specify the number of features to select and the regression algorithm used to help choose the feature. Taking runtime into consideration, we kept estimator as catboost, and tried to find the best no of features by tuning this hyperparameter using in the range 15-23 using GridSearchCV. We found 19 as the optimal number of parameters based on test rank results returned. We then extracted these 19 features using RFE which are most important in the prediction of resale price: 'region', 'floor_area_sqm', 'latitude', 'longitude', 'lease_commence_date', 'min_mrt_dist','cbd_dist', 'min_hawk_dist', 'min_com_dist', 'min_mall_dist', 'mrt', 'adults_per', 'sale_month', 'sale_year', 'storey_mid', 'lease_period', 'flat_type_ord', 'flat_model_ord', 'mature_town'.

B. Train/Validation Split

We kept the distribution of test and validation set similar to get the best understanding of how our various models are performing. Which means, if our train set has 50% of flat in west, 30% of flat in central and 20% in east, we maintained this ratio in our validation set. For this, we considered region as a criteria to do Stratified Sampling of train and validation set. We split our training data in 80-20 percent ratio into train and validation set.

C. Model Selection

To get some intuitions, we experimented with a few regression models to start with. Different regression models from Sci-kit Learn library were analyzed to identify models that perform the best with given data. Following features from the dataset were used to train the models: region, floor_area_sqm, latitude, longitude, lease_commence_date, min_mrt_dist, cbd_dist, min_hawk_dist, min_com_dist, min_mall_dist, mrt, adults, sale_month, sale_year, storey_mid, lease_period, flat_type_ord, flat_model_ord, mature_town,

malls, pschools. The models were initially trained with default parameters followed by tuned hyper parameters. It was observed that the tuning of hyper parameters significantly improved the performance of the models.

Table II gives the summary of the models, hyper parameters used for each of the models, time taken for training and corresponding RMSE values. The best model is selected based on: quality of fit to the validation data set and run time.

TABLE II
ANALYSIS OF REGRESSION MODELS

Model	Hyperparameters	Time	RMSE
Catboost	learning_rate:4 depth:8	6m	16233.44
XGBoost	max_depth:40 min_samples_split:25 criterion:mse random_state:0	14m 18s	17571.50
GradBoost	max_depth:40 min_samples_split:25 criterion:mse	6m 35s	18737.67
RandomForest	max_depth:40 min_samples_split:25 criterion:mse	16m 14s	18490.21
Decision Tree	max_depth:40 min_samples_split:25 criterion:mse random_state:0	6s	22374.66
KNN	n_neighbors:5	1m 44s	29773.76
Linear Regressor		561ms	55202.99

It was observed that Catboost Regressor (RMSE=16233.44) and XGBoost Regressor (RMSE=17571.50) gave the best results. The time taken for training by Catboost model was also less compared to other ensemble models evaluated. The LinearRegressor model gave a comparatively low result (RMSE=55202.99). This suggests that there could be some non linear elements in the dataset which the Linear Regression model could not capture.

Based on the above observations, Catboost Regressor and XGBoost Regressor were selected to carry out further experiments. Catboost regressor is an algorithm that performs gradient boosting on decision trees. It offers multiple advantages like improved accuracy, fast prediction and categorical feature support. Since the dataset contained a large number of categorical features and Catboost offered fast training, it was initially preferred to conduct experiments.

Both Catboost and XGBoost work on the technique of gradient boosting on decision trees. The catboost model grows oblivious trees which provides simple fitting and efficient CPU utilization. Oblivious trees are grown in such a way that all nodes at the same level test for the same predictor with same condition and index of the leaf is calculated using bitwise operations. XGBoost improves the base gradient boosting framework through system optimization and algorithmic enhancements.

Although Catboost regressor and XGBoost regressor are able to provide low root mean square error value compared to other models, determining the relationship between prediction features and target variables is challenging.

D. Hyper Parameter Tuning using K-Fold Cross Validation

As described in subsection C, we found Catboost as the best model followed by XGBoost. For Hyperparameter tuning of Catboost, we performed K-Fold cross validation with k=5 and tuned two parameters: learning_rate and max_depth. We found optimal learning_rate and max_depth as 0.4 and 8 respectively. For XGBoost, we performed hyper parameter tuning using GridSearchCV. Although, it returned best parameters as max_depth: 60, min_samples_split: 12, it was taking a very long run time, so we selected max_depth=40, min_samples_split=25 instead.

We also relied on K-Fold cross validation while experimenting with performance of the catboost and XGBoost on different feature combination.

