AIRBNB PRICE PREDICTION ANALYSIS IN SINGAPORE

Team Members:
Jia Xiang, Kowsalya, Sanofer, Rukaiyaa, Luqman, Si Kai, Tanisha

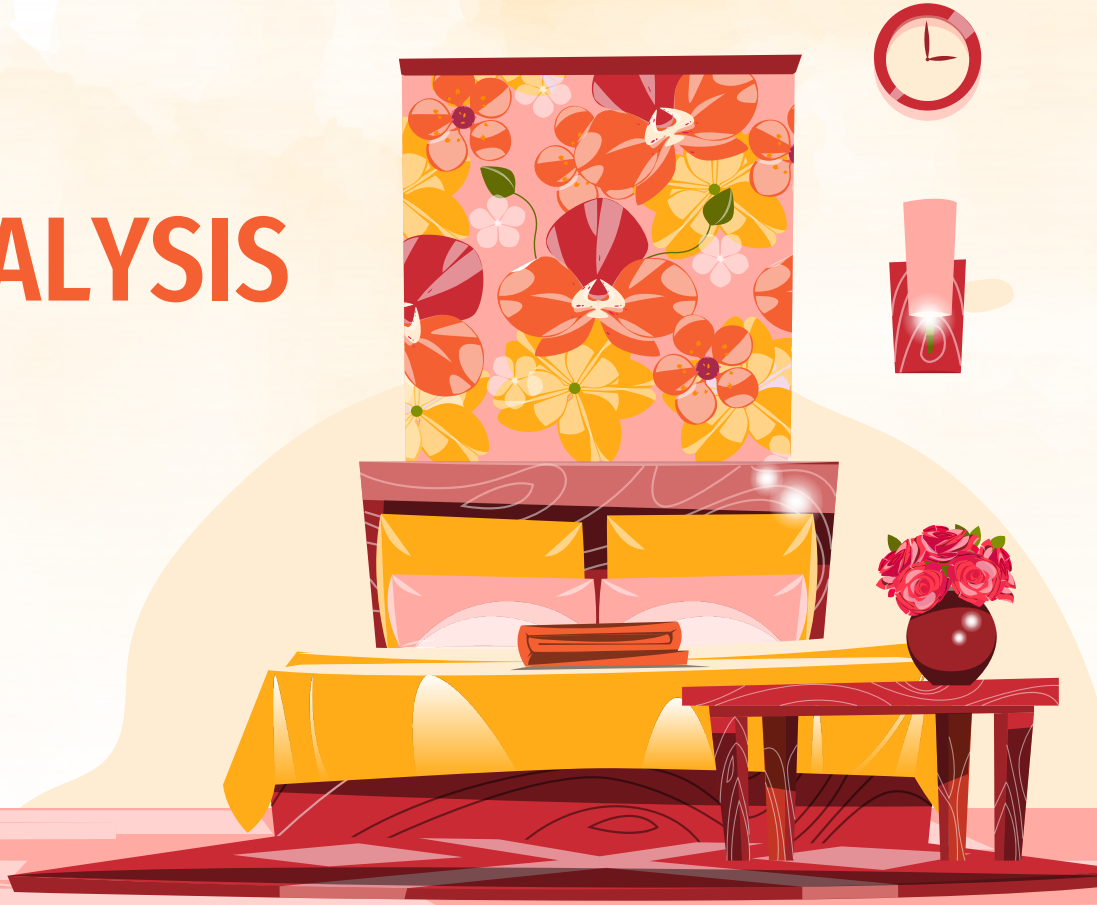


TABLE OF CONTENTS

01

INTRODUCTION

Background and problem statement

02

MOTIVATION

What is the reason we are doing this?

03

LITERATURE REVIEW

Previous related works

04

DATASET

Description of data used

05

METHODOLOGY

Machine Learning technologies used

06

PREPROCESSING

EDA, Feature engineering

07

FINAL MODEL

Regression model, final results

08

CONCLUSION

Overview of project and future work

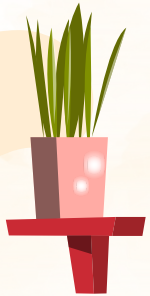
09

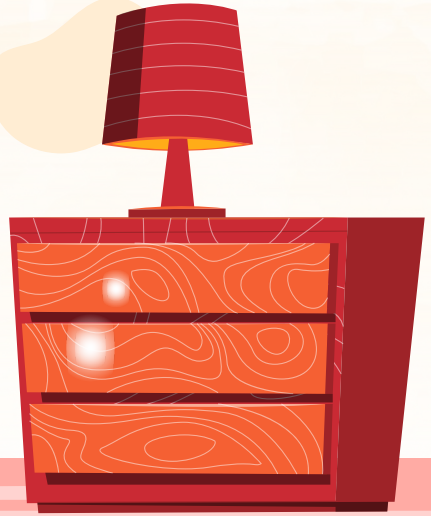
REFERENCES

Citations of works referred to

01

INTRODUCTION





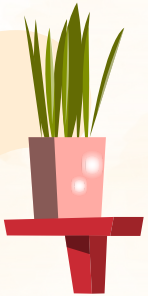
- Founded in 2008
- An American company based in San Francisco for short-term homestays and experiences
- The company acts as a broker and charges a commission from each booking.
- As of Dec 2021 - Company revenue of 5.99 billion USD and > 6000 employees worldwide

