Capstone Project Final Report

Prediction of hotel cluster based on historical browsing data

Mentor: Ryan Rosario

Mentee: Tanzina Zaman

1 INTRODUCTION

1.1 Problem definition:

The project proposal is based on a sponsored competition from Kaggle. The client, Expedia wants to predict which hotel cluster is booked by the customer, based on customer's previous bookings and clicks. A click means a user clicked a link to see hotel details on a hotel information site page. The service provider has predefined hotel clusters based on historical price, customer star ratings, geographical locations relative to city center, etc. New hotels which have no historical data are considered as outliers.

1.2 Business need:

Hundreds of people look for hotel either for work purpose or for vacation. And numerous options appear when a potential customer search for hotel that will fit his need. The aim of this project is to help the client to offer a recommendation system which will predict the types of hotel the potential customer may chose based on his previous browsing history. This would make the hotel selection process easier for a potential customer as he would get recommendation for hotels that will match his need and guidance towards right direction.

2 DATA DESCRIPTION

For this project, dataset from a sponsored competition on Kaggle[1] is used. In the competition, the dataset is divided into training and test set based on the date of the observation is recorded. Entries from 2013 and 2014 are in training set, while test data are from 2015. Two datasets are differentiated by a key feature which describes whether the user booked that hotel or browsed only. Both booking and browsing events of the users are included in training set, but the test set only has the booking events. An additional dataset is supplied in the competition which contains the

feature extracted from hotel reviews. In this project, only the given training set is considered due to large volume.

The training set is considered as the parent source of information in this project which contains 37 millions of observation and 24 fields. Among the 24 fields, one is considered as target attribute and rest are used as feature attribute. These fields include a timestamp as well as the following information: Expedia point of sell, customer country, region and city, the distance between the customer and the searched hotel, whether or not the customer was browsing through a mobile device etc. These fields provide the information related to hotel search. To understand the data fields, a quick glance at the Expedia website would be helpful.

The first thing that draws any customers attention while using Expedia website is 'Going to' tab. This tab provides the information of destination type, the corresponding continent, country and the market of a hotel. The second important feature of the site are 'check in' and check out' tab, which maps to two more feature attributes of the dataset, check in and check-out date respectively. The three tabs next to 'check out' tab, map to the number of adults, children and rooms specified during the hotel search. If the customer uses 'Add a flight' tab, then the transaction is considered as a package.

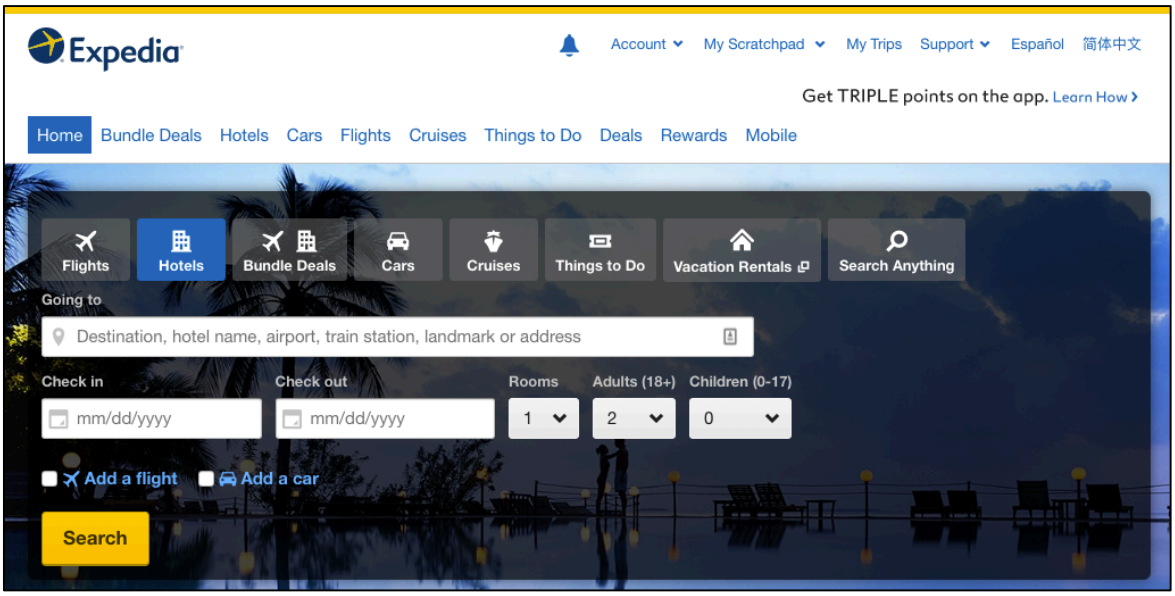


Figure 1: An example of Expedia interface to search or book hotels online

3 EXPLORATORY DATA ANALYSIS

The amount of data has a very important influence on data analysis technique to be chosen in such case. Here, the dataset has more than 37 million rows and 24 fields. The click and booking events in the dataset are not equally distributed and this is quite natural as people use to browse a lot before they finally decide to book any hotel. The below bar-chart shows the count of click and booking events.

