MEF UNIVERSITY

PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES: AN APPLICATION TO VACATION RENTAL PROPERTIES

Capstone Project

Oğuz Ay

İSTANBUL, 2021

GCRIS

GCRIS

MEF UNIVERSITY

PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES: AN APPLICATION TO VACATION RENTAL PROPERTIES

Capstone Project

Oğuz Ay

Advisor: Asst. Prof. Dr. Hande Küçükaydın

İSTANBUL, 2021

GCRIS

MEF UNIVERSITY

Name of the project: Price Prediction Using Machine Learning Techniques: An Application To Vacation Rental Properties
Name/Last Name of the Student: Oğuz Ay
Date of Thesis Defense: 25/01/2021

I hereby state that the graduation project prepared by Oğuz Ay has been completed under my supervision. I accept this work as a "Graduation Project".

25/01/2021
Asst. Prof. Dr. Hande Küçükaydın

I hereby state that I have examined this graduation project by Oğuz Ay which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

25/01/2021

Prof. Dr. Özgür Özlük

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Hande Dr. Küçükaydın	……..
2. Prof. Dr. Özgür Özlük	……..

GCRIS

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Oğuz Ay	25/01/2021	

GCRIS

EXECUTIVE SUMMARY

PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES: AN APPLICATION TO VACATION RENTAL PROPERTIES

Oğuz Ay

Advisor: Asst. Prof. Dr. Hande Küçükaydın

JANUARY, 2021, 23 pages

Pricing is a subjective process that highly depends on person. There is no general rule to price a house. That is why there is both overpriced and underpriced rental houses in rental listings in websites such as AirBnB. In order to reduce the effect of subjective pricing, a general machine learning model is built in this project to make more objective price predictions.

In the literature, there are different machine learning models to make numeric predictions. Physical features of houses are used as an input to make inferences about the price of a house. These machine learning models can identify the relations between features and the price and make the predictions with respect to features of a new listing house that has not been priced before.

In this project, six different machine learning models are developed. These are linear regression, ridge regression, support vector regressor, random forest regressor, light gradient boosting machine regressor and extreme gradient boosting regressor. The performances of all models are compared, and the best model is selected for hyper-parameter tuning to make more accurate predictions.

Key Words: Price Prediction, Regression, Parameter Tuning

ÖZET

MAKİNA ÖĞRENİMİ TEKNİKLERİYLE FİYAT TAHMİNLEME: KİRALIK TATİL MÜLKLERİNE BİR UYGULAMA

Oğuz Ay

Proje Danışmanı: Dr. Öğr. Üyesi Hande Küçükaydın

OCAK, 2021, 23 sayfa

Fiyatlandırma, büyük ölçüde kişiye bağlı olan öznel bir süreçtir. Bir evi fiyatlandırmanın genel bir kuralı bulunmamaktadır. Bu nedenle AirBnB web sitesinde olduğu gibi ev kiralama süreçlerinde hem aşırı hem de düşük fiyatlı kiralık evler bulunmaktadır. Kişiye olan bu bağımlılıkları önlemek için bu projede genel bir makine öğrenimi modeli oluşturularak konut fiyatlarını kestirmek için objektif bir yöntem bulunması hedeflenmiştir.

Literatürde sayısal tahminler yapmak için farklı makine öğrenimi modelleri vardır. Evlerin fiziksel özellikleri, bir evin fiyatı hakkında çıkarımlarda bulunmak için girdi olarak kullanılmaktadır. Bu makine öğrenimi modelleri, konut özellikleri ve fiyatı arasındaki ilişkileri belirleyebilir ve daha önce fiyatlandırılmamış yeni bir evin özelliklerine ilişkin fiyat tahminlerini yapabilir.

Bu projede altı farklı makine öğrenimi modeli geliştirilmiştir. Bunlar doğrusal regresyon, ridge regresyon, destek vektör regresyonu, rassal orman regresyonu, hafif gradyan artırma regresyonu ve aşırı gradyan artırıcı regresyondur. Tüm modellerin performansları kararlaştırılmış ve en iyi model hiper parametre ayarlaması için seçilmiştir.

Anahtar Kelimeler: Fiyat Tahminleme, Regresyon, Parametre Ayarı

TABLE OF CONTENTS

LIST OF FIGURES

GCRIS

LIST OF TABLES

GCRIS

1. INTRODUCTION

Price prediction has been one of widely encountered challenges in various fields. Using past information, forecasting can be made about the unknown or future data such as predicting the price of a property. The aim of this project is to use AirBnB listings of condos in major US cities and to forecast the prices of these listings. For this purpose, a linear regression model is used a benchmark model. Then, various models including ridge regression, support vector machine, random forest regressor, LGBM regressor and XGB regressor are made use of and the performance of each model is compared with the performance of other models and the benchmark model using the mean squared error metric. A listing in AirBnB consists of major condo properties, which are not limited to the number of bedrooms, bathrooms, the location of the property.

House price predictions are realized by comparing the price of a house with similar properties. However, this traditional way of pricing requires experience and knowledge of an agent. This may lead to confliction between agents since each agent might come up with a different price. With the help of machine learning algorithms, more objective prices can be obtained (Zhao et al., 2019). Therefore, this project aims to obtain a machine learning model which can objectively predict the price for AirBnB listings.