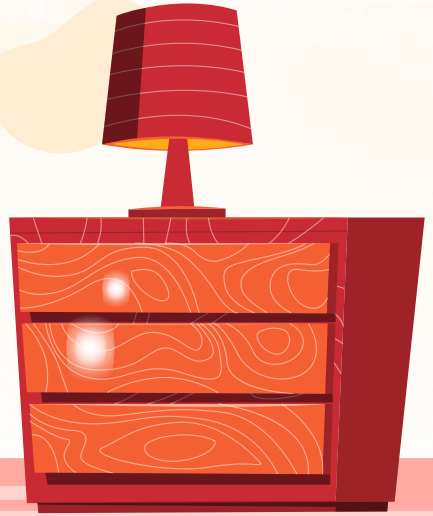
PROBLEM STATEMENT

Finding the optimal model to predict pricing of an Airbnb listing in Singapore

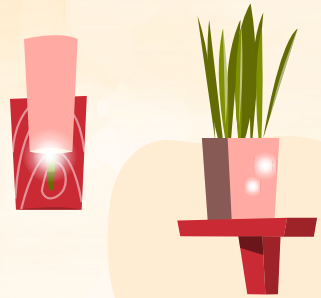
02

MOTIVATIONS





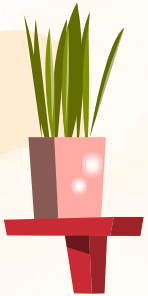
MOTIVATIONS

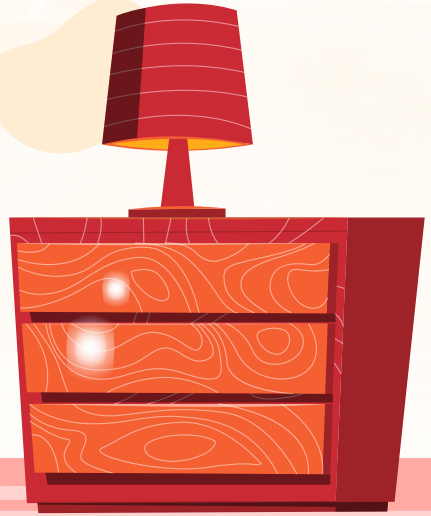


1. Understand and adapt to the preferences and demands of consumers looking for rental spaces to stay
2. Clearly explain the derivation of such prices to be more objective in pricing decision
3. Avoid overpricing or underpricing listings

03

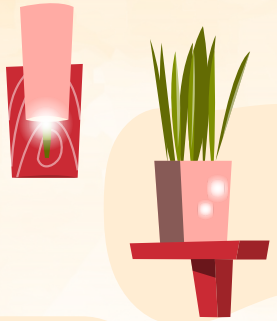
LITERATURE REVIEWS





Literature Review 1

"Predicting Airbnb Prices with Machine Learning and Location Data"



Overview

Several listing features to predict price, added bonus of predictor based on space: proximity to certain venues eg: bar, market

Approach & Models Used

Log-transformed data to normally distribute price, one hot-encoding categorical features, and standardised using StandardScaler()

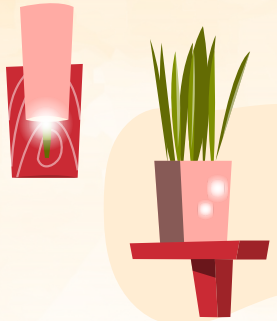
Spatial Hedonic Price Model (OLS Regression), Gradient Boosting with XGB Regressor

Conclusion

XGBoost performs better than the Spatial Hedonic Price Model, however it only predicts 66% of the variation in price. This model was built without the reviews column.

Literature Review 2

"Use of Dynamic Pricing Strategies by Airbnb Hosts"



Overview

Tool to maximise profit by adjusting the price of a product or service continuously to meet the fluctuation demand

Approach & Models Used

Has a unique position in the sharing economy;

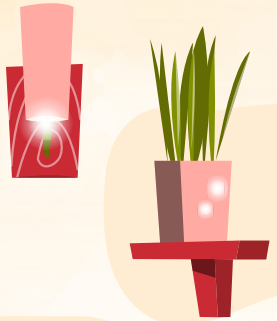
Hosts set daily, weekly, and monthly room rates and control prices over the time

Conclusion

Airbnb hosts make limited use of dynamic pricing; push for host motivation to better understand determinants of a listing when pricing to better adhere to fluctuations in price

Literature Review 3

"Airbnb Price Prediction using Machine Learning and Sentiment Analysis"



Overview

The pricing of property and the evaluation of price of a property is one of the most challenging issues owners/customers face in regards to gauging the value of a property and ensuring optimal pricing.