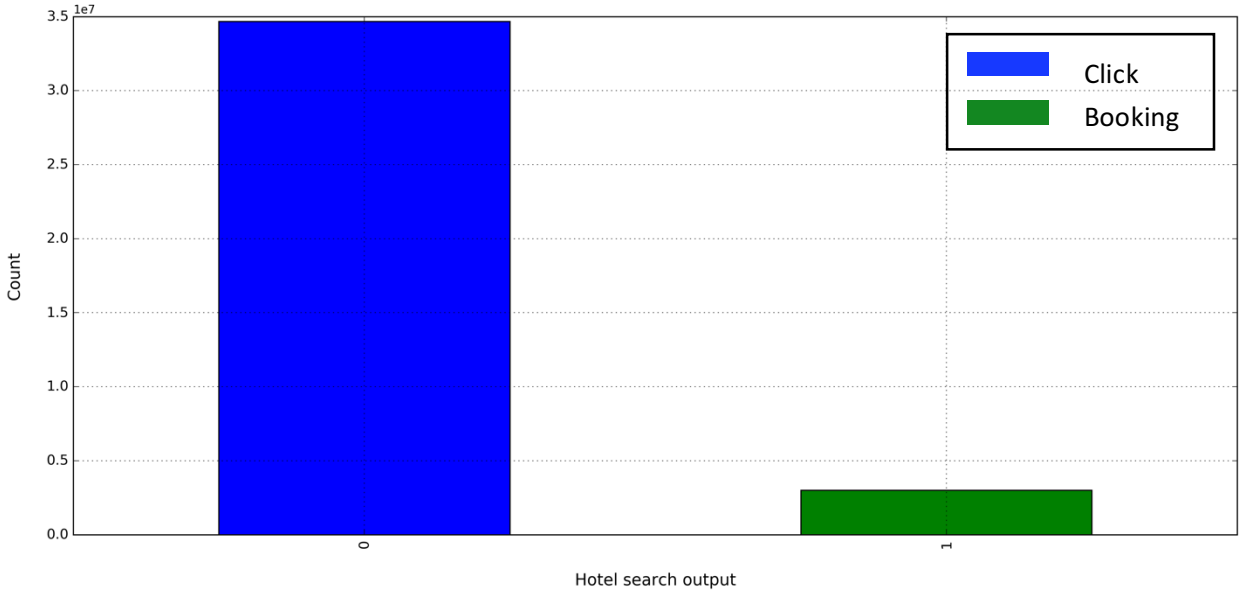


Figure 2: Count of click and booking events present in the dataset

People tend to search for hotels almost all over the year, but the figure 4 states that the search trend gradually increases during summer season and lasts till holiday season in the end of year.

