2020

Stripping customer feedback on Las Vegas hotels through data analysis

STUDENT ID: 102010742
BHANUSREE ALAPARTHY

Ph: 0451227373
Date: 29-5-2020

Abstract

This report give analysis on hotels in Las Vegas. Based on the TripAdvisor score by taking 504 reviews from 21 hotels all the analysis is done. The study shows features affecting the score which helps the hotels to maintain based on important features and by various graphical visualizations it helps the tourists to enhance the journey in prior and can know the services they provide. It also enhances the season effect on the hotels which helps travellers to know which the best season is to visit and lastly it helps to predict the score by using different types of models and comparisons, this helps to gain the future analysis of the tourists visit.

Introduction:

In the present situation online travel agencies plays a major role in travel booking, which includes transport and accommodation. Online reviews have increased its use, in fact many potential hotel customers prefer to see feedback before their purchase decisions. As every consumer have their access to internet and it became very easy to express their positive or negative feedback, it means people give their review based on their experience and opinions. In this report main aim is to predict score by using predictive analysis and by using different modelling classifiers. It also shows few visualizations which helps to know all the columns and their relationships between them and used feature importance to know the important columns which are affected due to score. Later modellings methods and test and training are used to know the accuracy and score of the data.

Dataset Information:

Dataset consists of quantitative and categorial data from reviews collected from 21 hotels which are in Las Vegas Strip. This data set is extracted from TripAdvisor website and it has got 504 reviews in total, in which reviews are collected from different parts of the world. Las Vegas Strip Dataset consists of 504 rows with 20 columns. The present dataset consists of all the data on review features, user features and hotel features information. All the reviews are featured from the Traveller type, period of stay, review month or weekday. It also got user information regarding users like user country , user continent , member years and their reviews and votes on hotels. The hotel data consists of their hotel name , number of rooms and gives whether it has got gym, casino , spa, internet ,tennis court and other facilities.

Task 1

Problem Formulation

In the Las Vegas strip dataset, review score is a major factor to be considered as every person rate the hotel and new customer try to go to the hotel based on the review score. TripAdvisor must get the analysis how the people are rating based on different features. Different sets of people go to hotels and stay for certain period and give ratings and it differs from each one perspective. All the features are analysed and calculated based on the Score. But the problem comes for the hotels to maintain same score and standards in the future. For that score is predicted by analysing all the features which influences Score. So that they can know their future scores in prior and can maintain it.

Data acquisition

The data set of the Las Vegas Strip is extracted from the UCI machine learning repository with all the descriptions and file has its headers for each column and score is set as class label in the data set and

it satisfies all the criteria given. It has got more then 150 rows and more then 5 columns and has got categorical columns.

Data preparation

At the initial stage, the dataset is in comma separated columns as it is not in table. So firstly, it was created into table and changed the column names for convenience .

Importing the libraries:

All the libraries are importing like NumPy, pandas, sklearn ,seaborn, mat plot for visualizations and imported other feature libraries which are needed for modelling .

Loading the data:

The dataset named trip1.csv is loaded and read into the juypter notebook and all the 504 rows and 20 columns are read.

Cleaning the dataset:

Removing null values*:* checked whether the dataset contains null values and fount null values in no of rooms, user continent, member years, review month and review weekday. As I got 504 rows it makes sense to remove the null values rows.

Checking duplicates*:* All the duplicates are checked, and it has not got any duplicates in the rows or columns.

Changing the datatypes*:* checked whether all are they are in their respective datatypes . changed the datatypes for member years , number of years and hotel stars for easy convenience to visualize and understand the data

Checking unique values: checked the unique values for each column helps to know whether it has blank spaces, spelling errors or other upper- or lower-case errors.

Descriptive statistics for numerical and object columns separately to analyse and know more about the data

Feature importance :

As the dataset consists of 20 attributes and has taken score as class label , where it helps to do predictive analysis . More than taking all the 20 attributes for analysis it will be more convenient to select top 10 columns related to score. To get the most contributing featured columns for score following steps are undertaken

- Firstly, selected and defined all the categorical columns in the dataset
- Transformed all the categorical columns into numbers, by doing this conversion it helps to calculate and compare all the columns.
- Removed the score label form the data so that while finding featured columns the score label will not include, and it will be easy to compare.

- Prepared train and test labels
- Executed recursive feature elimination and applied random forest feature to the class and fount the most contributing features of the Score.
- fig1.1 it shows the most featured columns of the Score.

```
*Most contributing features in Score*
()
['Period_of_stay' 'Traveler_type' 'Pool' 'Tennis_court' 'Spa' 'Casino'
 'Free_internet' 'hotelstar' 'User_continent' 'Review_weekday']
```

Figure 1.1 Most contributing features of score

Task 2 Data Exploration

Based on the most contributing columns of the score , 10 columns are taken for exploration