Approach & Models Used

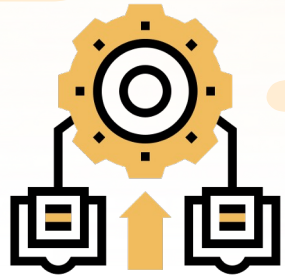
Sentiment Analysis (investigating into customer reviews)
Feature importance analysis (manually selecting features to reduce model variance)

Linear regression, Tree-based models, SVR, KMC, NNs

Conclusion

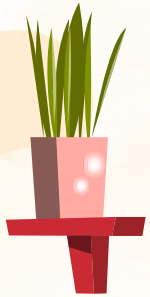
Excessive amounts of features will result in high variance and a weak performance of the model.

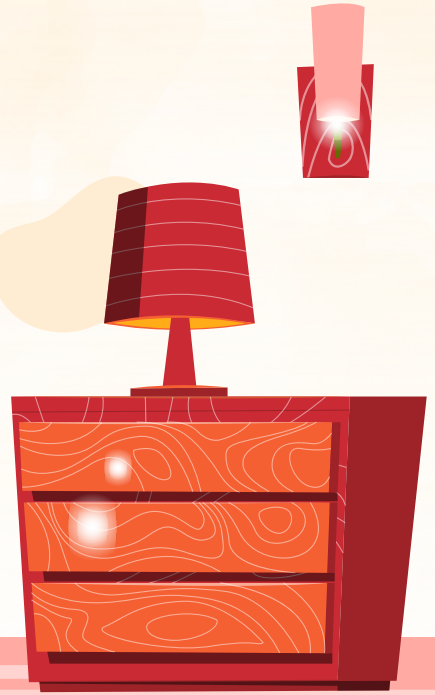
Using the SVR method produced the best results: R Square = 69%
MSE = 0.147



04

DATASET





Airbnb Listings and Reviews (Dataset)

About Dataset: The datasets below contain information about all Singapore Airbnb listings that were live on the site on July 2019. The reviews dataset includes all reviews from those listings! Reviews_translated.csv was generated from the original dataset to produce a set with all the reviews translated to English!

Source: **https://www.kaggle.com/datasets/sarvasaga/airbnb-singapore-listing**

Dataset	Number of Columns	Number of Rows
Reviews for Airbnb listings.csv	6	10919
Airbnb listing.csv	98	8293
Reviews_translated.csv (*generated from reviews dataset)	7	10893

MRT, Malls, Tourist Attractions (Dataset)

About Dataset: The following datasets focus on mall coordinates, tourist attractions, and MRT exits throughout Singapore.

Sources:

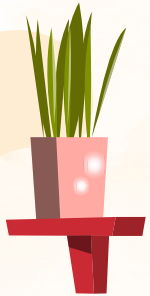
- https://github.com/ValaryLim/Mall-Coordinates-Web-Scraper
- https://data.gov.sg/dataset/tourist-attractions
- https://data.gov.sg/dataset/lta-mrt-station-exit

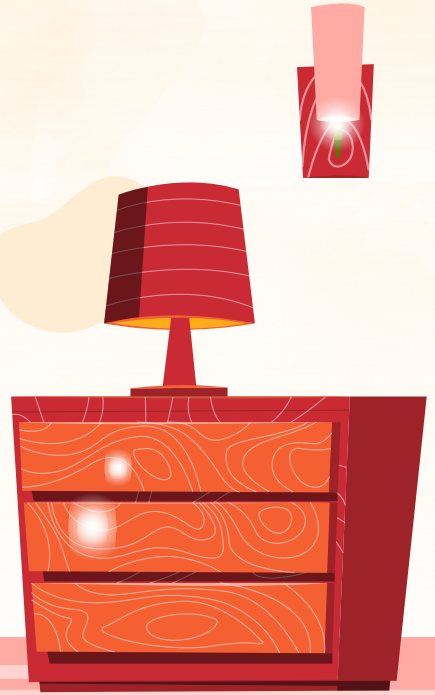
Dataset	Number of Columns	Number of Rows
mall_coordinates_upated.csv	3	183
TOURISM.csv	23	106
lta-mrt-station-exit-geojson.csv	6	474

05

METHODOLOGY

(Tools & Resources)





Data Manipulation and Modelling

working with arrays and performing element-wise operations on them

data science/data analysis and machine learning tasks

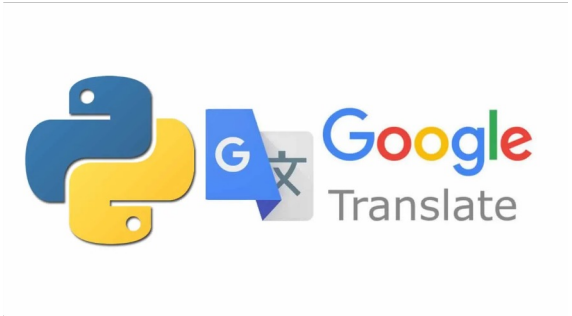
provides a selection of efficient tools for regression machine learning models among others

Data Visualization

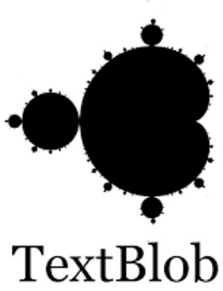
creating static, animated, and interactive visualizations - create 2D graphs and plots

provides a high-level interface for drawing attractive and informative statistical graphics

