NIIT (Stackroute)

March 2022 – July 2022

AIRBNB HOME RENTAL ANALYSIS

By

Indrajeet Gupta

Vishal Tyagi

Saurav Avhad

Shubham Upadhyay

Under the Guidance of

Dr. Nidhi Chahal

TABLE OF CONTENTS

INTRODUCTION

Airbnb was born in 2007 when two Hosts welcomed three guests to their San Francisco home, and has since grown to over 4 million Hosts who have welcomed more than 1 billion guest arrivals in almost every country across the globe. Every day, Hosts offer unique stays and experiences that make it possible for guests to connect with communities in a more authentic way. Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data - data that can be analyzed and used for security, business decisions, understanding of customers and providers (hosts) behaviour and performance on the platform and much more. United States, France, Spain, Italy, United Kingdom, Australia, Canada, Japan are the top countries where airbnb is used.

Seattle dataset is used for the analysis. This dataset describes the listing activity and metrics in Seattle from 2008 till 2015. Seattle is a seaport city on the West Coast of the United States. It is the seat of King County, Washington. With a 2020 population of 737,015,[2] it is the largest city in both the state of Washington and the Pacific Northwest region of North America.

Our analysis focuses on understanding the Seattle dataset and meet business objective which involves providing different unique insights for hosts (property owner) & customer which can be useful for airbnb to know their end-to-end customer, hosts to take better decision while investing on property in Seattle. Along with performing exploratory data analysis, building a time-series model to forecast property revenue based on past data and identify if there are any trends and seasons when travelers book these accommodations. And lastly, build a Classification model that will predict the ‘property type’ a customer is most likely to select, given a set of input features which will act as property type recommendation.

There are two files which has 3818 rows, 33 columns & 3818 rows, 92 columns respectively.

Tools used for completing the analysis are Excel, Python, SQl, Tableau, HTML.

PROBLEM STATEMENT

1. To get various insights into the home rental business from the perspective of
   - Property owner
   - Customer

2. Identify if there are any trends and seasons when travelers book these accommodations. Analyze and try to forecast the property prices during specific seasons

3. Build a Classification model that will predict the 'property type' a customer is most likely to select, given a set of input features.

DATA DICTIONARY

Manual data dictionary is created with help of different sources to understand the dataset and its features. Listing the top features which are used in this analysis.

Id	text- (unique id allotted for each house)
listing_url	text- (reference URL)
scrape_id	numerical- (common id for scraping)
last_scraped	date- (date of scraping)
picture_url	URL – (that redirects to image as view of the house)
host_id	number – (unique id of each host)
host_name	name of the host (maybe the owner's name)
host_since	date since the person is host
host_response_time	response time of the host mentioned in textual format
host_response_rate	response rate of host mentioned (in number 0 to 100)
host_neighbourhood	names of nearby area where house is located
host_total_listings_count	no of listing host has on the airbnb platform
host_identity_verified	tells if host id is verified on airbnb platform or not
property_type	talks about property type like apartment, bungalow, cabin)
room_type	talks about room type- only 3 types
accommodates	no of people can live in house (1 to 16)
bathrooms	no of bathrooms (0 to 8) also has blank data
bedrooms	no of bathrooms (0 to 10) also has blank data
beds	no of beds in a house (1 to 16)
bed_type	talks about bed type
square_feet	numerical data, talks about sq. feet area of house
price	numerical data, rental price of the house (in dollars)
security_deposit	security deposit needed for renting (in dollars)
guests_included	no of guest's customer can bring in house
minimum_nights	min no of nights customer have to consider for booking
maximum_nights	max no of nights customer can consider for booking
availability_365	the number of available days for rent in the next 365 days.
Number_of_reviews	no of reviews available for that particular house
first_review	date of first review
last_review	date of last review
cancellation_policy	categorical data -tells how strict or simple the policy is

EXPLORATORY DATA ANALYSIS

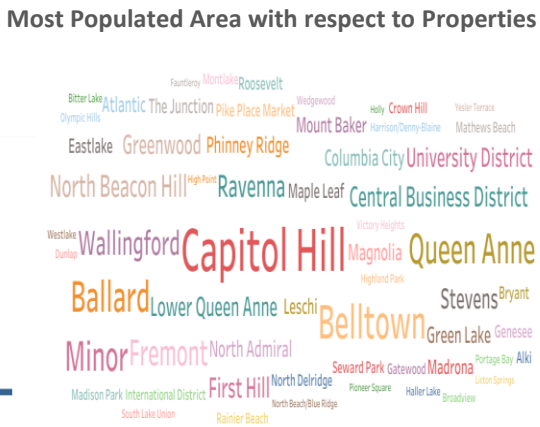
Pre - Processing

After shortlisting relevant features for further analysis, dataset was found with 25 missing rows which is around 0.6% of overall dataset, thus it is dropped.

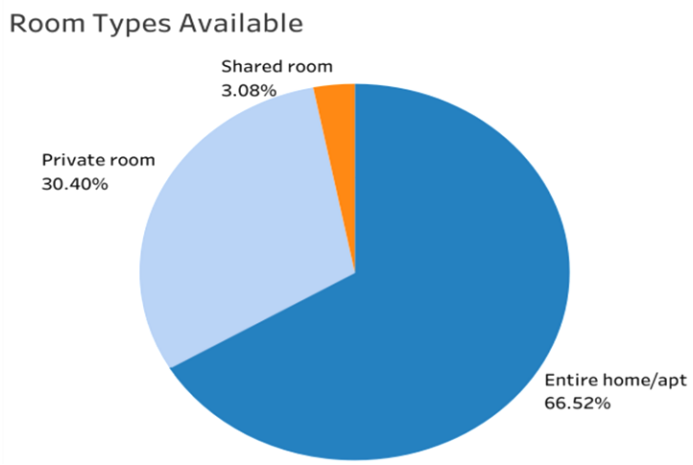
Part 1.1: Analysis with Respect to Customer