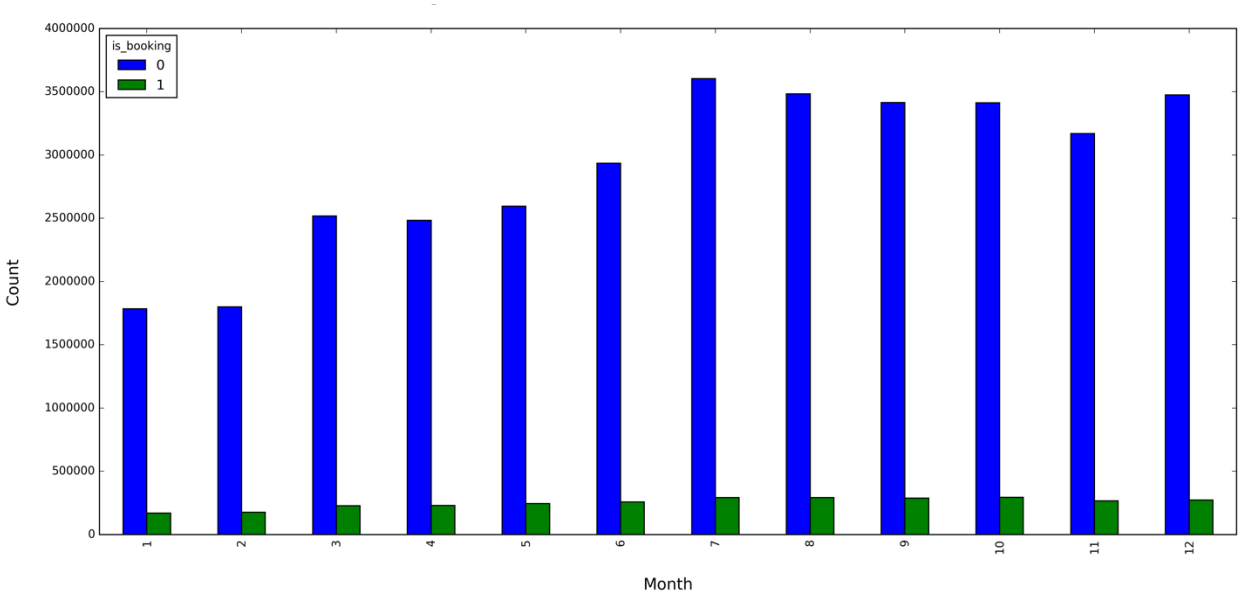


Figure 3: Trend of hotel search varies over month.

People also tend to search more during the start of a week, rather than end of the week. In figure 5, the ratio of booking events is presented as a ratio of total search event.

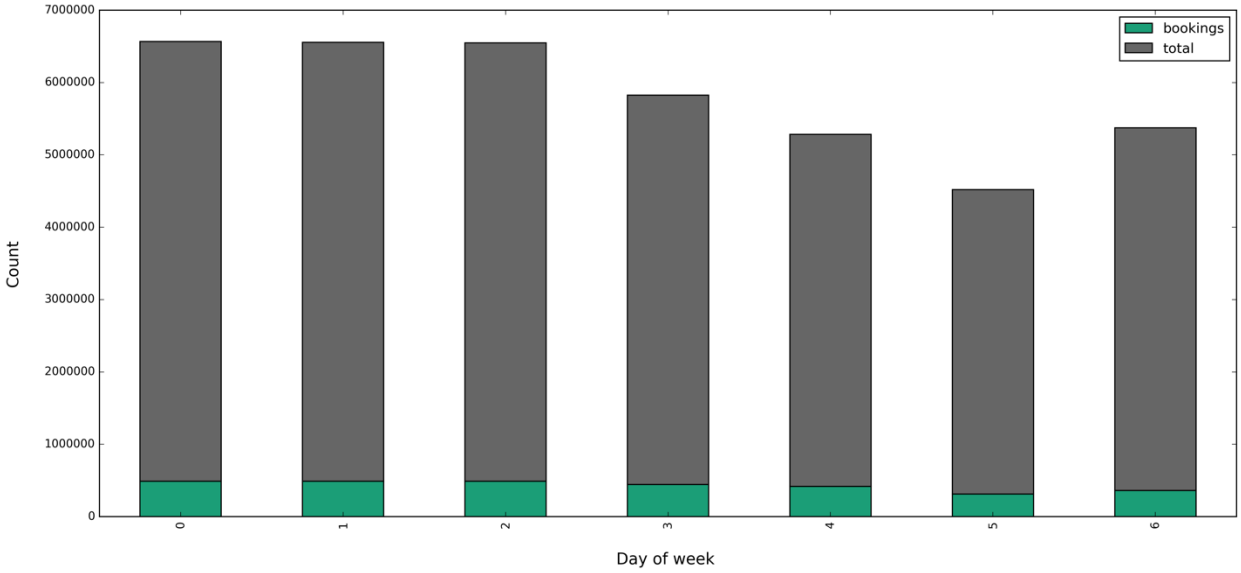


Figure 4: Trend of search and booking event over the days of week

Besides the monthly and daily trend, the location of hotel searched is another interesting finding. Even though the name of the continents is not given, user mostly search for hotels which are in continent 2 and 6 (figure 6).