Sentiment Analysis



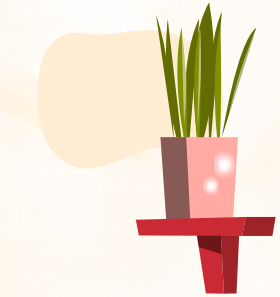
fast and reliable, auto language detection, bulk translations

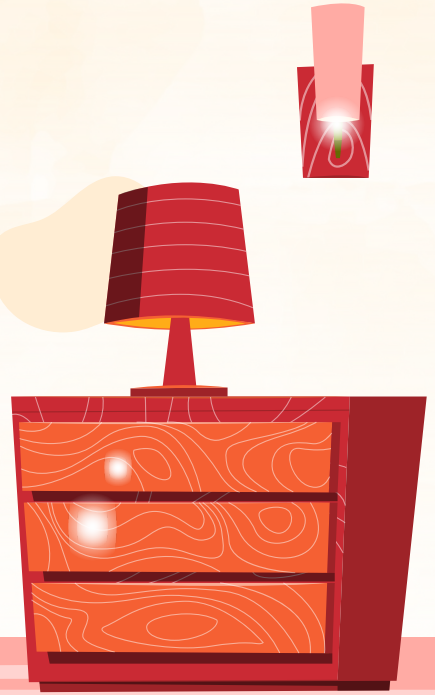


provides easy-to-use sentiment analysis functionality

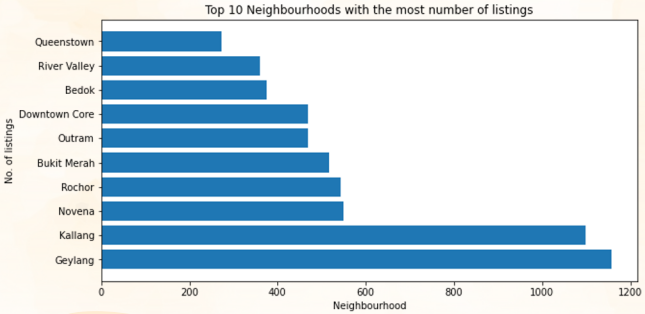
06

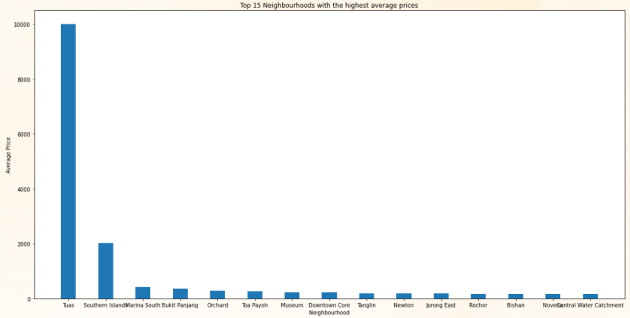
DATA PREPROCESSING

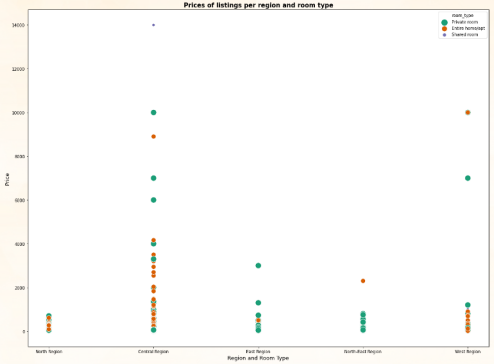




EXPLORATORY DATA ANALYSIS







Geylang and Kallang has the most number of listings.

Queenstown being the least number of listings.

Most neighbourhoods have around the same average price range.

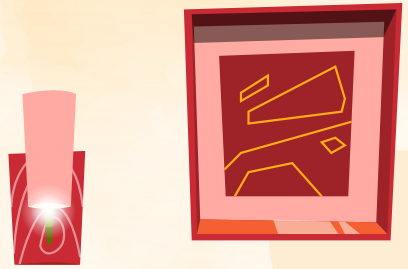
Tuas with an extremely high amount of price.

North region has the smallest range

West Region has the most outliers

Central Region has the biggest range

DATA PREPROCESSING



REMOVAL

Remove irrelevant attributes like IDs and URLs and values with less than 90% of total rows.

IMPUTATION

Rows with only a less than 10% of missing values will be imputed depending on the use case.