Visual 1

Visual 2



Visual 3



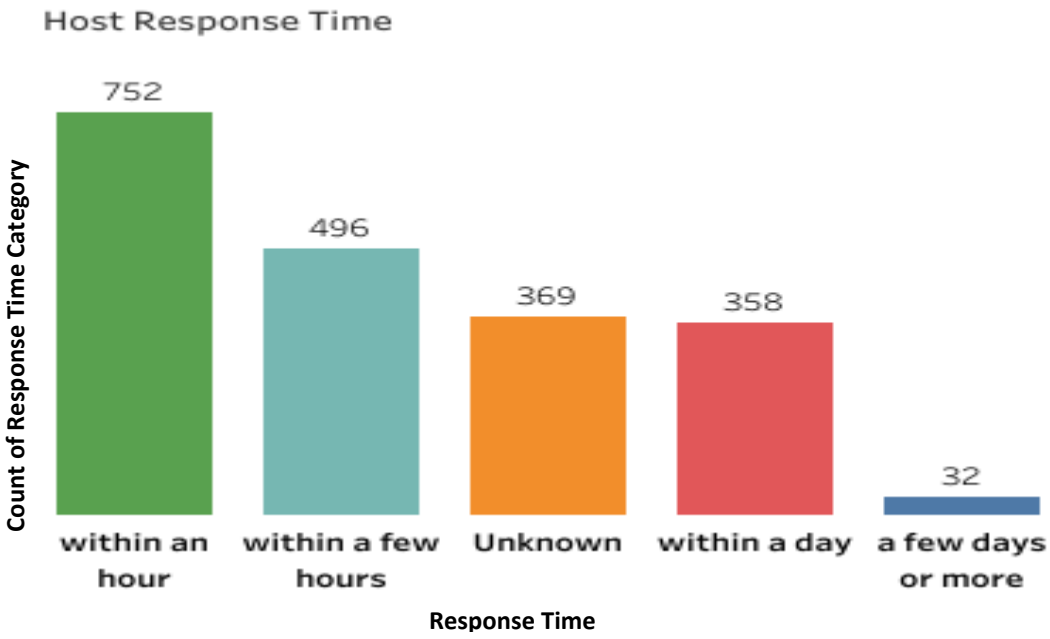
Visual 1: House and apartment are the top most property types available in Seattle.

Visual 2: Capitol Hill is the most populated area with respect to properties.

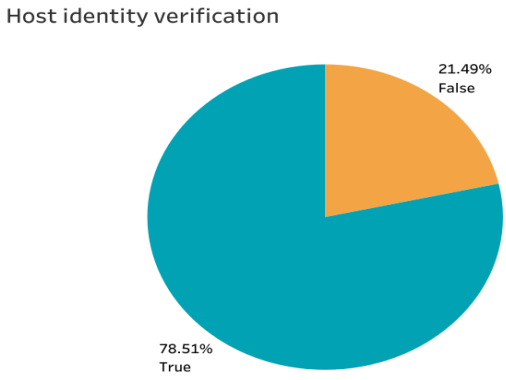
Visual 3: 97% of the data belongs to either Home/Apartment or Private room in Seattle.

Part 1.2: Analysis with Respect to Host

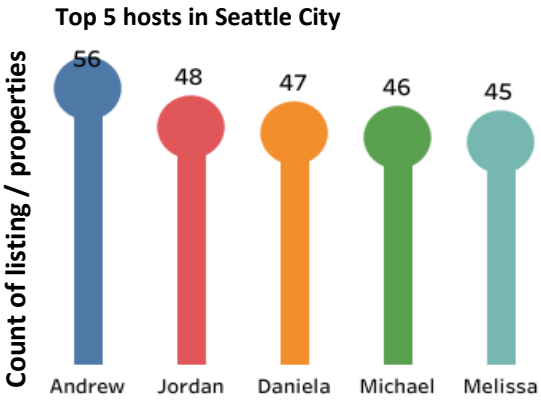
Visual 4



Visual 5



Visual 6

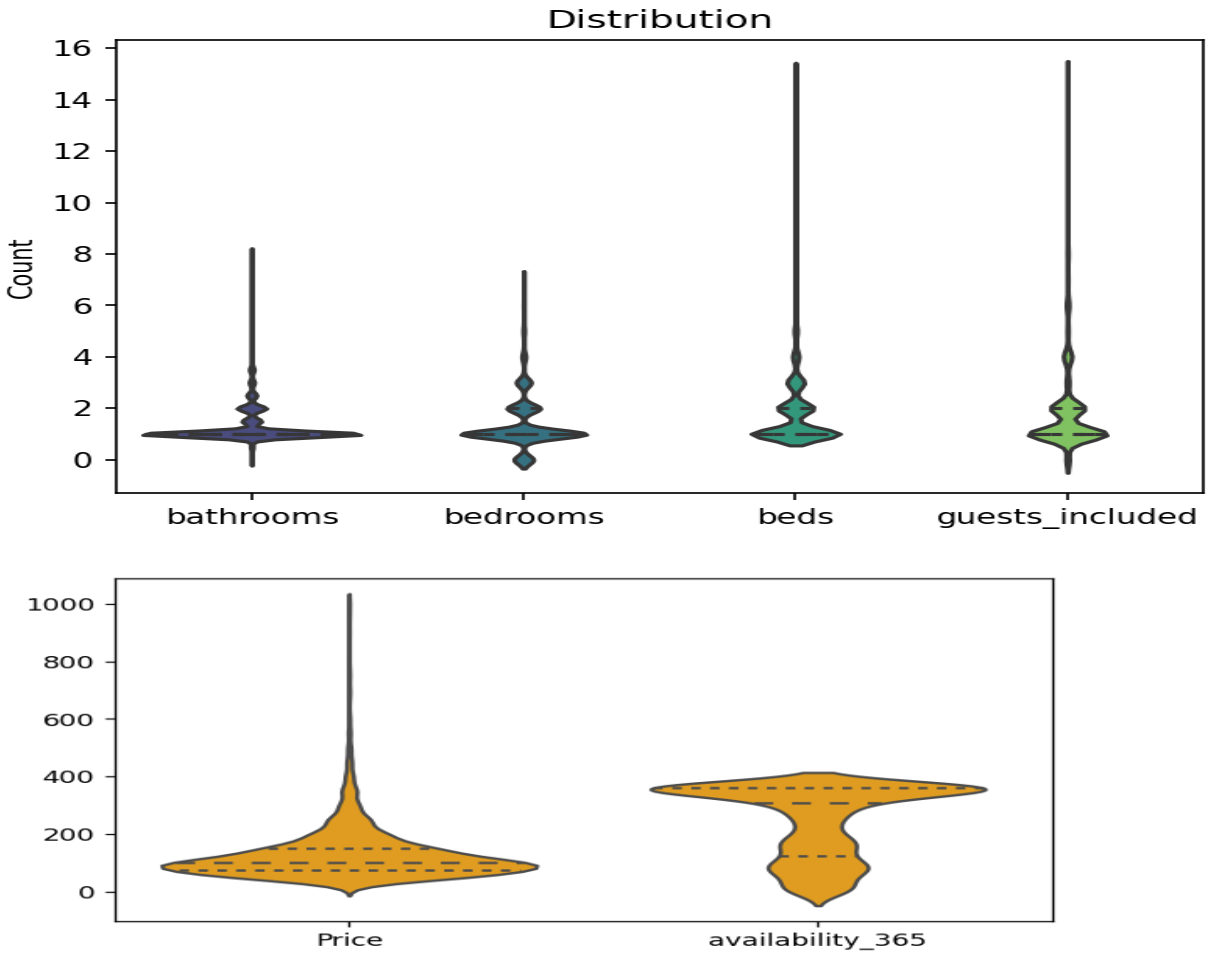


Visual 4: Mostly hosts response within an hour.