There are two basic approaches in price prediction of houses. The first is the prediction of exact prices and the second one is predicting the price range rather than the price itself. The latter approach is actually a classification method (Ma et al., 2020). In this project, the price itself is predicted rather than the range of price.

In the AirBnb dataset, there are different property listings to be used for price prediction. It is known that some features such as house type, number of bedrooms, number of bathrooms, amenities have a vast impact on price (Limsombunchai, 2004). Our dataset includes those features, and the second objective of the project is to determine which of these features in the AirBnB dataset are more important than the others.

2. LITERATURE REVIEW

Pricing a house is a crucial task. It is discussed that there is a positive correlation between the sales of houses and price of the listings (Caplin & Leahy, 2011) This is also valid in rental pricings. If the rental price of a house is high compared to the similar ones, the number of rentals can drop drastically. That is why when making decisions about the price of a house, one should think carefully.

In literature, researchers use different approaches when predicting a house price. Some researchers focus on economic indexes in order to form a time-series regression and predict the prices that move with time (Kilpatrick, 2000). Another approach focuses on predicting the price on the current bid price of houses.

Apart from approaches mentioned, another well-known approach is called Hedonic price method (HPM) (Li and Brown, 1980). In HPM, the focus is on both internal characteristics of the house as well as the external factors. Although being a convenient model, it is hard to use, since it requires a large number of data sets with a large number of different features (Visser et al., 2008). It is impracticable to obtain a clean and usable such data in most cases. That makes HPM algorithms hard to implement in every data.

Many studies prove the effect of physical features of houses on their prices. Physical features such as open space have shown to have a positive impact on house prices (Lutzenhiser and Netusil, 2008; Abelsom et al., 2013). Although the positive impact of open space on house prices is obvious, there may be some unexpected correlation between physical characteristics of a house and its price. Visser et al. (2008) state that the number of bedrooms have a negative impact on the price per metersquare of the house. This surprising result brings the question of whether there are any other surprising relations between physical characteristics and the price. In this project, different physical features of a house are evaluated to make price predictions.

It is also widely believed that regression approaches are the best fit for determining the relation between house prices and housing characteristics. Advanced price predictions models made using linear regression in order to make price predictions that give high accuracy (Lu et al. 2017). However, some believe that decision trees have a better performance in understanding this relation (Fan et al., 2006). In this project, both approaches are tested to conclude this uncertainty.

Moreover, for house price predictions, some believe that support vector regressors have a high accuracy (Milunovich, 2020). On the other hand, in the literature, random forest model has been the most effective method on predicting the prices (Sedkaoui and Benaichouba, 2019). One other advantage of random forest model is that it is able to cope with missing values of the features and categorical variables (Antipov and Pokryshevskaya, 2012). Singh et al. (2020) develop three main models: linear regression, random forest and gradient boosting and shows gradient boosting performs the best among them. Thus, in this project some of these models such as support vector regressor and random forest regressor are employed to see which one is the best performer for AirBnB dataset.

GCRIS

3. ABOUT THE DATA

3.1. Data Description

For this project, AirBnB listings of condos in major US cities are used. This public dataset consists of nearly 74,100 listings and 27 different features that are used for training. The dataset that is used for testing has been split by the data provider.

Table 1 shows the features of the model. At first glance, there are some important features to predict the price. These are number of bathrooms, house type, room type as it is previously mentioned. However, some features are irrelevant for price prediction such as listing id, description and name. All these features are engineered in order to be used in model. The data is formed by Mizrahi (2017).

GCRIS

Table 1: Data Features

#	Feature Names
1	id
2	log_price
3	property_type
4	room_type
5	amenities
6	accommodates
7	bathrooms
8	bed_type
9	cancellation_policy
10	cleaning_fee
11	city
12	description
13	first_review
14	host_has_profile_pic
15	host_identity_verified
16	host_response_rate
17	host_since
18	instant_bookable
19	last_review
20	latitude
21	longitude
22	name
23	neighbourhood
24	number_of_reviews
25	review_scores_rating
26	thumbnail_url
27	zipcode
28	bedrooms
29	beds

3.2. Data Preparation

One of the most important part of the project is to prepare the data for training purposes. The feature 'id' is a unique indicator of the listing and it has no predictive effect on the price. Therefore, it is removed from the dataset.

Null values can be misleading when using in machine learning models and must be dealt with prior to building the model. Table 2 shows the number of null values for each feature.

Table 2: Number of Null Values in each Feature

Features	Number of Null Values
id	0
log_price	0
property_type	0
room_type	0
amenities	0
accommodates	0
bathrooms	200
bed_type	0
cancellation_policy	0
cleaning_fee	0
city	0
description	0
first_review	15864
host_has_profile_pic	188
host_identity_verified	188
host_response_rate	18299
host_since	188
instant_bookable	0
last_review	15827
latitude	0
longitude	0
name	0
neighbourhood	6872
number_of_reviews	0
review_scores_rating	16722
thumbnail_url	8216
zipcode	966
bedrooms	91
beds	131

The first set of features include number of bathrooms, bedrooms and beds. A zero value for these features indicate that there is no such room in the property. Thus, these variables are set equal to zero in the model. Information about host such as whether the host

has a profile picture or whether the host has verified his identity is important for predictive purposes, since it has reassuring impact on people. Thus, the null values in 'host_has_profile_pic' and 'host_identity_verified' are assumed to be zero. These features have two values: *t* or *f*. These stand for true and false, respectively. They are also converted to 1 and 0 to make them numeric variables, i.e. 1 is used for true and 0 for false. After these operations, the number of null variables become zero. The feature 'cleaning_fee' also has two values: True or False. These variables are also converted to 1 or 0. 'neighbourhood' feature may have an impact on price. So, this feature is also made use of in the model. The rest of the null variables are taken as 'Unknown'. For 'review_scores_rating' and 'host_response_rate' features, null variables are taken as zero. After cleaning the data, all null values are converted to numeric values.