2.1 univariate analysis

Descriptive statistics

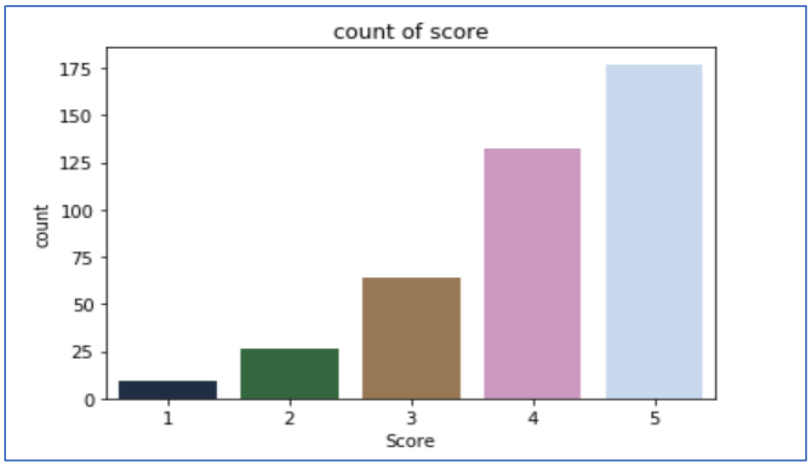
	count	mean	std	min	25%	50%	75%	max
No_of_reviews	408.0	43.848039	72.075080	1.0	12.0	22.0	47.25	775.0
No_of_hotelreviews	408.0	14.764706	23.086283	0.0	5.0	9.0	17.00	263.0
Helpful_votes	408.0	29.767157	46.142629	0.0	8.0	16.0	31.00	365.0
Score	408.0	4.083333	1.019667	1.0	4.0	4.0	5.00	5.0
hotelstar	408.0	4.235294	0.807538	3.0	4.0	4.0	5.00	5.0
No_of_rooms	408.0	2540.529412	1177.605288	315.0	1467.0	2916.0	3348.00	4027.0
Member_years	408.0	-0.068627	89.674624	-1806.0	2.0	4.0	7.00	13.0

Figure 2 Descriptive statistics

Fig 2 shows the Descriptive statistics for all the numerical values

- Score is the class label; the values ranges from 1 to 5 and it has got the average value of 4
- Hotel star rating is ranging from 3 to 5 and the mean value is 4, It states that all the hotels are rated high. As many of the hotels got high ratings , the score also will be high based on hotel stars
- As most of the hotel have high scores it can be known that the quality and infrastructure of the hotels are good as it has got on an average 2540 rooms in the hotel.

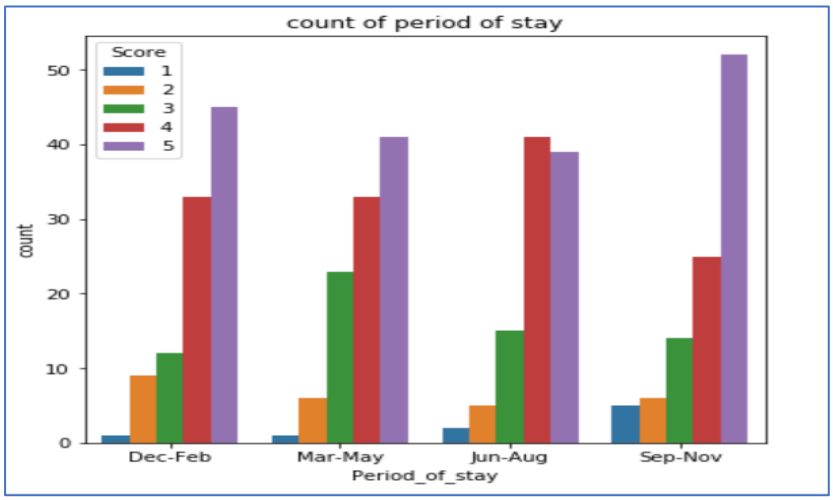
Count of Score




```
5    177
4    132
3     64
2     26
1      9
Name: Score, dtype: int64
```

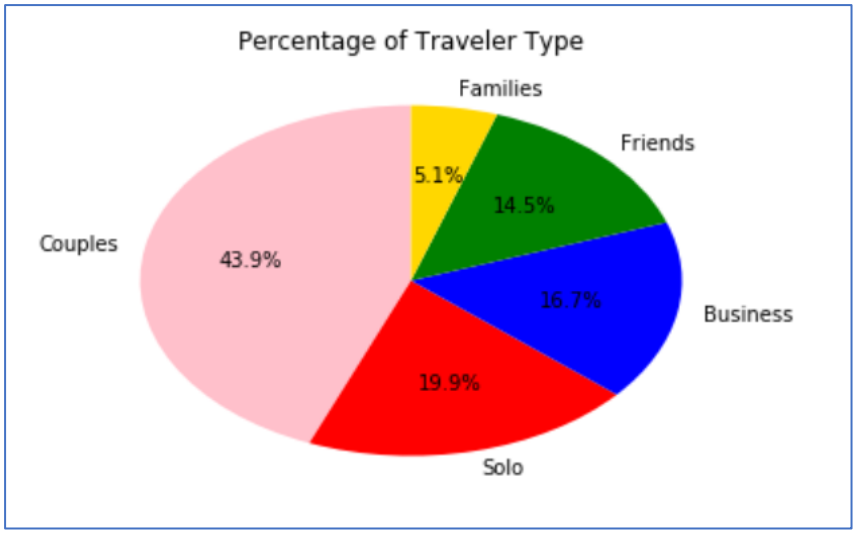
Above bar graph shows the count of score, from this it helps to know how most of the scores are given for every hotel. It shows that majority of the scores are 4 and 5 .