Visual 5: More than 75% hosts are verified on the platform.

Visual 6: It shows the top 5 hosts in Seattle who have maximum properties listed on Airbnb.

Distribution of Numerical Data

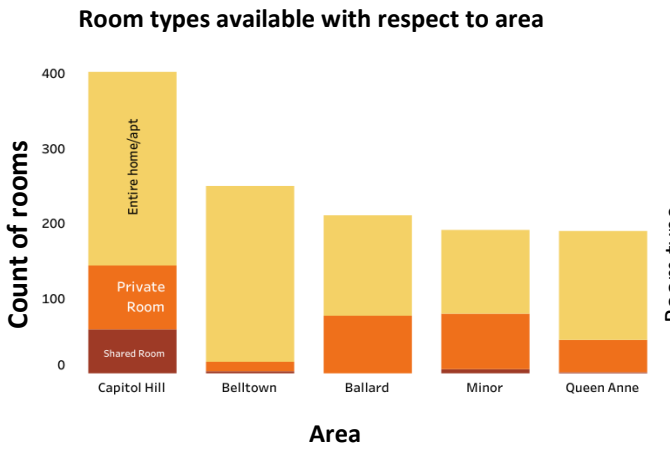


For Numerical features the violin plots are used which gives better insights as it depicts summary statistics as well as density of points.

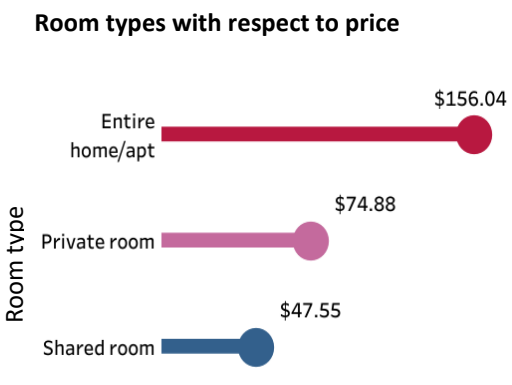
We can see the distribution of respective features and long tail here is called as whisker which means there are some data points which are having large distance than all regular data points.