Another important task in building a model is to correct the data types of the features. For this purpose, data types of the variables have to be checked. As shown in Table 3, there is no need to make conversions for the data types since numeric variables such as number of bathrooms, cleaning fee and number of bedrooms are integer or float and text variables such as property type, room type and bed type are object.

GCRIS

Table 3: Data Types

#	Features	Data Type
0	id	integer
1	log_price	float
2	property_type	object
3	room_type	object
4	amenities	object
5	accommodates	integer
6	bathrooms	float
7	bed_type	object
8	cancellation_policy	object
9	cleaning_fee	integer
10	city	object
11	description	object
12	first_review	object
13	host_has_profile_pic	integer
14	host_identity_verified	integer
15	host_response_rate	integer
16	host_since	object
17	instant_bookable	integer
18	last_review	object
19	latitude	float
20	longitude	float
21	name	object
22	neighbourhood	object
23	number_of_reviews	integer
24	review_scores_rating	float
25	thumbnail_url	object
26	zipcode	object
27	bedrooms	float
28	beds	float

3.3. Exploratory Data Analysis

Understanding the price distribution is important since it is the objective to predict. Figure 1 shows the distribution of the logarithm of prices in all listings. As the graph shows, the majority of the prices lies between 4 and 5, which indicate that most of the prices range from $55 to $150. It has a normal like distribution.

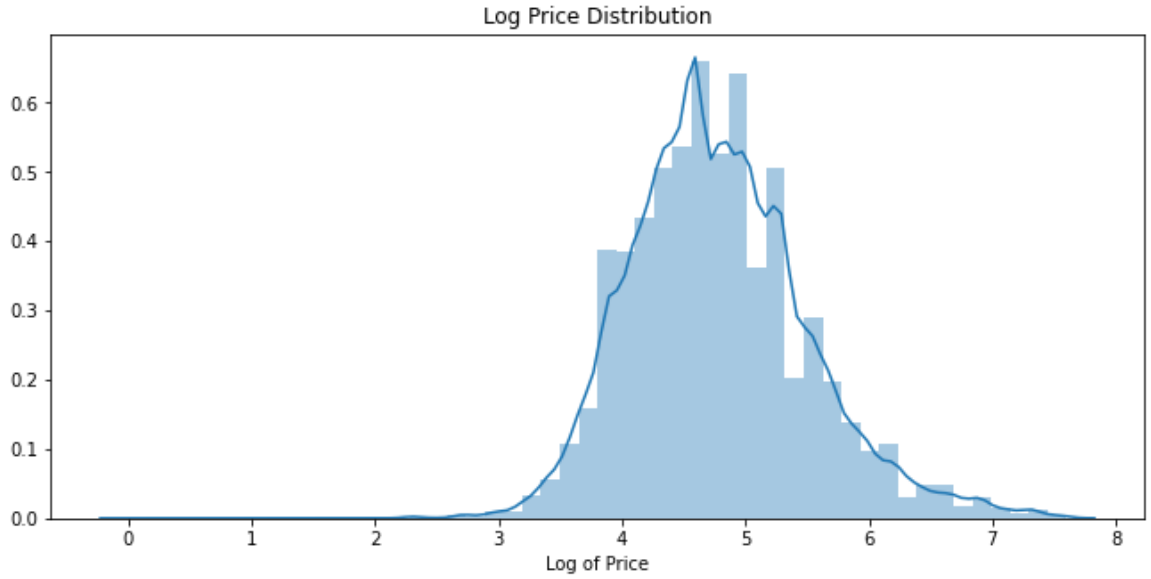

Figure 1: Price Distribution

Room type is a feature showing whether the rent is a home/apartment, a private room, or a shared room and is one of the major indicators for a price. Thus, the log prices for each room type are examined. Figure 2 shows that the median of the prices is around 5 for home/apartment, around 4.5 for private rooms, and around 4 for shared rooms. So, it can easily be concluded that the home/apartment rentals are more expensive than private rooms and private rooms are in turn more expensive than shared rooms. Thus, the feature 'Room Types' is a good indicator of price and hence, it is used in our model. It can also be seen from Figure 2 that the majority of prices of home/apartments lies between 4.8 and 5.5 approximately, for private rooms between 4 and 4.8 and for shared room between 3.8 and 4.5, which shows that the prices of different room types do not overlap.