OUTLIERS

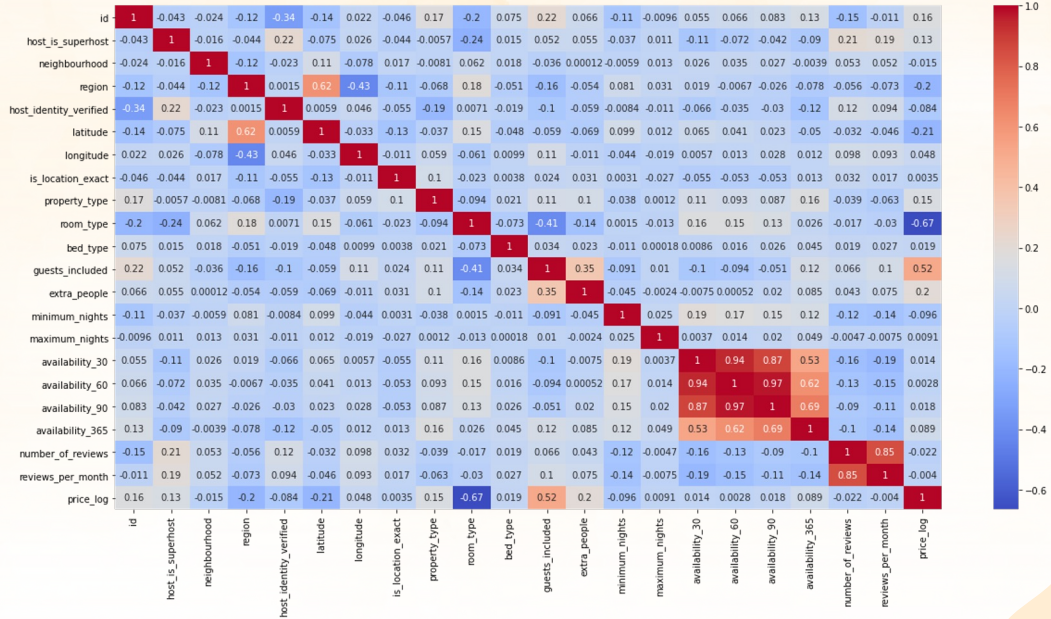
Remove outliers in prices as most listings are around $100-$400.

CATEGORICAL

Apply label encoding to all categorical variables.

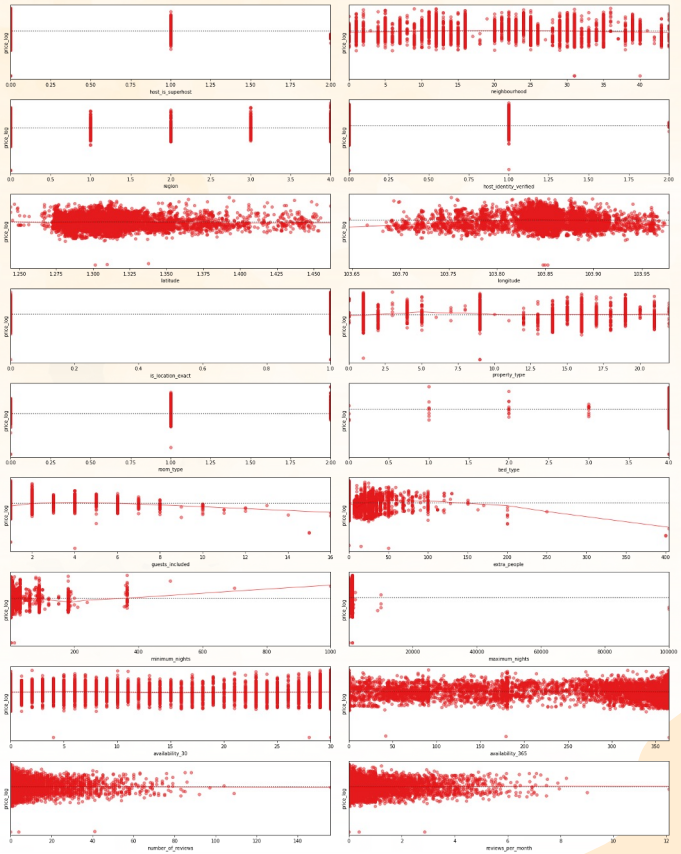
CORRELATION MATRIX

- No values with a high correlation with price
- Top 3 correlated values are
  - guests_included (0.52)
  - extra_people (0.2)
  - property_type (0.15)
- Multicollinearity exists between variables hence reduced to “availability_30” and “availability_365”.



RESIDUAL PLOTS

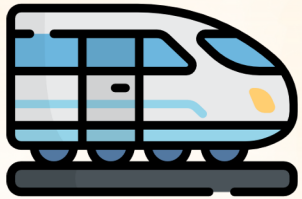
- Detect outliers and non-linearity in data.
- Most variables do not have a red horizontal line which suggest non-linearity.
- There is no independent variables with a strong relationship to price.



FEATURE ENGINEERING

K-Means CLustering

Group listings with MRT exits and Tourists Attractions nearby



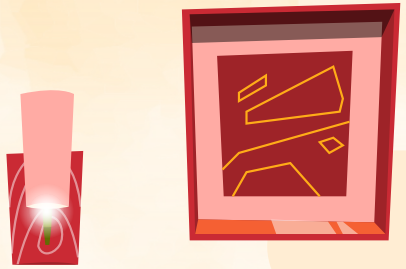
Distance Calculation

Distances between MRT and listing; number of nearby malls and attractions

Sentiment Analysis

Average sentiment score for review of each listing

K-MEANS CLUSTERING

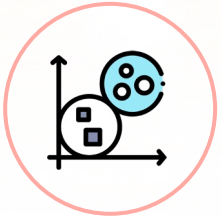


STEP #1

Merge coordinates of MRT exits and tourist attractions

STEP #2

Perform elbow method to obtain the optimal number of cluster = 5.