Part 2.1: Analysis with Respect to Customer

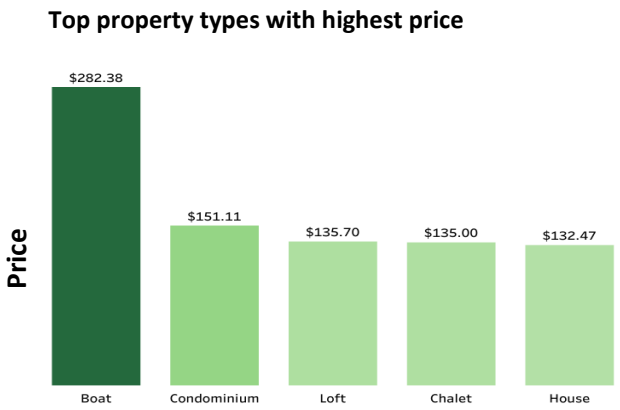
Visual 9



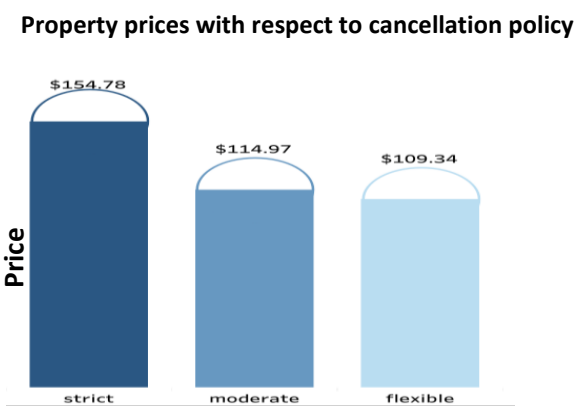
Visual 10



Visual 11

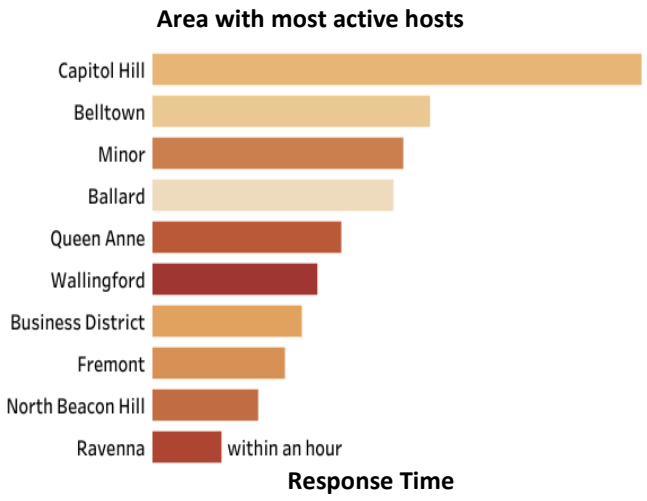


Visual 12

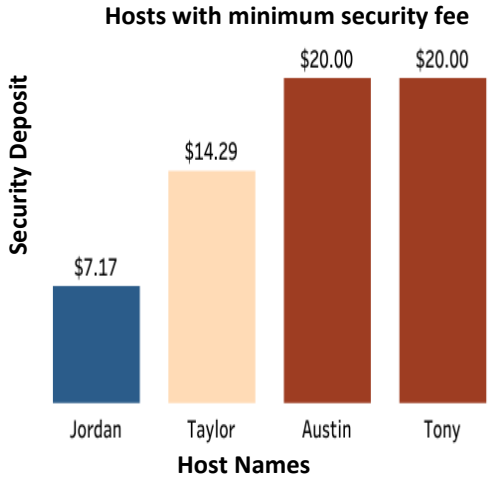


Part 2.2: Analysis with Respect to Host

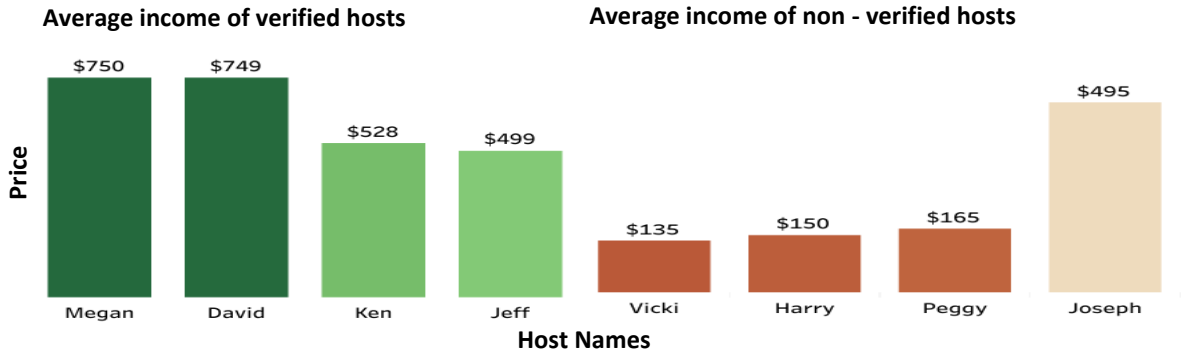
Visual 13



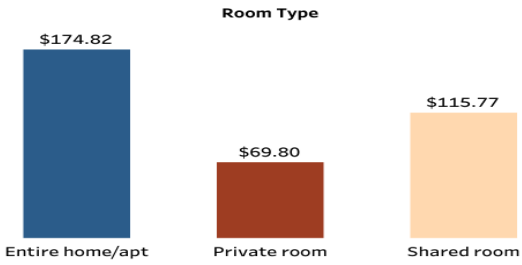
Visual 14



Visual 15

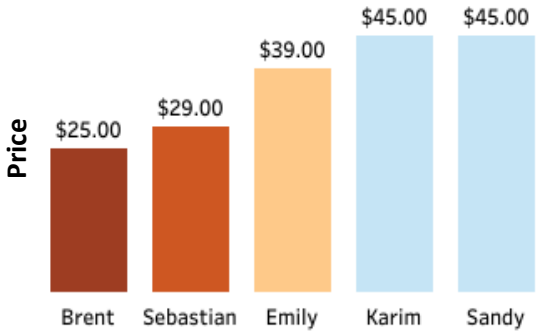


Visual 16

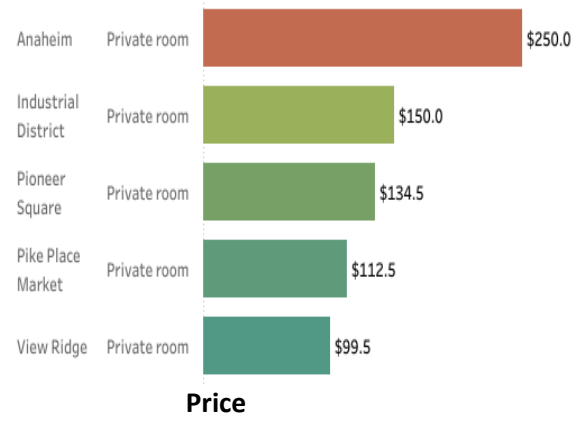


Part 3: Analysis with Respect to Customer

Visual 17



Visual 18

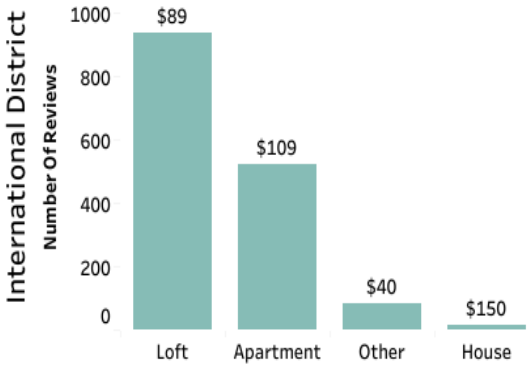


Visual 17: Capitol Hill as most populated area - Presenting top 7 verified hosts offering cheapest shared rooms.

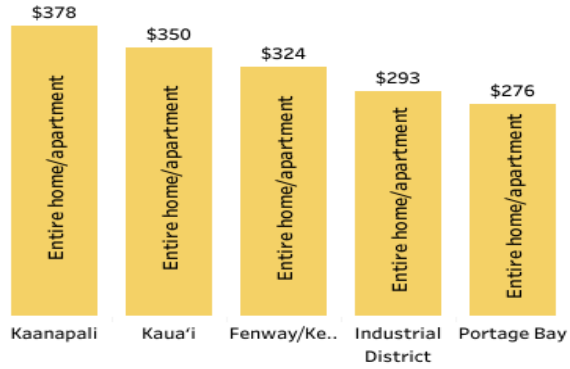
Visual 18: Presenting the top 5 posh areas for private rooms having 1 minimum nights.

Part 3: Analysis with Respect to Hosts

Visual 19



Visual 20



Visual 19: **International District (area) with maximum people engagement** - presenting the most preferred property types along with their prices.

Visual 20: **From all over the Seattle city**- Presenting the top 5 areas where verified hosts are earning maximum due to respective room types.

Exploratory Data Analysis – Insights / Conclusion

❑ For Customers

- ✓ Someone looking for shared room which are cheapest and owner must be verified to avoid any security issues can contact *Brent, Sebastian or Emily* who offers such kind of rooms in Capitol Hill.
- ✓ Person or company planning business trip to Seattle and wants to save money in possible ways & same time wants to rent top quality private rooms can look in the areas like *Anaheim* / *Industrial district* / *Pioneer square* (as they are offering luxurious private rooms & minimum no of nights to be booked is 1).

❑ For Hosts

- ✓ Most of the property type offered by different hosts are houses or apartment but people are showing more response for *Loft* (property type) & specially in the international district area of Seattle so new upcoming hosts can invest in such properties.
- ✓ More than 75% hosts are verified & earning greater than non-verified hosts.
- ✓ Some of the top areas where verified hosts are earning maximum are *Kaanpali, Kaua'I, Fenway/Kenmore, Industrial district, etc & 'Entire Home/apartment's the most preferred room type among them.*

PROPERTY REVENUE FORECAST

OBJECTIVE

Build a time series model that will forecast the future property prices based on the past data.

INTRODUCTION TO TIME SERIES

Time series can be defined as a sequence or series of data points that are ordered in time.

It makes predictions based on past data, and it predicts the future with all the previous observations taken into a consideration. Time-series adds a definite order of dependence between observations.

In time series, any variable that changes as time goes on it is acceptable. It is normal to use a time series to track progress over some period of time. This can be tracked over a short term or long term.

FORECASTING IN DIFFERENT DOMAINS

1. Retail: Sales, profit, and returns on new products
2. Stock Market: Stock Prices
3. Weather: Temperature, humidity etc.
4. Airline: Number of passengers
5. Supply Chain: Cost of raw materials, time to purchase materials, etc.
6. Human Resource: Manpower allocation in various departments
7. Inventory Management: Orders of products

Available Past Data: From October 2008 to December 2015

Features Selected: Host Since (date) and Property Price (revenue)