Period of stay based on scores



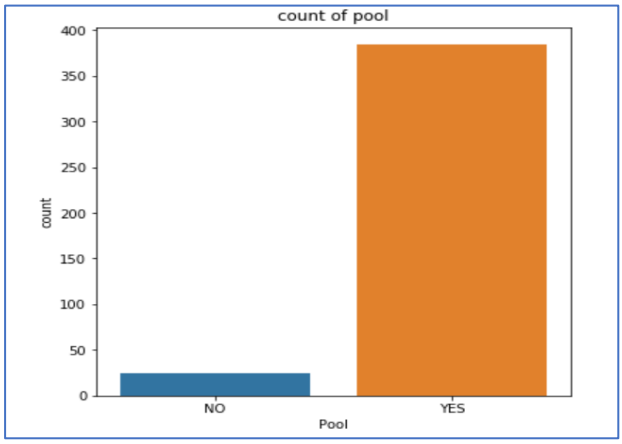
As it is known that most of the scoring comes as 4 or 5 , the period of stay is also ranging more for 4 and 5 . Above figure shows the count of stay based on the ratings provided. From this it helps to estimate how the Scores are ranging in each period.

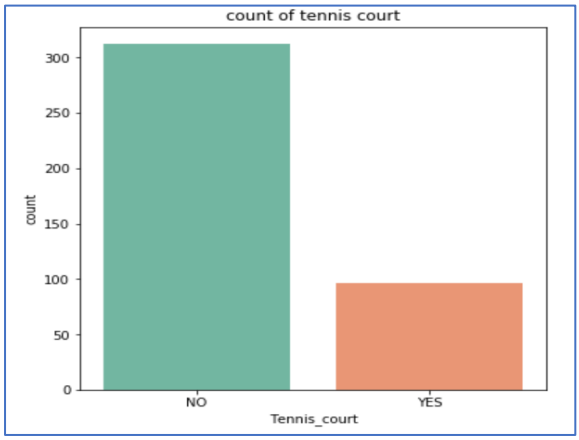
Percentage of Traveller type

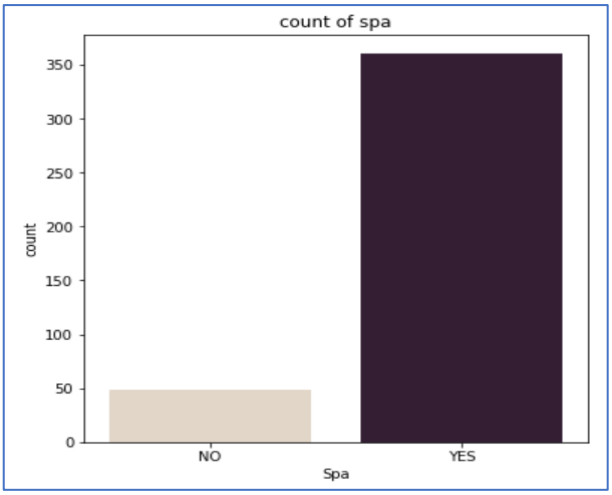


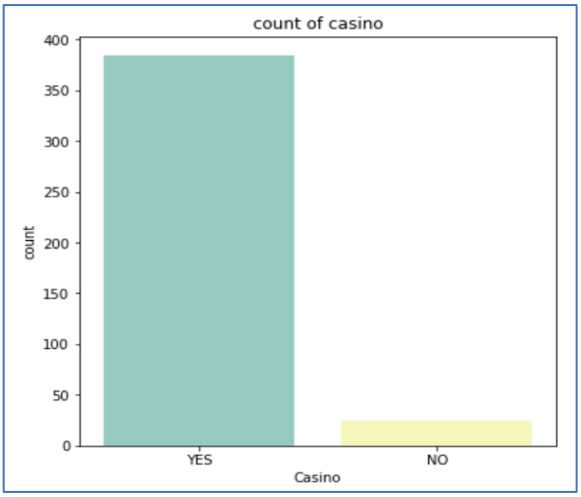
Above figures illustrates the percentage of the traveller types like couples, solo, families, business, and friends . It is clearly shown that couples entrees are more than other traveller types and families do visit very less. It is good to implement some honeymoon packages for couples ,as couple often travel more as families visit is less hotels need to advertise the services provided for children and as a family service.

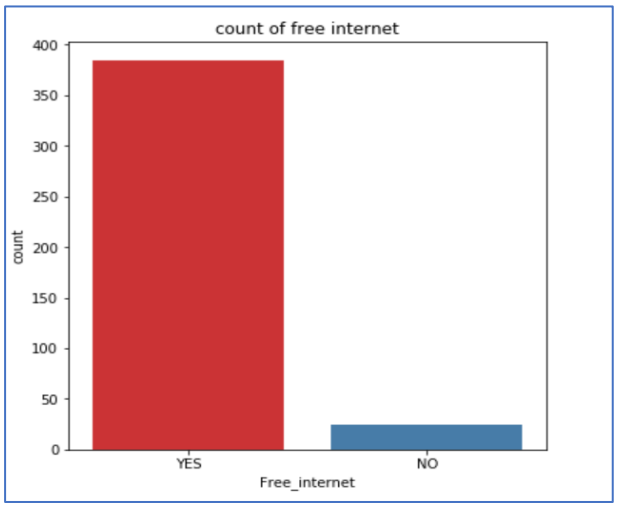
Different types of services provided by the hotels