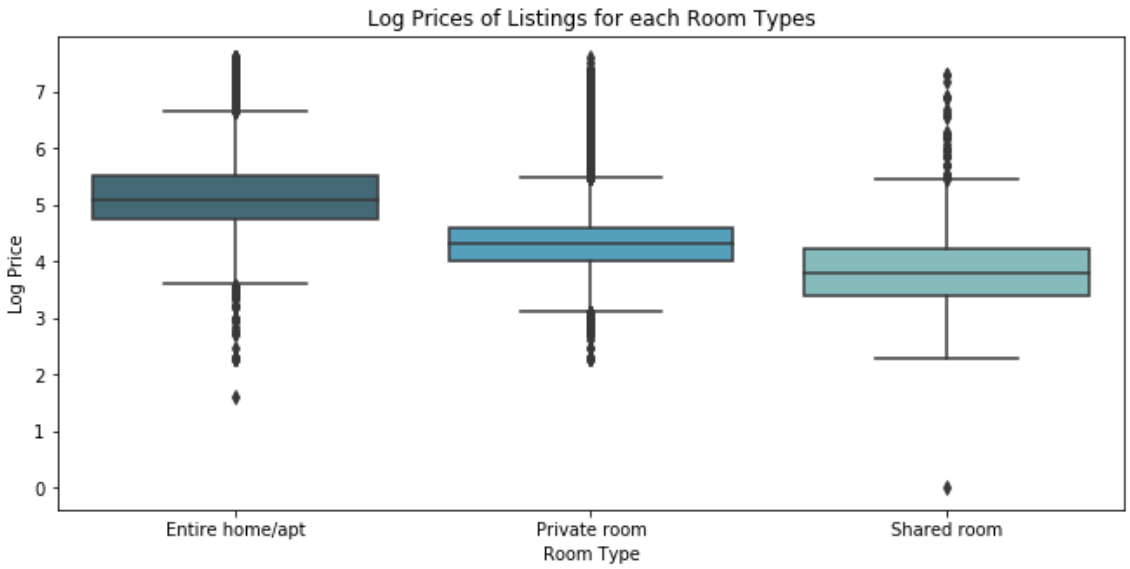

Figure 2: Prices of Room Types

Property type has also a major impact on price. The log prices for each property types are examined. As Figure 3 displays, it can be seen which property types are more expensive than others. In some property types such as timeshare and vacation home, the median of prices are higher than the median of prices for property types such as hostel and dorm. Moreover, some property types such as hut and chalet have less variance than some property types such as earth house and villa. As a result, this shows that property type can be a good indicator to predict the price.