E. Ensemble Model

Ensembling is a very useful techniques in machine learning in boosting up the accuracy of the overall model. The rationale behind is that, different classifier is learning the pattern of the data set from different perspective. We wanted to exploit the results of both Catboost and XGBoost models. Hence, we experimented with combining the results of both by using an Ensemble Method called Stacking. Fig 16 is the architecture of stacking ensemble model that we used. We split our train set into 80% training and 20% validation set by doing Stratified Sampling on region. We train Catboost and XGB regressors on our training data and get different predictions for our validation set. We use both the predictions to train a Linear Regression model, which learns from the two predictions and the ground truth of the validation set. The intuition behind this is that the Linear Regression model is deciding the weights on the predictions made by the different regressors. On different scenarios, either of Catboost and XGBoost models perform better than the other, so the Linear Regression model will optimize and find the best combination. By gathering and learning from the results of both the regressors, we could see some performance gain.

F. Results and Discussion

The results of our initial experiments with various models are mentioned in Model Selection section. This section describes our experiments and results after selecting Catboost as our first preferred model choice and then implementing a stacked ensemble method which was mentioned in section E. As mentioned earlier, before doing RFE, we experimented with different feature combination and used K-Fold cross validation to validate our results. Taking the run time into account, we performed most of these experiments only on Catboost and then used the best validation results on ensemble model to get predictions for the test set. Table III shows our various experiments and the performance on test set on Kaggle.

We define FeatureSet as the initial features that we considered and it includes the following features: floor_area_sqm, lease_commence_date, latitude, longitude, min_mrt_dist, cbd_dist, min_hawk_dist, min_com_dist,

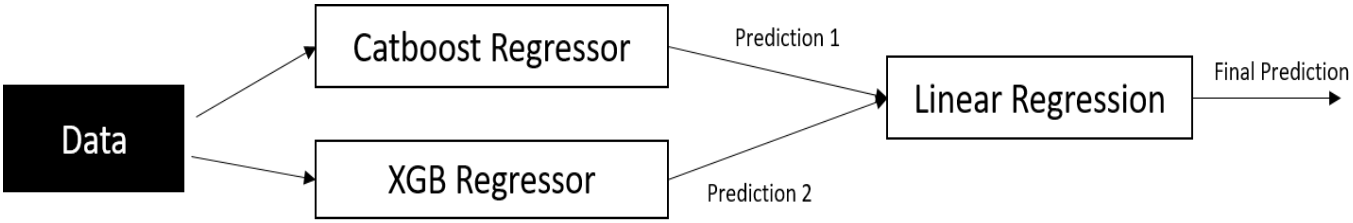

Fig. 16. Ensemble model architecture

TABLE III
TEST SET PERFORMANCE ON KAGGLE

Model	Features	RMSE	Remark
Catboost	FeatureSet + children	16593.183	
Catboost	FeatureSet + region	16383.67	
Catboost	FeatureSet + region + malls	16383.67	sqrt transform resale price
Stacked Ensemble	FeatureSet + region + malls + pschools	16008.07	sqrt transform resale price
Stacked Ensemble	FeatureSet + region + malls + pschools	16021.55	Removed outliers
Stacked Ensemble	FeatureSet + adult percentage	15946.03	sqrt transform 19 best features

min_mall_dist, mrt, adults, children, retired, storey_mid, sale_year, sale_month, lease_period, flat_type_ord, flat_model_ord, mature_town.

As can be seen from the Table III, we made the following observations:

- Transforming the resale price using sqrt to handle skewness improves predictions.
- There is a performance gain by using Stacked Ensemble as it uses best combination from both Catboost and XGBoost models
- Performance degrades after removing the outliers (discussed in section IV). This means, our initial hypothesis that removing records with a very high resale price might bias the model and degrade the performance is true.
- The model gives best performance on the 19 features returned by our RFE model which includes flat details along with proximity to mrt, hawker, cbd, commercial center, mall and demographics info like adult percentage

Now, let me discuss our best model in detail. As mentioned, our best model used the 19 features returned by our RFE algorithm, which are mentioned in A. First, we trained a catboost model and got an average RMSE value 16450.2 on our validation set. Fig 17 shows the predicted vs true value of this model prediction.

Having some initial results on catboost, we started some experiments with the ensemble model. We split our train

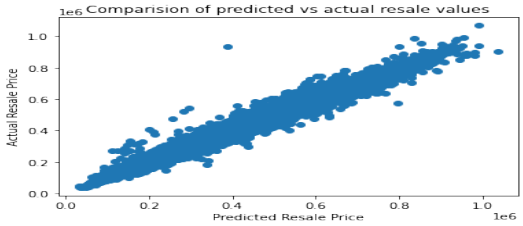

Fig. 17. Graph showing predicted vs true resale price on validation set for Catboost model

data into 90% train and 10% test sets. Then we used the train set to do K-fold cross validation with k=5. In each fold run, we use train subset to train Catboost and XGBoost and predict on the validation set. We use these validation predictions and the ground truth to train the Linear Regressor. We then predict resale price on the held out test set using trained Catboost, XGBoost and Linear Regression models. The idea was to get some intuition behind how various models are performing. It would also give us information regarding whether the ensemble approach is actually helping or not. The average RMSE on test set from Linear Regression is 16342.4. The validation results of various models on 5 folds are mentioned in Table IV.