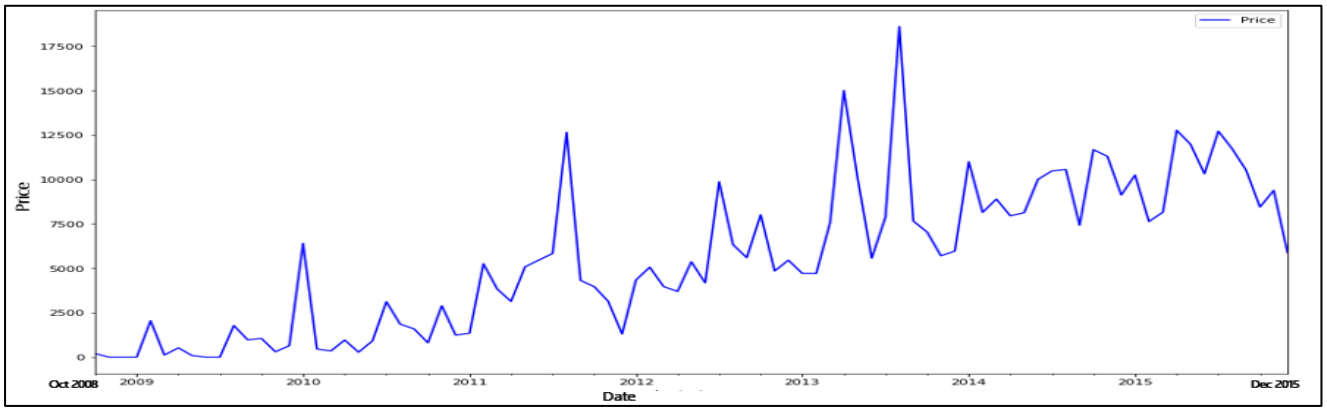


Fig. 3.1: Year Wise Revenue

COMPONENTS OF TIME SERIES

1. **Trend**: This is the long-term increase or decrease in the data. The trend can be increasing or decreasing as well as linear or nonlinear.
2. **Seasonality**: The regular pattern of up and down fluctuations in a time series. It may be a short-term variation occurring due to seasonal factors.

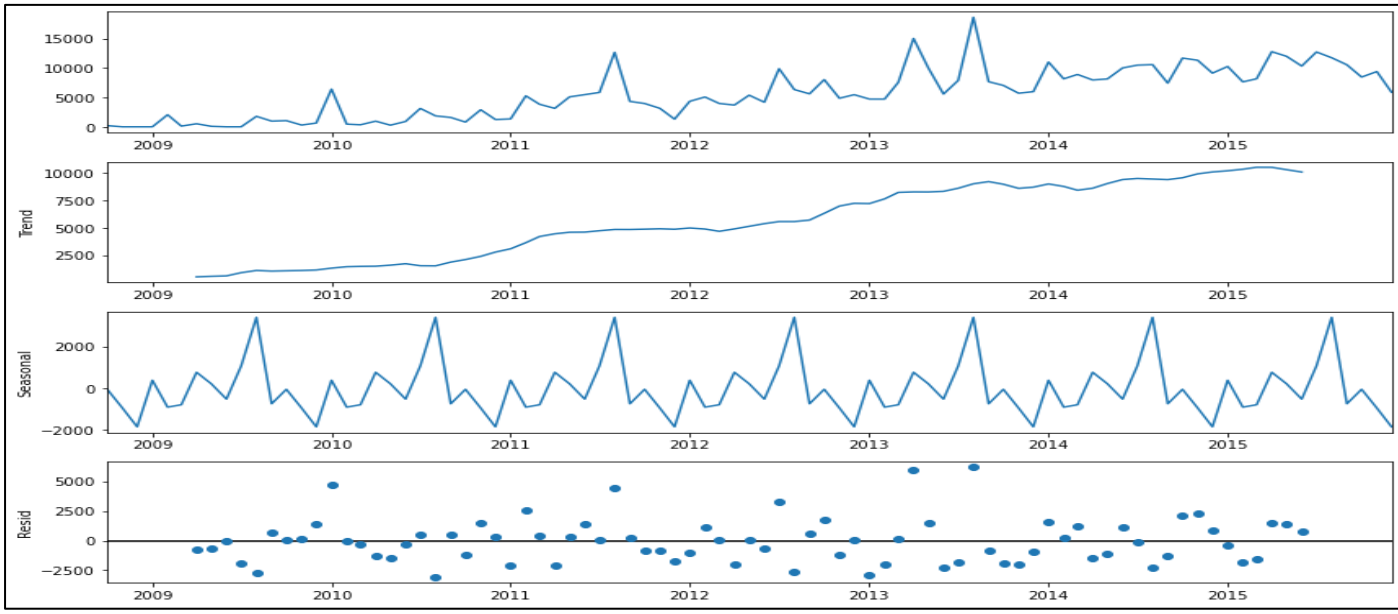


Fig. 3.2: Components of Time Series

- Data shows gradual upward trend and seasonality.
- Each year, during August, bookings increases & thus its showing peak which confirms presence of seasonality.

METHODS OF FORECASTING USING TIME SERIES

1. **Triple Exponential Smoothing (SES):** Holt-Winters Exponential Smoothing is used for forecasting time series data that exhibits both a trend and a seasonal variation.
2. **Seasonal Autoregressive Integrated Moving-Average (SARIMA):** The SARIMA is fitted for a univariate time series with the trend or seasonal components.
3. **Seasonal Auto-Regressive Integrated Moving Average with exogenous factors (SARIMAX):** The SARIMAX model is best suited when data exhibits both trend and seasonality by taking some exogenous variables into consideration.

Model 1: Holt-Winters Method (Triple Smoothening Method)

Model Summary

Dep. Variable:	Price	**No. Observations:**	61
Model:	ExponentialSmoothing	**SSE**	378539433.204
Optimized:	True	**AIC**	986.098
Trend:	Additive	**BIC**	1019.872
Seasonal:	Additive	**AICC**	1002.384
Seasonal Periods:	12	**Date:**	Wed, 27 Jul 2022
Box-Cox:	False	**Time:**	11:40:41
Box-Cox Coeff.:	None		