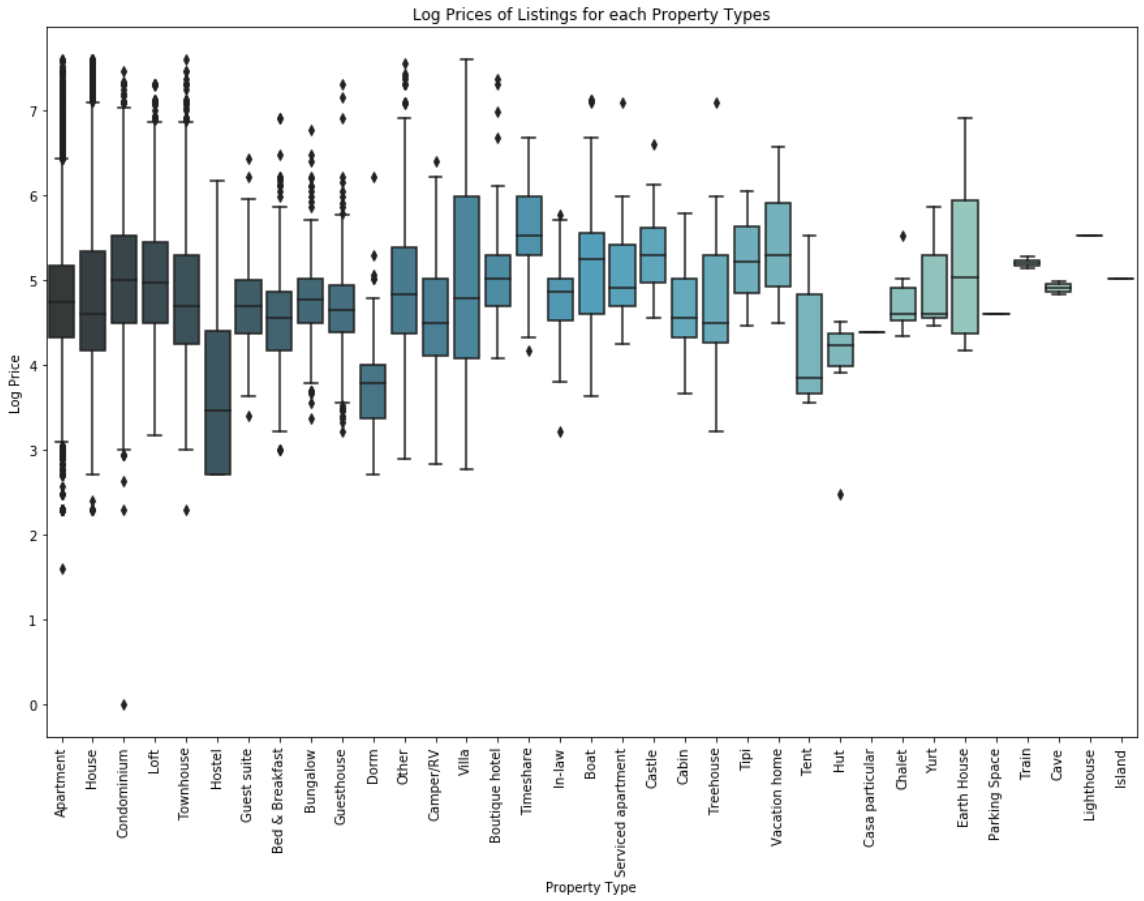

Figure 3: Prices of Room Types

The location of the condos is also an important indicator of price. In order to see this correlation, the price distribution in New York City is displayed by Figure 4. As the color gets closer to yellow, the price increases and as the color gets darker, the price decreases. The listings in Manhattan district are yellow and the listings in Brooklyn district are darker. Obviously, the prices of listings in Manhattan are higher than the prices of listings in Brooklyn. Thus, we conclude that the longitude and latitude which shows the location of the property are also a major indicator of price.

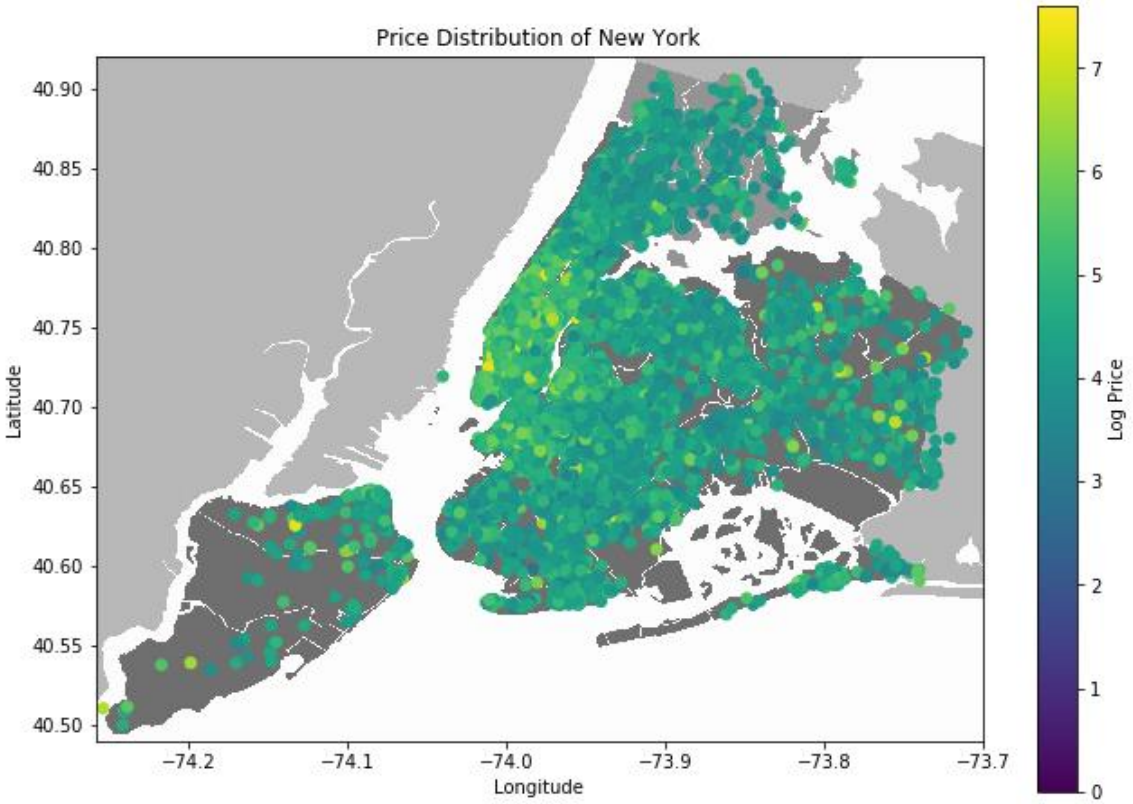

Figure 4: Price Distribution in New York

Next we try to figure out if there is a price trend for property types in each city. To this end, the average prices of each property type are shown for various cities in Table 4. For example the average price for listing of apartments in Boston is 167.72 USD. As it can be observed, there is a boutique hotel trend in New York City which can be understood from high prices. However, for more calm cities like Washington, the trend is to rent loft condos.