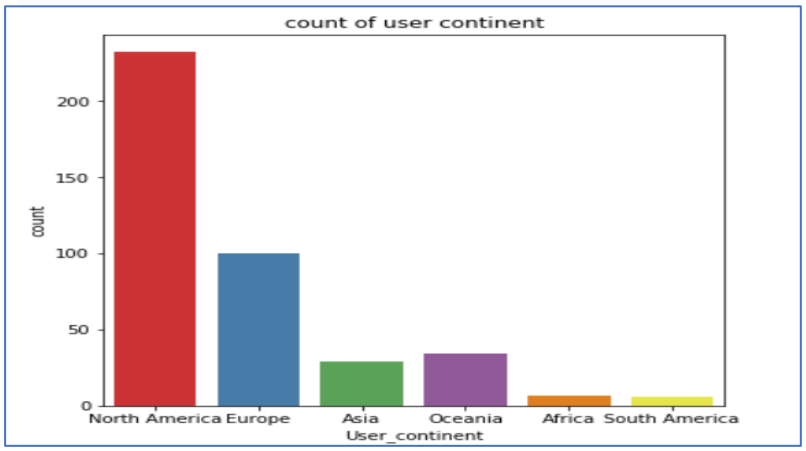






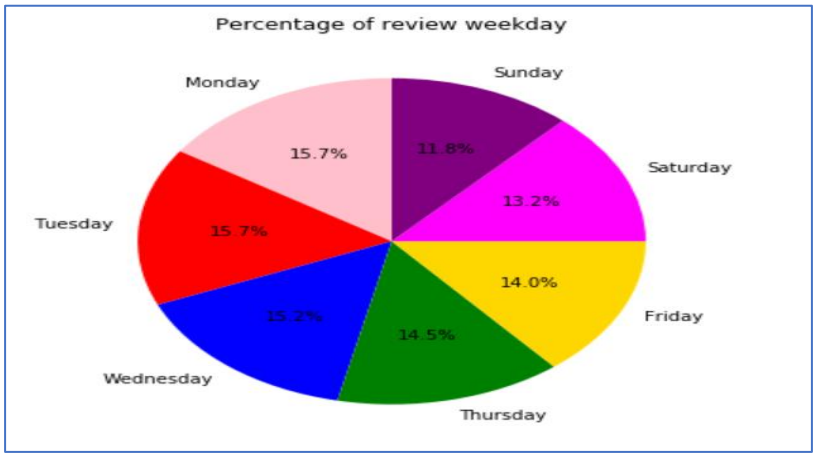
Different types of services are provided by the hotels like pool, tennis court, spa, casino, and free internet. It shows that majority of the services are provided but very few numbers of hotels provide the tennis courts as it takes lot of space and need environment there are less chances of providing tennis courts for the hotels

Types of continents of users



As hotels are from Las Vegas most of the visitors will be Americas so majority of the user continents are from north America

percentage of review weekday



Slightly more percentage of entrees comes more on Monday and Tuesday

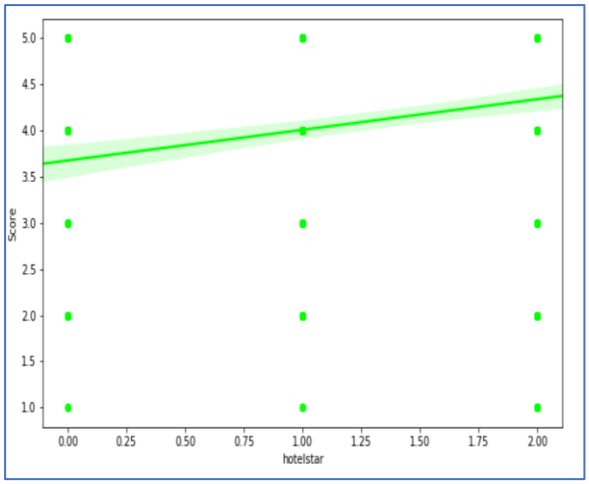
2.2 Relationships between columns

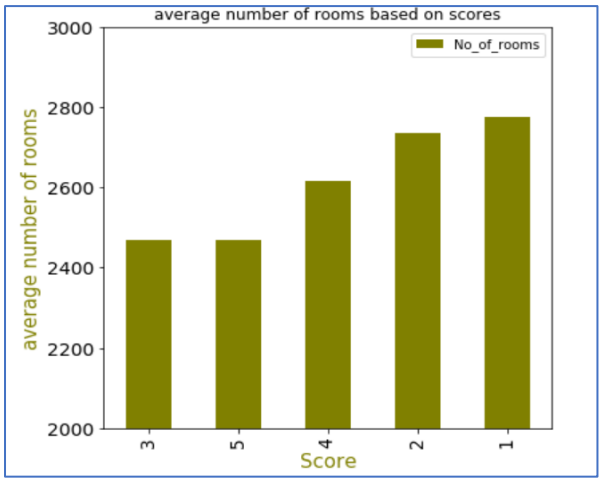
	Period_of_stay	Traveler_type	Pool	Tennis_court	Spa	Casino	Free_internet	hotelstar	User_continent	Review_weekday	Score
Period_of_stay	1	-0.042	-0.035	-0.0049	-0.024	0.0022	0.0022	-0.013	-0.034	-0.084	-0.0094
Traveler_type	-0.042	1	-0.11	-0.027	-0.063	0.019	0.038	-0.056	-0.058	0.00079	0.019
Pool	-0.035	-0.11	1	0.14	0.68	-0.062	-0.063	0.38	-0.047	0.031	0.21
Tennis_court	-0.0049	-0.027	0.14	1	0.2	0.14	0.14	-0.16	0.022	-0.021	0.068
Spa	-0.024	-0.063	0.68	0.2	1	0.68	-0.091	0.56	-0.087	-0.0066	0.18
Casino	0.0022	0.019	-0.062	0.14	0.68	1	-0.062	0.38	-0.072	-0.04	0.031
Free_internet	0.0022	0.038	-0.063	0.14	-0.091	-0.062	1	0.073	0.088	0.0006	0.19
hotelstar	-0.013	-0.056	0.38	-0.16	0.56	0.38	0.073	1	-0.031	-0.019	0.26
User_continent	-0.034	-0.058	-0.047	0.022	-0.087	-0.072	0.088	-0.031	1	0.022	0.1
Review_weekday	-0.084	0.00079	0.031	-0.021	-0.0066	-0.04	0.0006	-0.019	0.022	1	-0.11
Score	-0.0094	0.019	0.21	0.068	0.18	0.031	0.19	0.26	0.1	-0.11	1