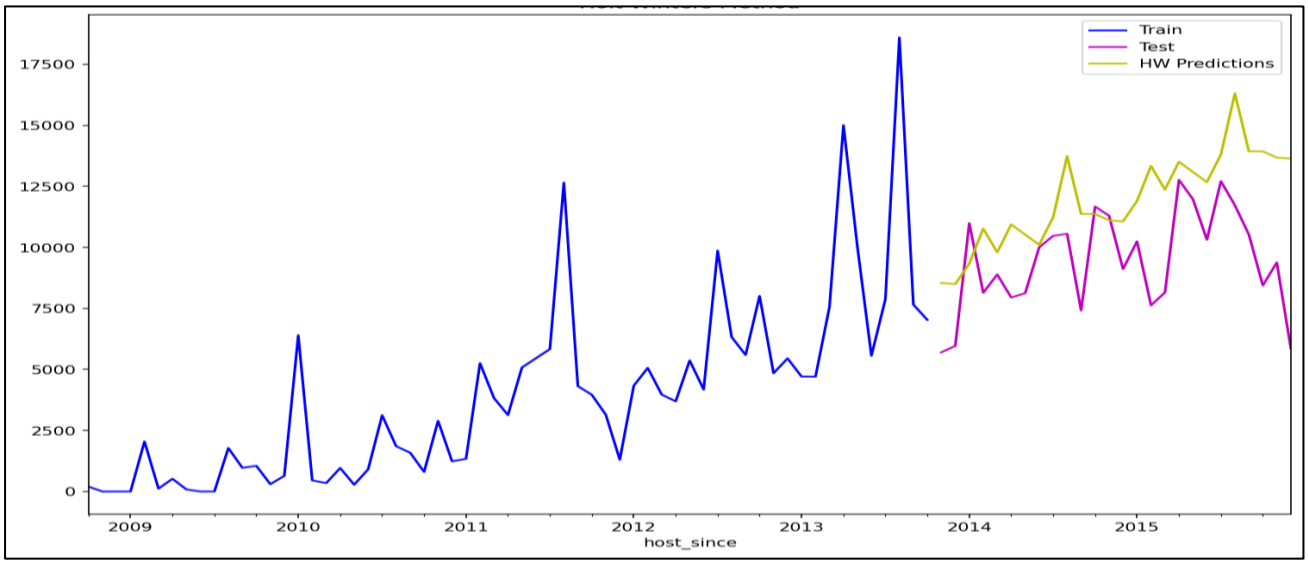


Fig. 3.4 Graph of Host Since VS Actual and Predicted Price for Triple Exponential Method

RESULT

AIC (Akaike Information Criterion)	986.1
MAE (Mean Absolute Error)	2647.98
RMSE (Root Mean Square Error)	3241.71

Model 2: SARIMA Model

Model Summary

Dep. Variable:	Price	No. Observations:	61
Model:	SARIMAX(0, 1, 2)x(0, 1, 2, 12)	Log Likelihood	-447.381
Date:	Wed, 27 Jul 2022	AIC	904.763
Time:	11:40:45	BIC	914.119
Sample:	10-01-2008	HQIC	908.298
	- 10-01-2013		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	-0.7047	0.244	-2.885	0.004	-1.183	-0.226
ma.L2	-0.2221	0.265	-0.839	0.402	-0.741	0.297
ma.S.L12	-0.9681	0.418	-2.315	0.021	-1.788	-0.148
ma.S.L24	0.6737	0.542	1.243	0.214	-0.389	1.736
sigma2	5.965e+06	2.25e+06	2.650	0.008	1.55e+06	1.04e+07

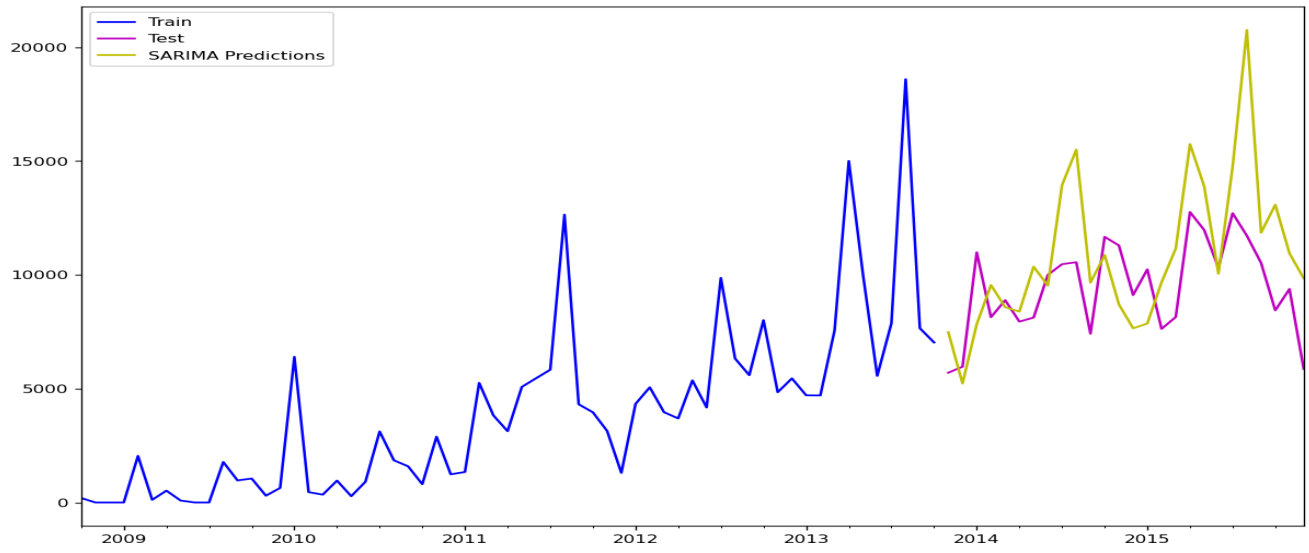


Fig. 3.5 Graph of Host Since VS Actual and Predicted Price for SARIMA model

RESULT

AIC (Akaike Information Criterion)	1375.9
MAE (Mean Absolute Error)	2356.05
RMSE (Root Mean Square Error)	2962.43