STEP #3

Perform K-Means clustering to the merged dataset

STEP #4

Assign cluster group to each listing and remove MRT and Tourist attractions rows

K-MEANS CLUSTERING

	latitude	longitude	type
0	1.44255	103.79580	0
1	1.33235	103.78521	0
2	1.44246	103.79667	0
3	1.34541	103.95712	0
4	1.34567	103.95963	0

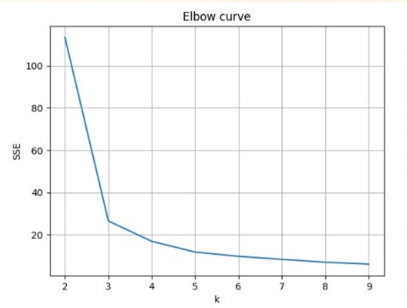
LISTINGS

	latitude	longitude	type
0	1.338511	103.870941	1
1	1.338583	103.870508	1
2	1.319235	103.861904	1
3	1.331067	103.868704	1
4	1.331148	103.869333	1

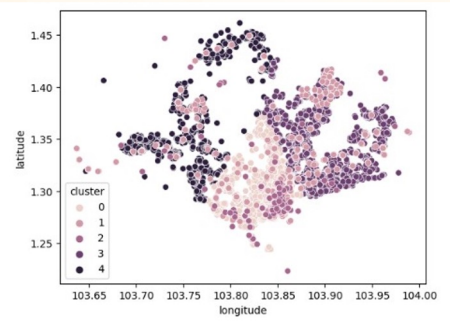
MRT EXITS

	latitude	longitude	type
0	1.283510	103.844350	2
1	1.280940	103.847630	2
2	1.310070	103.899420	2
3	1.277219	103.837336	2
4	1.275490	103.841420	2

TOURIST ATTRACTIONS



ELBOW METHOD



CLUSTER GROUP

DISTANCE CALCULATION

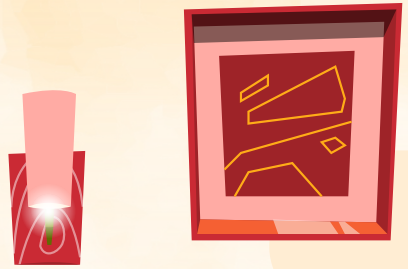
Calculating distances to places of interest using Haversine Formula

Finding distance of nearest mrt exit to listing

Finding number of attractions within 5 km radius

Finding number of malls within 2 km radius

SENTIMENT ANALYSIS



STEP #1

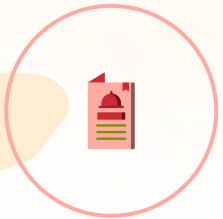
Translate review of each listing into English

STEP #2

Obtain sentiment scores of each reviews through TextBlob

STEP #3

Grouped each review by their listing id and average sentiment scores

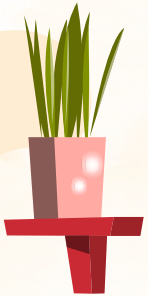


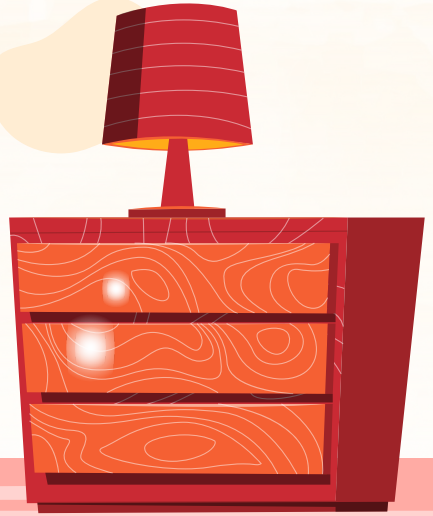
STEP #4

Concatenate the average sentiment score to each listing

07

FINAL REGRESSION MODEL





MODEL PREPARATION

SCALING

Due to the different ranges of numerical values of the various variables, we performed scaling to ensure values are on a comparable scale

```
from sklearn.preprocessing import MinMaxScaler

# create a scaler object for min-max scaling
scaler = MinMaxScaler()

#host_is_superhost, neighbourhood, region, host_identity_verified
# is_location_exact, property_type, room_type, bed_type,

# select only the numerical columns of the dataframe because its not recommended to scale categorical or binary variables
# num_cols = ['guests_included', 'extra_people','minimum_nights','maximum_nights','availability_30','availability_365','numbe
num_cols = ['minimum_nights','maximum_nights','availability_30','availability_365','number_of_reviews',
            'reviews_per_month','price_log','nearest_mrt','num_attractions','num_malls','sentiment_scores']
listings_numeric = listings_model_data[num_cols]

# scale the numerical columns using min-max scaling
scaled_num = scaler.fit_transform(listings_numeric)

# replace the original numerical columns with the scaled ones
listings_model_data[num_cols] = scaled_num
```