Table 4: Average Prices ($) of Cities for each Property Type

Property Type \ City	Boston	Chicago	Washington	Los Angeles	New York City	San Francisco
Apartment	167.72	122.76	199.84	124.10	139.85	215.80
Bed & Breakfast	105.71	111.00	143.82	97.25	134.02	150.20
Boat	358.80	367.67	82.67	181.10	146.00	404.00
Boutique hotel	149.00	215.50	142.38	104.00	347.81	202.14
Bungalow	-	99.33	172.50	133.03	144.17	227.00
Cabin	-	-	110.00	111.18	176.67	154.71
Camper/RV	-	-	-	129.32	-	160.40
Casa particular	-	-	-	80.00	-	-
Castle	-	-	199.00	283.75	175.00	228.33
Cave	-	-	-	125.00	-	147.00
Chalet	-	-	-	105.67	150.00	-
Condominium	182.66	149.78	286.35	150.34	215.97	276.10
Dorm	65.00	50.33	43.95	42.10	73.00	84.33
Earth House	-	-	-	383.33	275.00	-
Guest suite	99.88	66.33	71.40	104.37	140.15	150.41
Guesthouse	154.50	163.15	122.80	123.46	100.34	159.21
Hostel	128.25	60.00	36.00	37.39	76.75	79.50
House	139.66	142.53	238.24	190.23	131.20	241.31
Hut	-	-	-	63.63	-	-
In-law	80.75	45.00	101.50	77.33	116.00	152.45
Island	-	-	-	150.00	-	-
Lighthouse	-	-	-	250.00	-	-
Loft	203.71	168.87	303.93	183.53	207.03	273.78
Other	203.75	303.21	242.31	197.06	196.07	266.42
Parking Space	-	-	-	100.00	-	-
Serviced apartment	129.00	-	135.00	143.80	225.63	1,200.00
Tent	-	-	-	72.38	250.00	98.00
Timeshare	265.00	-	-	-	328.59	288.40
Tipi	-	-	-	232.33	-	-
Townhouse	197.80	165.89	238.12	141.52	190.40	195.72
Train	-	-	170.00	195.00	-	-
Treehouse	-	-	25.00	368.50	-	227.50
Vacation home	-	389.00	-	266.00	265.67	-
Villa	125.00	33.00	160.00	424.55	130.85	99.00
Yurt	-	-	-	160.67	95.00	206.00

4. METHODOLOGY

First, a linear regression model is built as a benchmark. Then, several models such as ridge regression, support vector regressor, random forest regressor, LGBM regressor and XGB regressor are developed. The performance of each model is compared with each other and with the linear regressor benchmark model using the mean squared error (MSE) and adjusted *r*-squared values.

4.2. Feature Scaling

For predictive models, feature scaling is realized to make all the features in the same scale. Although this is crucial for most of the models, some of them such as linear regression do not need scaling of the features due to the characteristics of the model. Basically, distance-based models such as *k*-nearest neighbors (*k-NN*), support vector machine (SVM) or algorithms using regularization need these pre-processing steps, whereas algorithms that rely on rules such as decision trees or algorithms that aim to fit the best fitted line such as linear regression do not need scaling. However, in order to compare the results of the different algorithms, min max scaling, standard scaling and normalizer scaling methods are performed for the features. These scaling methods are similar, but they have different approaches. Min max scaler converts the features' scale between 0 and 1, standard scaler converts the features to have a mean of 0 and a standard deviation of 1 and in normalization scaling the aim is to make the sum of squares of each row to be 1. As a result of these three scaling approaches, min max scaled dataset gives a mean squared error of 3.73×10^{19}, whereas standard scaled and normalized dataset provide an MSE of 8.68×10^{23} and 1.42×10^{18}, respectively. This implies that normalization is the best method to perform feature scaling.

4.3. Solution Methods

Linear regression model is a simple model that focuses on fitting a simple linear line between two variables in order to make its predictions.

As it can be concluded from the mean squared errors mentioned before, the error terms for linear regression are high which means that there is no linearity for the overall data. Nevertheless, if the different scaling methods are compared, normalization method outperforms the others, since it has a smaller mean squared error. Algorithms such as linear regression use normalization assumption. Thus, it needs a normalization a priori.

We also devise more sophisticated algorithms to be able to obtain better results. For this purpose, regularization is added to the linear regression by forming a ridge model, the best fitted line is taken in hyperplane by SVR (Sharp, 2020), a bagging model is added by random forest regressor and a boosting model is formed by Light GBM and XGB regressor. They are selected for having different logic and methods to make predictive models.

Firstly, ridge regression model is formed with default parameters. Ridge regression is a regularization method that adds a penalty function to the model aiming to reduce the complexity of the model (Xiaohong et al., 2020). Three scaling methods are priorly performed. Min max scaler which converts the features between 0 and 1 has a mean squared error of 0.223. Then standard scaler is performed which takes the features to have a mean of 0 and a standard deviation of 1. Unlike in min max scaling, all features have a mean of 0. This model has a mean squared error of 0.223 which is the same as in min max scaling. Lastly, normalization is realized. As the default parameters suggest, *l2* normalization is performed. In this method, the sum of squares of each row must be 1. The mean squared error worsens as it takes the value of 0.436.

Secondly, support vector regression is implemented. Support vector regression models form an optimization problem that attempts to find the narrowest area that surrounds all the data points while minimizing the prediction error which is the distance between the predicted and the real outputs (Awad and Khanna, 2015). The mean squared errors for min max scaling, standard scaling and normalizer scaling are 0.199, 0.199 and 0.241 respectively. Like in ridge regression, this model has same error for min max scaling and worsens with normalizer scaling with an error of 0.241.

Thirdly, random forest regression model is formed. It is a tree-based ensemble model with different trees with random sub data sets (Cutler et al., 2011). The mean squared errors with min max scaler, standard scaler and normalizer are calculated as 0.218 for all models. The scaling method is irrelevant for this model, but it gets worse compared to support vector regressor.

Lastly, an LGBM regressor is developed. This is also an ensemble machine learning model that is constructed from decision tree models which are added one at a time and learn from its errors and improve itself (Brownlee, 2020). The best mean squared error obtained for this model is with min max scaling and standard scaling with an error of 0.194. Moreover, an XGB regressor is developed. XGB is also an ensemble machine learning model that is

faster than other ensemble models with the help of parallel computing and with its split-finding and regularization, it gives accurate predictions (Wade, 2020). This model has a similar error of 0.195.