MODEL 3: SARIMAX

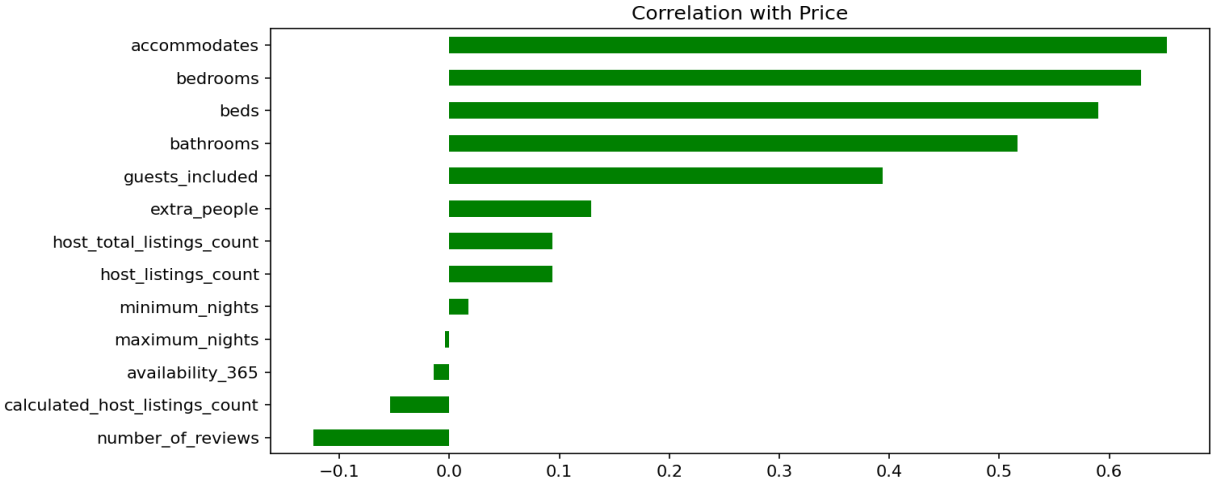


Fig. 3.6: Correlation of features with price

Dep. Variable:	Price	**No. Observations:**	61
Model:	SARIMAX(0, 1, 2)x(0, 1, 2, 12)	**Log Likelihood**	-387.306
Date:	Wed, 27 Jul 2022	**AIC**	794.611
Time:	11:40:48	**BIC**	813.323
Sample:	10-01-2008	**HQIC**	801.683
	- 10-01-2013		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
x1	24.4022	26.420	0.924	0.356	-27.380	76.185
x2	22.0447	49.543	0.445	0.656	-75.058	119.147
x3	24.5473	56.237	0.436	0.662	-85.676	134.770
x4	-18.0253	44.352	-0.406	0.684	-104.953	68.903
x5	4.6573	23.995	0.194	0.846	-42.372	51.686
ma.L1	-0.9174	0.554	-1.655	0.098	-2.004	0.169
ma.L2	-0.0815	0.391	-0.208	0.835	-0.849	0.686
ma.S.L12	-0.6764	0.580	-1.166	0.244	-1.813	0.460
ma.S.L24	0.2956	0.742	0.398	0.691	-1.159	1.751
sigma2	8.011e+05	0.000	6.4e+09	0.000	8.01e+05	8.01e+05

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	4.22
Prob(Q):	0.92	**Prob(JB):**	0.12
Heteroskedasticity (H):	4.13	**Skew:**	0.33
Prob(H) (two-sided):	0.01	**Kurtosis:**	4.29

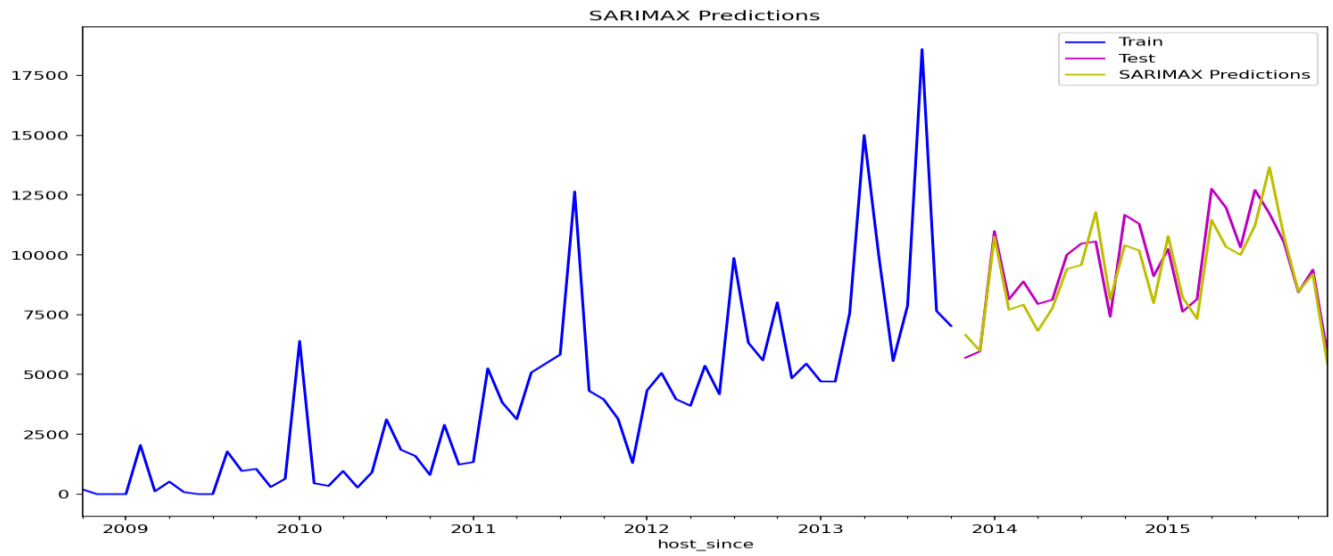


Fig. 3.7: Graph of Host Since VS Actual and Predicted Price for SARIMAX model

RESULT

AIC (Akaike Information Criterion)	794.5
MAE (Mean Absolute Error)	810.28
RMSE (Root Mean Square Error)	960.94

SUMMARY OF ALL MODDELS

MODEL	**AIC** (Akaike Information Criterion)	**MAE** (Mean Absolute Error)	**RMSE** (Root Mean Square Error)
Holt-Winters (Triple Exponent Smoothening)	986.1	2647.98	3241.71
SARIMA	1375.9	2356.05	2962.43
SARIMAX	794.5	810.28	960.94

SARIMAX model algorithm has yielded minimum error and AIC and thus finalized SARIMAX model for revenue forecasting.

FORECASTING OF NEXT ONE YEAR

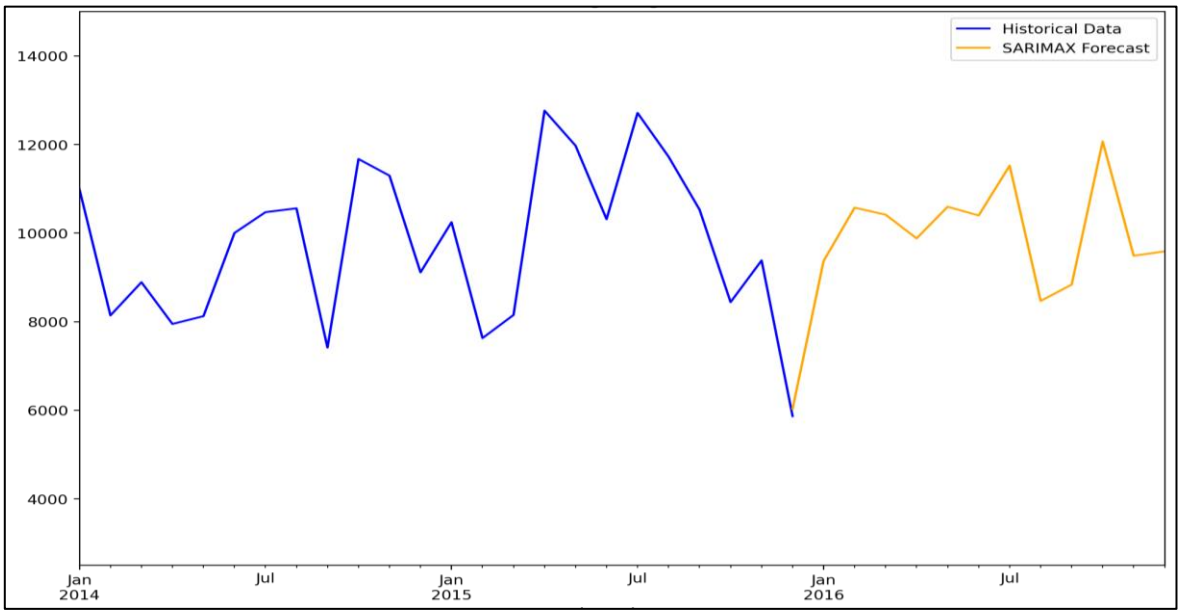


Fig. 3.8: Forecasting of next year using SARIMAX

CONCLUSION

Seasonality shows that more bookings were done in the month of August maybe due to the warm weather and specially to attend Sea fair festival where people majorly come to watch the hydroplane races and the Blue Angels which is a Seattle tradition.