MODEL PREPARATION

K-FOLD CROSS VALIDATION

Perform K-Fold cross validation with # of folds = 5, to have a more accurate evaluation of the generalization performance of the model.

```
from sklearn.model_selection import KFold
kfold_cv=KFold(n_splits=5, random_state=42, shuffle=True)
for train_index, test_index in kfold_cv.split(listings_variables,listings_price):
    X_train, X_test = listings_variables.iloc[train_index], listings_variables.iloc[test_index]
    y_train, y_test = listings_price[train_index], listings_price[test_index]
```

MODEL USED

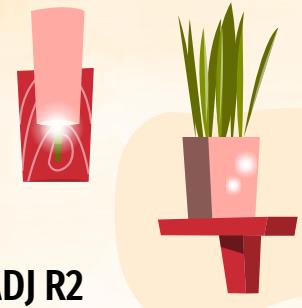
REGRESSION MODELS

- Linear Regression
- Lasso Regression
- Elastic Net (Enet) Regression
- Ridge Regression
- Decision Trees Regresion
- K-Nearest Neighbours Regression

BAGGING & BOOSTING

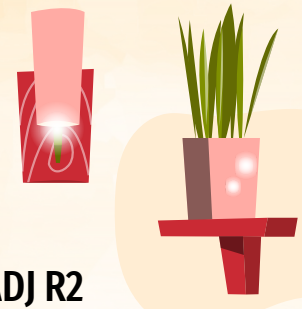
- Standard Bagging Method (Decision Trees Regressor)
- AdaBoost
- XGBoost

REGRESSION MODELS



MODEL	RMSE	MAE	R2	ADJ R2
Linear Regression	0.07	0.05	0.56	0.55
Lasso Regression	0.08	0.06	0.42	0.42
Elastic Net Regression	0.07	0.06	0.49	0.52
Ridge Regression	0.07	0.05	0.52	0.53
DT Regression	0.07	0.06	0.54	0.54
KNN Regression	**0.07**	**0.05**	**0.54**	**0.54**

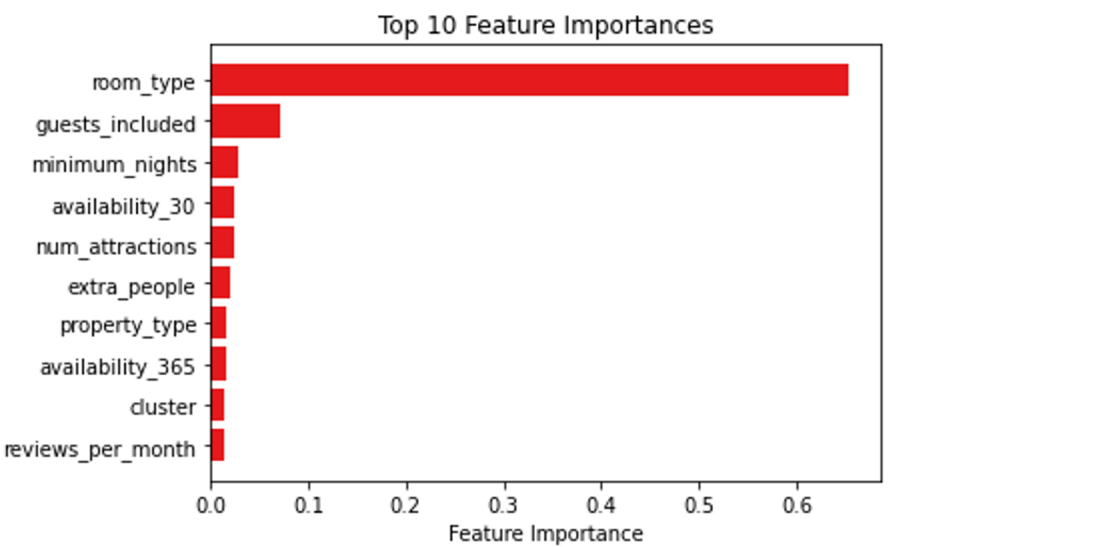
BAGGING & BOOSTING



MODEL	RMSE	MAE	R2	ADJ R2
Standard Bagging Method	0.07	0.05	0.55	0.54
AdaBoost	0.07	0.05	0.55	0.54
XGBoost(PCA)	0.07	0.05	0.55	0.55
XGBoost	0.06	0.04	0.70	0.70

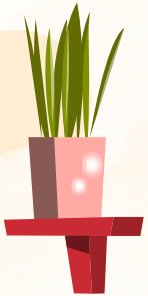
FEATURE IMPORTANCE

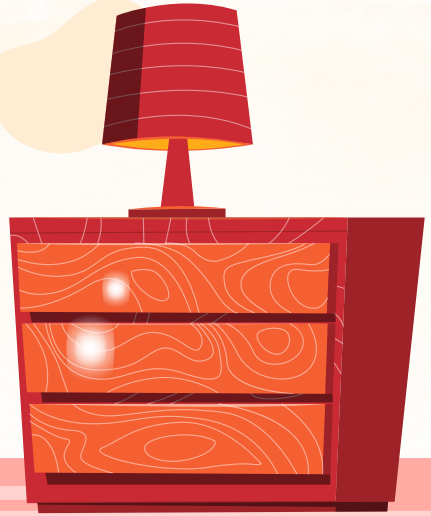
- Provides insight into which aspects of the listing contribute to the price.
- Feature importance plot conducted on XGBoost, best performing model
- The most important feature is “room_type”