This is the co-relation matrix for the most 10 contributed features of the Score.

	No_of_reviews	No_of_hotelreviews	Helpful_votes	Score	hotelstar	No_of_rooms	Member_years
No_of_reviews	1	0.59	0.77	-0.025	-0.029	-0.088	0.022
No_of_hotelreviews	0.59	1	0.73	0.013	-0.078	-0.09	0.022
Helpful_votes	0.77	0.73	1	0.019	-0.0056	-0.061	0.023
Score	-0.025	0.013	0.019	1	0.26	-0.05	-0.042
hotelstar	-0.029	-0.078	-0.0056	0.26	1	0.27	0.016
No_of_rooms	-0.088	-0.09	-0.061	-0.05	0.27	1	-0.016
Member_years	0.022	0.022	0.023	-0.042	0.016	-0.016	1

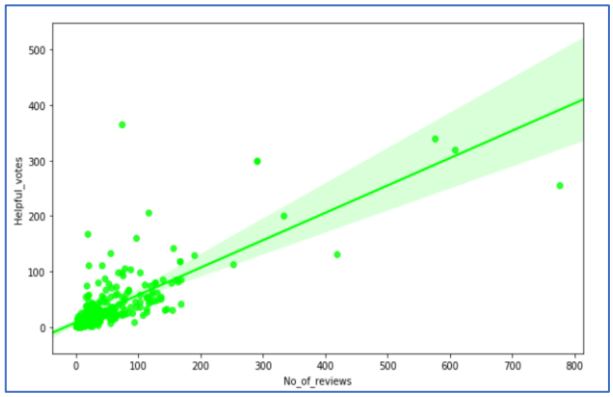
This the co-relation matrix for all the numerical columns

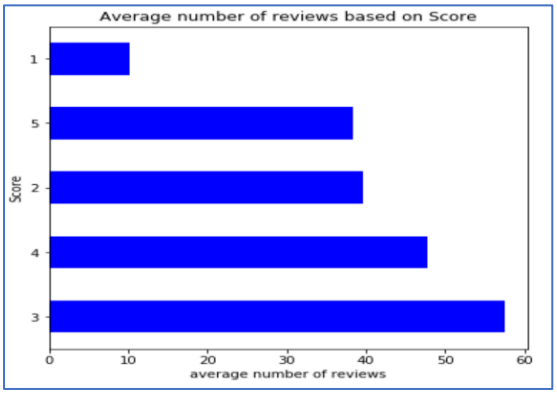




When we take co-relation between score and other columns the highest relationships come with hotel star and pool .As we can see in the graph the regression is done in between score and hotel star which is weakly positively co related . But it can be noted that as hotel star increases the score also increases .

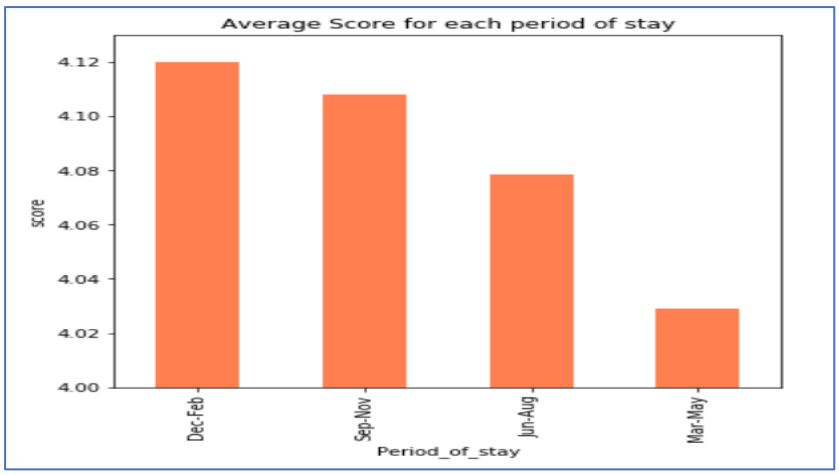
In the second graph it is noted that as the hotel star is more the number of rooms decreases. It is also can be said that as number of rooms increases the score and hotel stars decreases. It happens as small hotels tend to offer a friendlier and non-crowd environment . This suits the tourists enjoying the quiet stays.