TABLE IV
VALIDATION RESULT OF K-FOLD CROSS VALIDATION

Folds	Catboost	XGBoost	Linear Regression
Fold 1	16621	17925	16337
Fold 2	16659	17909	16365
Fold 3	16628	17963	16360
Fold 4	16666	17869	16368
Fold 5	16569	17895	16310

Clearly, we can see that stacked ensemble is improving results. So, we used this architecture to train on full data. It took 16 mins 15 seconds to train the ensemble model. This ensemble model trained on 19 features is the model which gave us our best score on Kaggle. Having these results, we do an in depth error analysis on our validation set to understand scenarios were our model is performing good and where it fails. Next section discusses our analysis in depth.

G. Error Analysis

We used the percentage error between predicted value and true value of validation set to perform error analysis. Some of

the features which are distances or area are standardized.

Highly correlated features are floor area, sale year, flat type. Medium correlated features are region, lease commencement date, n_neighboring_mrts, storey, flat model.

True Predictions analysis when error percentage is less than 1% can be seen in Fig. 18. Predictions are more accurate when flat_type_ord are generally smaller, sale year is older, lease commencement date is older and the region is in the north or west. Our model predictions are more accurate for cheaper HDBs as we can see in the True Price chart of Fig 18.

False Predictions analysis when the predicted price is 30% lower than actual price can be seen in Fig.18. There were a total of 20 predictions out of 86347 in the validation set whose predictions were 30% higher than the actual price. Model predicted lower scores due older flat models and lesser lease period. But these flats have a bigger floor area and better quality flat_type which could be the reason for their higher actual price. Price fluctuations are also caused due to economic factors such as inflation rate which is not considered in our model. The model tends to behave badly when two highly correlated features like flat_model and flat_type (Fig. 15) contradict with each other, for example in case of 5-room flat type which has a simplified flat model. Similarly, sale year and lease period are negatively correlated with each other, the lower the sale year - higher will be the lease period. However, in the cases where the sale year is very recent like 2020 but lease period left is still 85 years, the results predicted by the model are unreliable.

False Predictions analysis when predicted price is 30% higher than the actual price can be seen in Fig. 18. There were a total of 111 predictions out of 86347 in the validation set whose predictions were 30% higher than the actual price. Although the flat_models are quite old and flat_type are smaller, the predictions are higher due to proximity to CBD and closest mrt, presence of many mrts in the vicinity of 1.3 km.

We are missing the information about the busstops, which also determines the accessibility of HDBs. For example, for places like Woodlands some HDBs are seen with higher prices even though they are far from MRT. In such cases proximity to busstops with express buses lead to the increase in price of the HDB.

H. Conclusion

Based on our findings, we have concluded that HDB resale prices in Singapore are largely affected by key factors such as floor area, the maturity of the town and proximity of the flat to the CBD as well as to facilities such as mrt, malls. With the error analysis we identified that the model performs well when the HDBs are located in north/west. We also identified some of the limitations and future scope such as information about economic factors may help in understanding rate of price rise. Data about the bus stops may help us identify more information about accessibility.

I. Statement of Contributions

Dolly, Niranjana and Nishtha contributed equally to writing the report. Each one of them were involved in model training, hyperparameter tuning and experiments at various occasions. Dolly worked on EDA, feature selection using RFE, KFold cross validation and the final ensemble model architecture. Nishtha worked on auxillary data analysis and extraction, and error analysis. Niranjana worked on model analysis and evaluation, calculation of minimum distance from auxiliary data, outlier detection and removal, and time series analysis.

REFERENCES

[1] Resale Flat Price Dataset from Singapore government https://data.gov.sg/dataset/resale-flat-prices

[2] LEE Hui Xin Anne, TAN Sok Yi, TENG Jing Wen, "Visualising Singapore's HDB Resale Prices"

[3] Philip Chan (Nov 12, 2020), "Understanding and Predicting Resale HDB flat prices in Singapore". https://towardsdatascience.com/understanding-and-predicting-resale-hdb-flat-prices-in-singapore-1853ec7069b0

[4] Mengyong Lee (Jan 15, 2019), "https://towardsdatascience.com/data-driven-approach-to-understanding-hdb-resale-prices-in-singapore-31c3beecfd97". https://towardsdatascience.com/data-driven-approach-to-understanding-hdb-resale-prices-in-singapore-31c3beecfd97

[5] Distance Calculation using Geopy "https://www.geeksforgeeks.org/python-calculate-distance-between-two-places-using-geopy/"