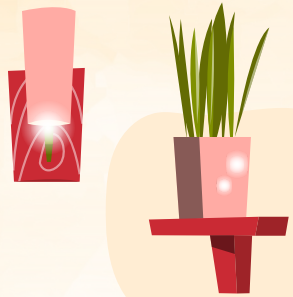
08

CONCLUSION AND FUTURE WORKS





CONCLUSION



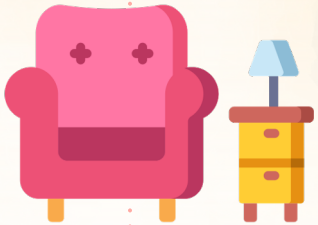
- No single factor can determine price correctly
- Combination of multiple factors correlate with price and are responsible for fluctuations
- Aggregate models perform better than individual models
- **XGBoost** proved to be the best model for our pricing problem

LIMITATIONS

Lack of Data

Have more listings in our dataset as we are only working with <8300 rows.

Feature analysis



Further feature engineering by understanding whether having certain amenities would affect price

Price fluctuation

Takes into account the price fluctuation of airbnb listings due to demand, time of booking to have a more stable model.

MOVING FORWARD

Data

Obtain more data on Airbnb Listings in Singapore

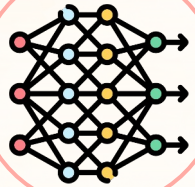
Feature

Look into amenities of each listing to determine any possible correlation with price

Time-series

Identify seasonal patterns in the data, to adjust the predictions for seasonal variations

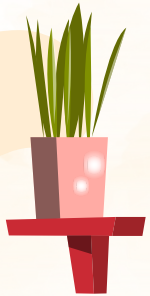
Deep Learning

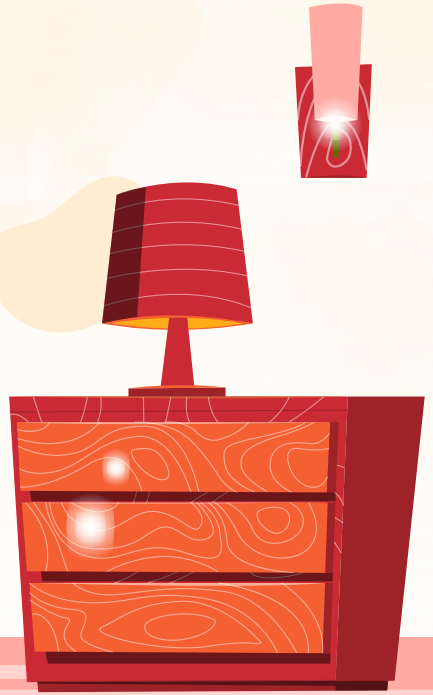


Utilise deep learning to better make sense of non-linearity in data and explore computer vision for listing images

09

REFERENCES





REFERENCES

- ks4s. (2019, July 26). Airbnb Singapore listing. Kaggle. Retrieved February 17, 2023, from https://www.kaggle.com/datasets/sarvasaga/airbnb-singapore-listing
- Rezazadeh Kalehbasti, P., Nikolenko, L., Rezaei, H. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2021. Lecture Notes in Computer Science(), vol 12844. Springer, Cham. https://doi-org.libproxy.smu.edu.sg/10.1007/978-3-030-84060-0_11
- Gibbs, Chris & Guttentag, Daniel & Gretzel, Ulrike & Yao, Lan & Morton, Jym. (2017). Use of dynamic pricing strategies by Airbnb hosts. International Journal of Contemporary Hospitality Management. 30. 00-00. 10.1108/IJCHM-09-2016-0540.
- Carrillo, G. (2020, January 29). Predicting airbnb prices with machine learning and location data. Medium. Retrieved February 17, 2023, from https://towardsdatascience.com/predicting-airbnb-prices-with-machine-learning-and-location-data-5c1e033d0a5a#788c
- Al-Tameemi, F. (2018, November 10). Airbnb listings analysis in Toronto [October 2018]. Medium. Retrieved February 18, 2023, from https://medium.datadriveninvestor.com/airbnb-listings-analysis-in-toronto-october-2018-2a5358bae007
- Karkala, D. (2020). Predictive Price Modeling for Airbnb listings. Predictive price modeling for Airbnb listings. Retrieved February 18, 2023, from https://www.deepakkarkala.com/docs/articles/machine_learning/airbnb_price_modeling/about/index.html
- Duygut, E. (2019, March 4). Airbnb NYC Price Prediction. Kaggle. https://www.kaggle.com/code/duygut/airbnb-nyc-price-prediction

Thank You!