Table 5: Mean Squared Error Performance of Algorithms

	Min Max Scaler	Standard Scaler	Normalizer
Ridge Regression	0.223	0.223	0.436
Support Vector Regressor	0.199	0.199	0.241
Random Forest Regressor	0.218	0.218	0.218
LGBM Regressor	0.194	0.194	0.198
XGB Regressor	0.195	0.195	0.200

Table 5 shows the summary of different models with different scaling. As it is discussed, the best performing model is LGBM regressor with min max scaling. These results are obtained by using cross validation with a cross-validation generator of 5. Therefore, no train-test split is performed to the dataset.

One other important metric for regression models is adjusted *r*-squared scores. As Table 6 shows the summary, LGBM regressor with min max scaling has an adjusted r squared of 0.621. Thus, LGBM regressor with min max scaling is the best performing model.

Table 6: Adjusted R Squared Performance of Algorithms

	Min Max Scaler	Standard Scaler	Normalizer
Ridge Regression	0.565	0.565	0.152
SVR	0.613	0.612	0.531
Random Forest Regressor	0.576	0.576	0.575
LGBM Regressor	0.621	0.621	0.614
XGB Regressor	0.620	0.620	0.609

4.4. Parameter Tuning of Algorithms

All algorithms are developed using the default parameter settings. Then a grid search model is built using different parameters. Since the algorithm cross-validates, all data is taken as input for the algorithm. Both train and test dataset is given to cross-validation one by one in order to make comparison between train and test dataset. The train dataset provides

a mean squared error of 0.1849 and test dataset gives a mean squared error of 0.1980 and the train dataset provides an adjusted r squared score of 0.6391 and test dataset gives an adjusted r squared score of 0.6200.

As both the MSE and adjusted r squared values show, the best predictive model is the LGBM regressor after tuning. As Table 7 implies, the MSE with default parameters is 0.20. In the best performing model, the boosting type is 'gbdt' as the default value. The regularization α of 0.2 worsens the error by 0.71×10^{-4} and the regularization λ of 0.1 worsens the error by 0.15×10^{-3}. But their mutual impact improves the error by 3.57×10^{-5}. The regularization alpha is well-known by name *l1* regularization and regularization lambda is also known by name *l2* regularization. The difference between *l1* and *l2* regularization is that *l1* tries to make less important features' weights to be zero, *l2* on the other hand encourages weights to be small but does not force them to be exactly zero. Overall, the tuned model has a mean squared error of 0.19.

Table 7: Best Parameters

Parameter	Default Value	Best Value	MSE Improvement
boosting_type	gbdt	gbdt	-
learning_rate	0.1	0.1	-
reg_alpha	0	0.2	-0.71×10^{-4}
reg_lambda	0	0.1	-0.15×10^{-3}
MSE	0.2	0.19	3.57×10^{-5}

As the best parameters improve the mean squared error, it also has a positive impact on adjusted *r*-square score. With the tuned parameters, adjusted *r*-square improved slightly from 0.621 to 0.622.

4.5. Feature Importance

In order to understand which features have a higher impact on price prediction, a further analysis is required. The model with the minimum error value is taken and the feature importance of the model is observed and sorted in descending order by feature importance attribute of the LGBM model. Table 8 shows the first ten important features.

Table 8: Feature Importance

Importance	Name of Feature
1	number_of_reviews
2	accommodates
3	host_response_rate
4	review_scores_rating
5	bathrooms
6	bedrooms
7	beds
8	Washington
9	San Francisco
10	cleaning_fee

As the result suggests, the feature which affect the price the most is the number of reviews of the post on the website. Generally, people have a tendency for choosing the properties that other people like which can be an indication of a herd psychology.

The other important features to predict the price are properties of home such as accommodates, number of bathrooms, number of bedrooms and number of beds which are predictable. As the number of bedrooms for example increase, so as the living space of the apartment.

The next set of features appear to be the location. The condos in Washington and San Francisco have a significant impact on price as the analysis shows.

As Adyan et al. (2017) suggest, there are three main influences for price of a house. These are physical conditions, concept and location which is consistent with the conclusions observed in this project.

5. CONCLUSION

The price policy of rental listings is ambiguous and subjective. There is no general rule behind the listing prices. The ambition behind this study is to see the best practices behind price prediction models and finding objective price policy. Thus, fair prices of the rentals are questioned throughout the study by making use of several machine learning models.

Linear regression, which is a widely known and simple algorithm, is used as a benchmark model to show the dependency between the house properties and the listing prices of the AirBnB listings. This model assumes a linear relationship. As the results imply, it cannot be said that there is a linear relation between the prices and the features of properties. The next step is to develop a model which predicts the prices better.

Several algorithms including ridge regression, support vector regressor, random forest regressor, LGBM regressor are designed to make a better prediction. After computational experiments, we find out that the best performing algorithm is LGBM regressor after its parameters are tuned.

As a further study, as it is mentioned in the beginning of the project, a classification problem can be considered, where the price range of a property is predicted rather than the price itself. Moreover, this project shows that bagging and boosting models have potential to predict better rather than other models. Thus, other algorithms based on bagging and boosting principles such as AdaBoost, boosted naïve bayes and bagging linear regression can be developed to see if there are any other improvements that can be obtained by changing the model.