[6] List of Mature towns "https://www.propertyguru.com.sg/"

[7] Jason Brownlee (August 28, 2020), "Recursive Feature Elimination (RFE) for Feature Selection in Python" "https://machinelearningmastery.com/rfe-feature-selection-in-python/ "

[8] Deepanshu Bhalla (June 1,2017),"Select Important Variables using Boruta Algorithm" "https://www.datasciencecentral.com/profiles/blogs/select-important-variables-using-boruta-algorithm"

[9] Valerie Lim (Feb 19,2020), "Using Regression Analysis to Predict HDB Resale Prices in Singapore" "https://www.datasciencecentral.com/profiles/blogs/select-important-variables-using-boruta-algorithm"

[10] Outlier Detection in Real estate Data "https://becominghuman.ai/outlier-detection-in-real-estate-data-4e7375e2c8ba

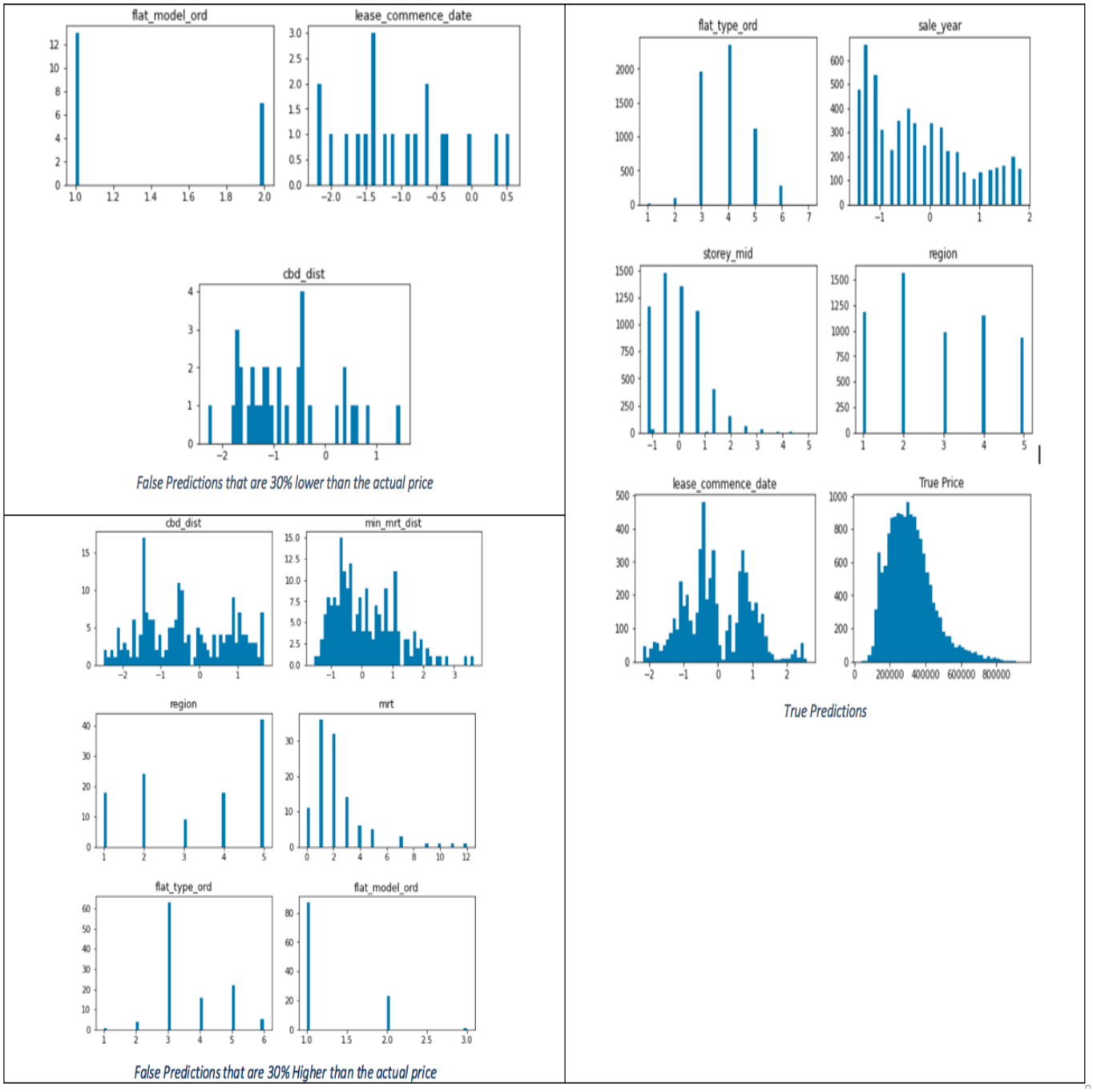

Fig. 18. Error Analysis