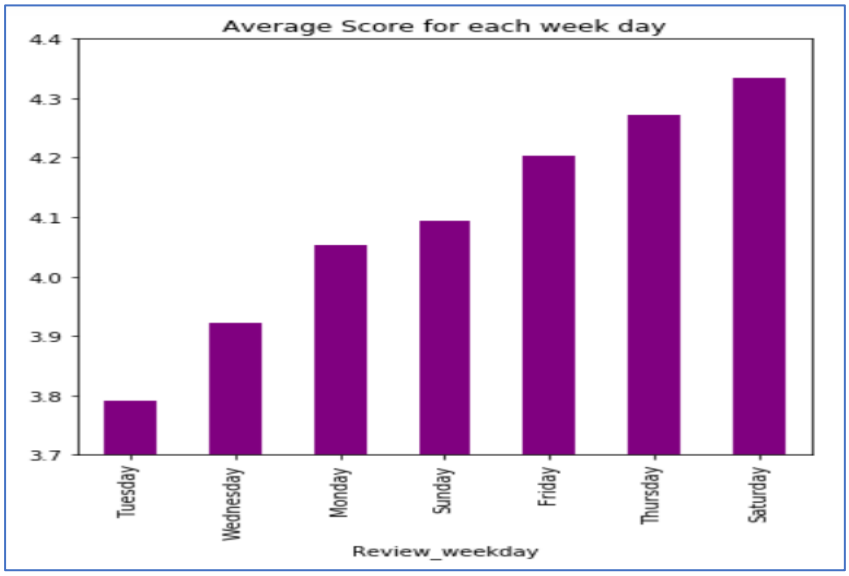
From the co relation matrix it shows that helpful_votes and No_of_reviews was near to 1. After plotting regression it shows that those are positively co-related. It shows that the reviews are more from 0 to 200 . It shows that trip advisor members have propabilty to look other people reviews and from the second plot it is observed that most number of reviews are more on 3 and 4 scores than 5 it states that it needs little more improvement on stay.

Ratings on period of stay:



Above plot states the relation between score and period of stay , it shows the seasonality effect on the Score. There are more number of ratings in the periods of dec-feb and sep-nov. considering that las vegas is located in hot dessert, it seems like it has got tourist attractions during the colder months like autumn and winter seasons.

Weekdays scoring:

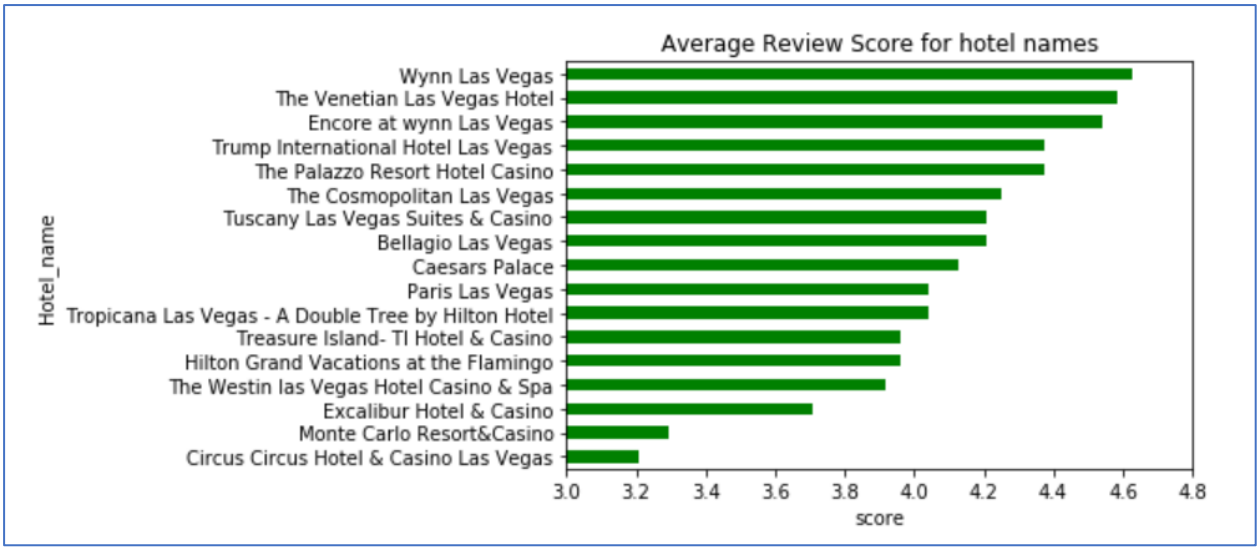


The plot shows that people give more feedbacks most on weekends, as Saturday is the first weekend day it is possible that people got a postive physchologically effect on people, which helps in getting high scores on Saturday.

2.3 Question

Which hotel scored the highest rating in the Las Vegas?

The hotel Wynn Las Vegas gives the highest scoring than other hotels in the Las Vegas . By calculating the average scoring with all the hotels, it tells which hotel gives the highest scoring . It also can be confirmed by calculating the mean with hotel star and review coming from that hotel , still it stands in top. Overall, it can be said that Wynn Las Vegas got the highest scoring in Las Vegas. People who aims for perfect hotel Wynn Las is best option.



Task 3 Data Modelling

3.1 splitting the data

Steps involved for splitting

- Firstly, it is needed to import the libraries like train, test and split that helps to split into the train and test subsets
- Splitting is divided into x and y for both training and testing.
- In X1 all the columns are taken and dropped the Score column.
- In Y1 only the score is taken .
- Later X1 and Y1 are spilled into x_test, y-test , x_train and y_train using train_test_split .
- After passing the test and train it is needed to set the test size based on the splitting
- For **Suite1: 50% for training and 50% for testing** the test size is set to 0.5
- For **Suite2: 60% for training and 40% for testing** the test size is set to 0.4
- For **Suite3: 80% for training and 20% for testing** the test size is set to 0.2
- As the testing size is setting based on that the training will be automatically set.