PROPERTY TYPE PREDICTION

OBJECTIVE

Build a classification model that will predict the 'property type' a customer is most likely to select, given a set of input features.

INTRODUCTION TO CLASSIFICATION

The Supervised Machine Learning algorithm can be broadly classified into Regression and Classification Algorithms. In Regression algorithms, we predict the output for continuous values, but to predict the categorical values, we need Classification algorithms.

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

Types of ML Classification Algorithms:

Classification Algorithms can be further divided into the Mainly two category:

Linear Models:

- Logistic Regression
- Support Vector Machines

Non-linear Models:

- K-Nearest Neighbours
- Random Forest Classification
- Gradient Boost Classifier

The data available in our target column – property_type has 16 distinct types of properties and huge imbalance in the distribution of data. To reduce this imbalance in data the following changes were made:

Types of Properties	Number of available types
House	1723
Apartment	1696
Townhouse	118
Condominium	91
Loft	40
Bed & Breakfast	37
Cabin	21
Other	21
Camper/RV	13
Bungalow	13
Boat	8
Tent	5
Treehouse	3
Dorm	2
Yurt	1
Chalet	1

- Property type Townhouse, Loft, Bed & Breakfast, Cabin, Camper/RV, Bungalow, Boat, Tent, Treehouse, Yurt, Chalet is merged with House.
- Property type Condominium and Dorm is merged with Apartment.

After making the changes we now have a binary classification task to perform

Types of Properties	Number of available types
House	1983
Apartment	1789

FEATURES SELECTION ALGORITHMS

The following algorithms were used to understand which features are most applicable to our binary classification task

- ExtraTreeClassifier
- Forward Selector
- Chi2 Selector

ExtraTreeClassifier	Price, room type private room, availability 365, number of reviews, bedrooms
Forward selector	Host total listings count, bedrooms, room type private room, room type shared room
Chi2 selector	Host total listings count, number of reviews, availability 365, Price

After all the possible combinations of features to get best performance the final features shortlisted for property type prediction are:

Features shortlisted for Property type prediction

- Host total listings count
- Bedrooms
- Price
- Room type Private room

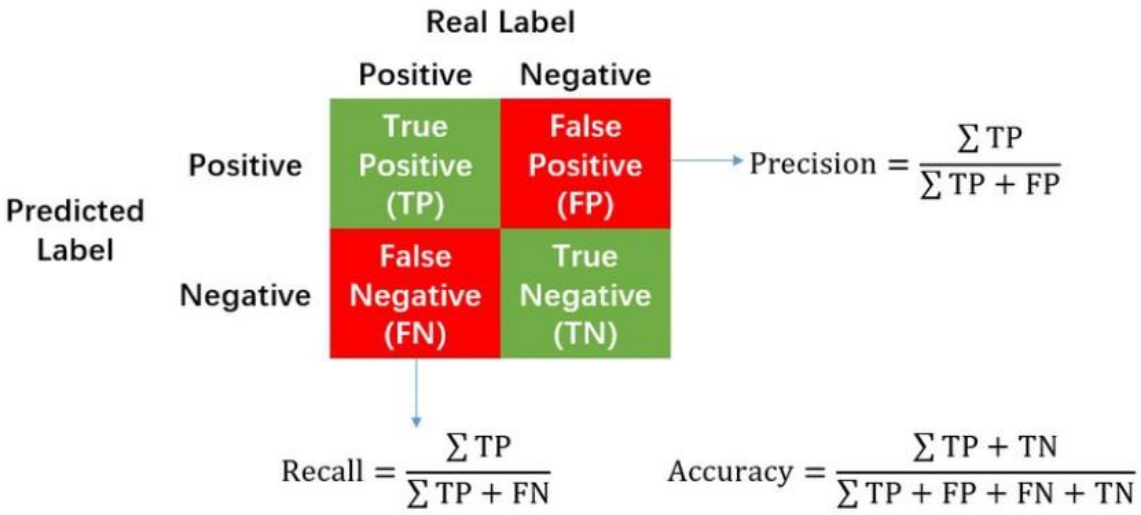
Evaluating a Classification model:

Confusion Matrix:

The confusion matrix provides us a matrix/table as output and describes the performance of the model.

It is also known as the error matrix.

The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions. The matrix looks like as below table:



The following is the performance of various classification algorithms:

1. **Logistic Regression:**
   Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

   Classification report:

	precision	recall	f1-score	support
Apartment	0.74	0.73	0.73	523
House	0.77	0.78	0.77	609
accuracy			0.75	1132
macro avg	0.75	0.75	0.75	1132
weighted avg	0.75	0.75	0.75	1132

2. **K Nearest Neighbours:**
   In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour.

Classification report:

	precision	recall	f1-score	support
Apartment	0.75	0.75	0.75	523
House	0.79	0.79	0.79	609
accuracy			0.77	1132
macro avg	0.77	0.77	0.77	1132
weighted avg	0.77	0.77	0.77	1132

3. **Random Forest Classifier:**
Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

Classification report:

	precision	recall	f1-score	support
Apartment	0.70	0.72	0.71	523
House	0.76	0.73	0.74	609
accuracy			0.73	1132
macro avg	0.73	0.73	0.73	1132
weighted avg	0.73	0.73	0.73	1132

4. **Support Vector Classifier:**
In machine learning, support-vector are supervised learning models with associated learning algorithms that analyse data for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVM maps training examples to points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

Classification report:

	precision	recall	f1-score	support
Apartment	0.77	0.72	0.75	523
House	0.77	0.81	0.79	609
accuracy			0.77	1132
macro avg	0.77	0.77	0.77	1132
weighted avg	0.77	0.77	0.77	1132

5. **Gradient Boost Classifier:**
   Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees.[1][2] When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest.[1][2][3] A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

   Classification report:

	precision	recall	f1-score	support
Apartment	0.75	0.76	0.75	523
House	0.79	0.78	0.78	609
accuracy			0.77	1132
macro avg	0.77	0.77	0.77	1132
weighted avg	0.77	0.77	0.77	1132

CONCLUSION

From all the 5 models, Gradient Boosting Classifier gives the maximum accuracy of 77.8%.

DEPLOYMENT

After committing all the required files on GITHUB and using FLASK API, the model is deployed on HEROKU platform:

https://property-type-recommender.herokuapp.com

REFERENCES

www.kaggle.com
www.medium.com
https://public.opendatasoft.com
https://towardsdatascience.com
https://www.niit.com/india/knowledge-centre/ML-algorithms