REFERENCES

Abelson, P., Joyeux, R., Mahuteau, S. (2013). Modelling House Prices across Sydney. *Australian Economic Review,* 46(3), 269-285. https://doi.org/ 10.1111/j.1467-8462.2013.12013.x

Alfiyatin, A.N., Taufiq, H., Febrita, R.E., Mahmudy, W.F. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications,* 8(10). https://doi.org/10.14569/IJACSA.2017.081042

Antipov, A., Pokryshevskaya E. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications,* 39(2), 1772-1778. DOI: https://doi.org/10.1016/j.eswa.2011.08.077.

Awad, M. & Khanna, R. (2015, January). Support Vector Regression. DOI: 10.1007/978-1-4302-5990-9_4.

Brownlee, J. (2020, June 10). *How to Use StandardScaler and MinMaxScaler Transforms in Python*. Retrieved from https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/

Brownlee, J. (2020, November 25). *How to Develop a Light Gradient Boosted Machine (LightGBM) Ensemble*. Retrieved from https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble

Caplin, A., Leahy, J. (2011). Trading Frictions and House Price Dynamics. *Journal of Money, Credit and Banking,* 43(s2), 283-303. https://doi.org/10.1111/j.1538-4616.2011.00436.x

Cutler, A., Cutler, D., Stevens, J. (2011, January). Random Forests. *Machine Learning*, 45(1), 157-176. DOI: 10.1007/978-1-4419-9326-7_5.

Fan, G., Ong, S.E., Koh, H.C. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies*, 43(12), 2301-2315. Retrieved from https://www.jstor.org/stable/43198334

Hargrave, M. (2020, November 18). *Hedonic Pricing*. Retrieved from https://www.investopedia.com/terms/h/hedonicpricing.asp

Kilpatrick, J.A. (2000). Factors Influencing CBD Land Prices. *Journal of Real Estate*, 28-29

Li, M.M., Brown, H.J. (1980). Micro-Neighborhood Externalities and Hedonic Housing Prices. *Land Economics*, 56(2), 125-141. https://doi.org/10.2307/3145857

Limsombunchai, V., Gan, C., Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences*, 1(3), 193-201. https://doi.org/10.3844/ajassp.2004.193.201

Lu, S., Li, Z., Qin, Z., Yang, X., Goh, S. M. (2017). A hybrid regression technique for house prices prediction. *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore*, 319-323, DOI: 10.1109/IEEM.2017.8289904.

Lutzenhiser, M., Netusil, N.R. (2001). The Effect of Open Spaces on a Home's Sale Price. *Contemporary Economic Policy*, 19(3), 291-298. https://doi.org/ 10.1093/cep/19.3.291

Ma, C., Liu, Z., Cao, Z., Song, W., Zhang, J., Zeng, W. (2020). Cost-Sensitive Deep Forest for Price Prediction. *Pattern Recognition*, 107. https://doi.org/10.1016/j.patcog.2020.107499

Mandot, P. (2017, August 17). *What is LightGBM, How to implement it? How to fine tune the parameters?*. Retrieved from https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

Milunovich, G. (2020). Forecasting Australia's Real House Price Index: A Comparison of Time Series and Machine Learning Methods. *Journal of Forecasting*, 39(7), 1098– 1118. https://doi.org/10.1002/for.2678

Mizrahi, R. (2017). *AirBnB listings in major US cities.* Retrieved from https://www.kaggle.com/rudymizrahi/airbnb-listings-in-major-us-cities-deloitte-ml

Sedkaoui, S., Benaichouba, R. (2019). How Data Analytics Drive Sharing Economy Business Models?. *International Academic Conference, Barcelona*. DOI: 10.20472/IAC.2019.052.057

Sharp, T. (2020, March 3). *An Introduction to Support Vector Regression (SVR)*. Retrieved from https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2

Singh, A., Sharma, A., Dubey, G. (2020). Big data analytics predicting real estate prices. *Int J Syst Assur Eng Manag*, 11, 208–219. DOI: https://doi.org/10.1007/s13198-020-00946-3

Visser, P., Van Dam, F., Hooimeijer, P. (2008). Residential Environment and Spatial Variation In House Prices in the Netherlands. *Tijdschrift voor economische en sociale geografie*, 99(3), 348-360. https://doi.org/10.1111/j.1467-9663.2008.00472.x

Wade, C. (2020, November 10). *Getting Started with XGBoost in scikit-learn*. Retrieved from https://towardsdatascience.com/getting-started-with-xgboost-in-scikit-learn-f69f5f470a97

Xiaohong, S., Huajiang, C., Bagherzadeh, S.A., Shayan, M., Akbari, M. (2020). Statistical estimation the thermal conductivity of MWCNTs-SiO2/Water-EG nanofluid using the ridge regression method. *Physica A: Statistical Mechanics and its Applications*, 537. https://doi.org/10.1016/j.physa.2019.122782

Zhao, Y., Chetty, G., & Tran, D. (2019). Deep Learning with XGBoost for Real Estate Appraisal. *IEEE Symposium Series on Computational Intelligence, SSCI 2019*, 1396-1401. DOI: https://doi.org/10.1109/SSCI44817.2019.9002790

GCRIS