3.2 choosing the models

Random forest and Gaussian process are the two models chosen to train the model

Identification of method:

- The method used for the random forest is “ensemble” sklearn.ensemble import RandomForestClassifier.
- The method used for the random forest is “Gaussian process” sklearn.gaussian_process import GaussianProcessClassifier

Parameters used for training models:

- Randomforest :Parameters for random forest are random state which helps in controlling the randomness and bootstrapping and other parameters which act as default prameters are
    1. Max depth is default parameter it expands untill the leaves are pure
    2. Max_samples is default and it helps to draw in shaping the samples

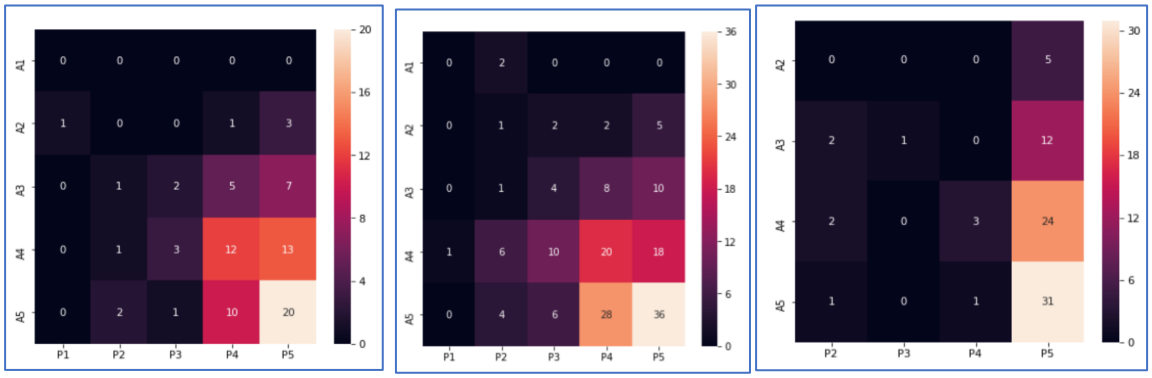
- Parameters for gaussioan process are random state used for number generation to intialize centers and other parameters which act as default are
    1. Kernel 1.0 is used as default

Evaluation for performance of model for training and testing the data sets :

Random forest

Confusion matrix for all the three suites

Suite 1: **Suite 2:** **Suite 3:**



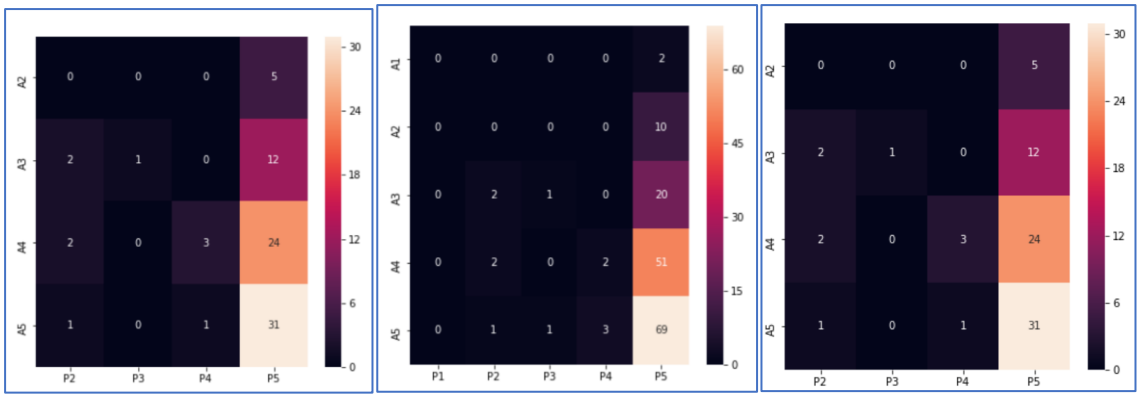
For clear picture see in juypter notebooks

Suite number	Training accuracy	Testing accuracy	Score	Precision	Recall	F1-Score	Support
1	0.97	0.41					
			1	0	0	0	0
			2	0	0	0	5
			3	0.33	0.13	0.19	15
			4	0.43	0.41	0.42	29
			5	0.47	0.61	0.53	33
2	1.0	0.37					
			1	0	0	0	2
			2	0.07	0.10	0.08	10
			3	0.18	0.17	0.18	23
			4	0.34	0.36	0.35	55
			5	0.54	0.49	0.50	74
3	0.97	0.41					
			1	0	0	0	0
			2	0	0	0	5
			3	0.33	0.13	0.19	15
			4	0.43	0.41	0.42	29
			5	0.47	0.61	0.53	33

Gaussian process classifier

Confusion matrix for all the three classifiers

Suite 1: **Suite 2:** **suite 3:**



For clear picture see in juypter notebooks

Suite number	Training accuracy	Testing accuracy	Score	Precision	Recall	F1-Score	Support
1	1.0	0.44					
			1	0	0	0	5
			2	0	0	0	10
			3	0.50	0.03	0.06	29
			4	0.43	0.04	0.08	67
			5	0.46	0.92	0.61	93
2	1.0	0.43					
			1	0	0	0	2
			2	0	0	0	10
			3	0.50	0.04	0.08	23
			4	0.40	0.04	0.07	55
			5	0.45	0.93	0.61	74
3	1.0	0.42					
			1	Not applicable	Not applicable	Not applicable	Not applicable
			2	0	0	0	5
			3	1.00	0.07	0.12	15
			4	0.75	0.10	0.18	29
			5	0.43	0.94	0.59	33

Results :

- Above two models are taken to evaluate the training and testing it shows on an average that testing accuracy for the models are 40 and 42 , shows that accuracy is very bad
- The accuracy is not so good enough as the score is ranging from 1 to 5 and all the hotels are very high class only 4 and 5 scores are more . So it is very difficult to distribute the samples in such cases overfitting can be done.
- In the confusion matrix while predicting the p5 it is showing the prediction but for other cases it is showing wrong prediction as the values are more so the prediction seems wrong in such cases.
- Precision shows the measures when the prediction is positive in most of the cases score 3 , 4 and 5

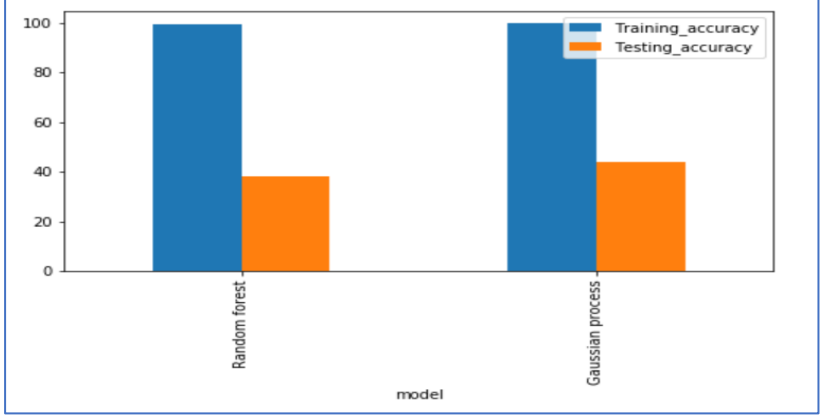
- Recall helps in measuring prediction positive classes in this case also 3 , 4 and 5 shows postive results but it does not get more then precision
- F1 nothing but the average of precision and recall by depending on that it can be estimated that 4 and 5 produces the good results.
- It can also be noted that suite 3 of gaussian model takes only 2 to 5 scores as while training score 1 does not get to fit any values
- Suggestion: oversampling can be done by rising the scores, it might increase the overall accuracy .

3.3 Comparision of models

Suite 1:

```
              model  Training_accuracy  Testing_accuracy
0     Random forest              99.51             38.24
1  Gaussian process             100.00             44.12

<matplotlib.axes._subplots.AxesSubplot at 0xfe76550>
```



Suite 2:

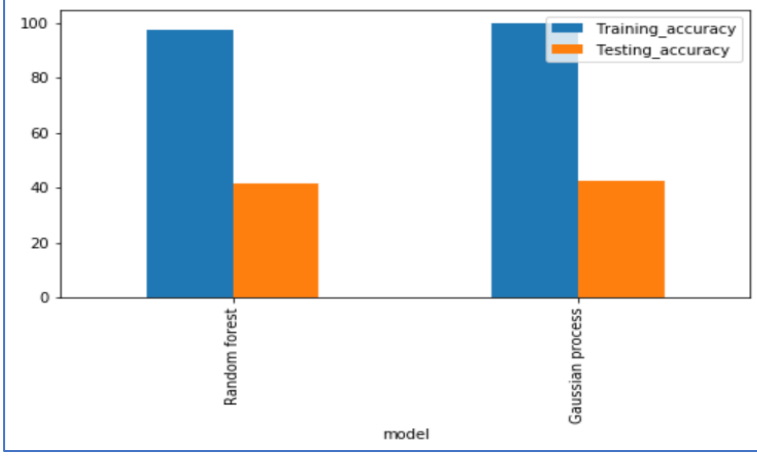
```
              model  Training_accuracy  Testing_accuracy
0     Random forest              100.0              37.2
1  Gaussian process              100.0              43.9

<matplotlib.axes._subplots.AxesSubplot at 0x14677208>
```

Suite 3:

```
              model  Training_accuracy  Testing_accuracy
0     Random forest              97.55             41.46
1  Gaussian process             100.00             42.68

<matplotlib.axes._subplots.AxesSubplot at 0x145e4240>
```



Firstly, I tried comparing with ROC and precision recall curve, but it takes only binary data as Score has got 5 options it not being applicable with them. Secondly done with box plot as it not good for visualization came up with a bar plot.

Comparing all three plots from three different suites it shows that it has got a good training accuracy. Whereas, testing accuracy is very bad due to the impact of score. Both the models predict almost equal accuracy, but gaussian process model shows slightly more than the random forest.

Discussions and conclusions:

It can be said that online feedback reviews as lot of impact on upcoming tourists who are visiting. All the hotels are high in quality and rating and it has got a good number of services. Overall rating can be known by the quantitative 1 to 5 . From the analysis it is known that it good for couples visit and during the weekends like Thursday, Friday, and Saturday it has got feedback. It is also known that there is good chance of increasing Score by maintaining a hotel star. After the modelling is done it suggested to do oversampling by increasing the score of 1 to 3 as a greater number of entries are in 4 and 5,So that the accuracy might increase. The gaussian process seems slightly good then random forest even though it has got a bad accuracy. Finally, all the hotels are in between 3 to5 stars based on the reviews and services it good to enhance the journey before travelling.