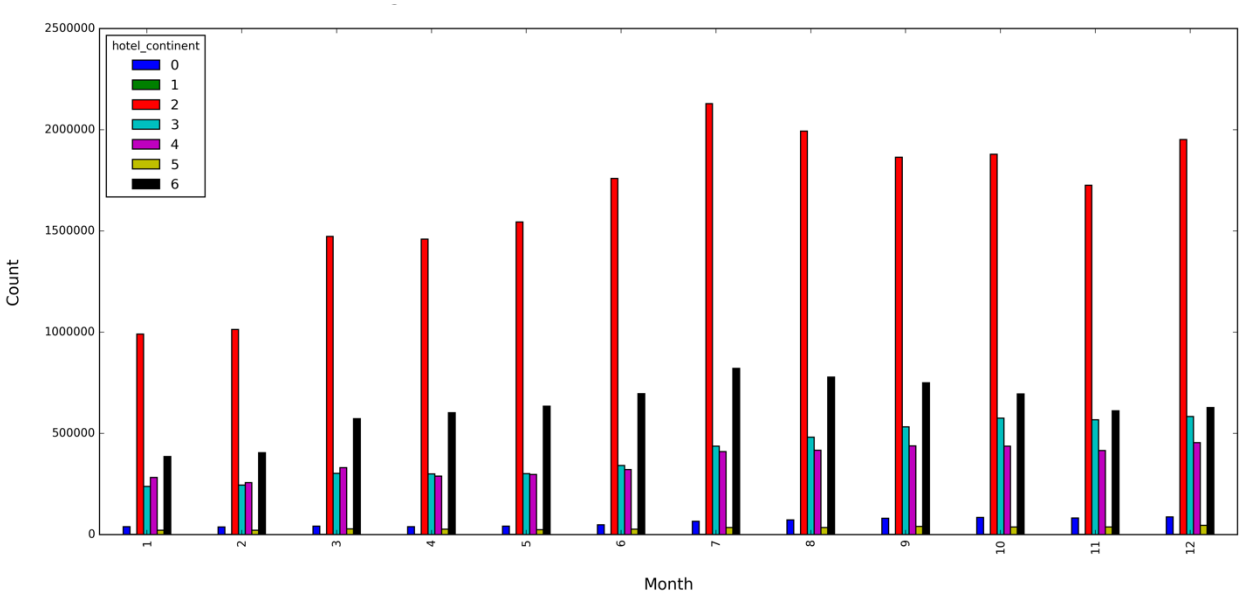


Figure 5: The search trend for hotels in seven continents over months.

4 DATA ANALYSIS

4.1 Approach 1

As the given data set has 37 million rows, it is also hard to load and analyze the entire dataset due to memory outage. For this reason, 1% of the train data is randomly selected and then split into training and testing set in a 70%-30% ratio.

4.1.1 Method

Here, the data set contains 24 attributes, where the id of hotel cluster is considered as target attribute and rest 23 attributes as feature. First of all, correlation among target attribute and feature attributes was checked before training any 'Linear regression' or 'Logistic regression' model. From the correlation coefficient of attributes (Table 2, appendix A) it is clear that, none of the regression techniques will be useful here. For this reason, classification methods are explored to see if these techniques can predict the id of hotel cluster based on user history.

During this analysis, it was checked if any attributes have missing entries. Those attributes along with the timestamp were not considered while the training set was fitting into machine learning model. First, Decision Tree is used for this purpose. The model doesn't give a good accuracy (0.12). Therefore, other metrics like precision, f1 score and recall have been explored. After that,

‘Support vector machine’ classifier was also trained using the training set. The evaluation metrics for both classifiers are given below. Attempts were taken to train using Naïve Bayes. But due to memory shortage of RAM, the process was terminated before it ends.

Table 1: Evaluation metric for classifier used

Evaluation Metric	Decision Tree	Support Vector Machine
Accuracy	0.12	0.036
Precision	0.1089	0.008
F1 Score	0.1112	0.008
Recall	0.1202	0.036

4.1.2 Limitation

A couple of observation can be made from this data analysis.

Down-sampling the data: The training set itself has more than 37 million rows and during preliminary analysis only 1% of the data was used. This could be a reason behind the failure of machine learning models. Besides, there are over 1 million unique users in the given dataset. Rather than randomly sampling the training and testing set, a different approach can be taken to preserve the full data of each user. A sample data set can be taken from the given data set selecting a certain number of users (user set) randomly. Then only picking the rows from the given dataset where that user id is in random user set. Once the sample dataset is picked, it can be split into training and testing set.

Adding new feature: Among the 24 attributes in the training set, the timestamp is an important attribute as it captures the information about the user. However, during this analysis timestamp was not used. Extracting the day of the week, day, month and year from this field will help to train the model in a better way.

Improve model: The machine learning models used during preliminary analysis did not give a good prediction. Multiple reasons might be responsible for this. One of them could be, presence of too many cluster. There are 100 clusters present in the training set. If only most popular cluster can be considered during training model, it could give us a better prediction. Taking the ten most

popular hotel clusters and turning them into binary classification problem might be helpful in better prediction.

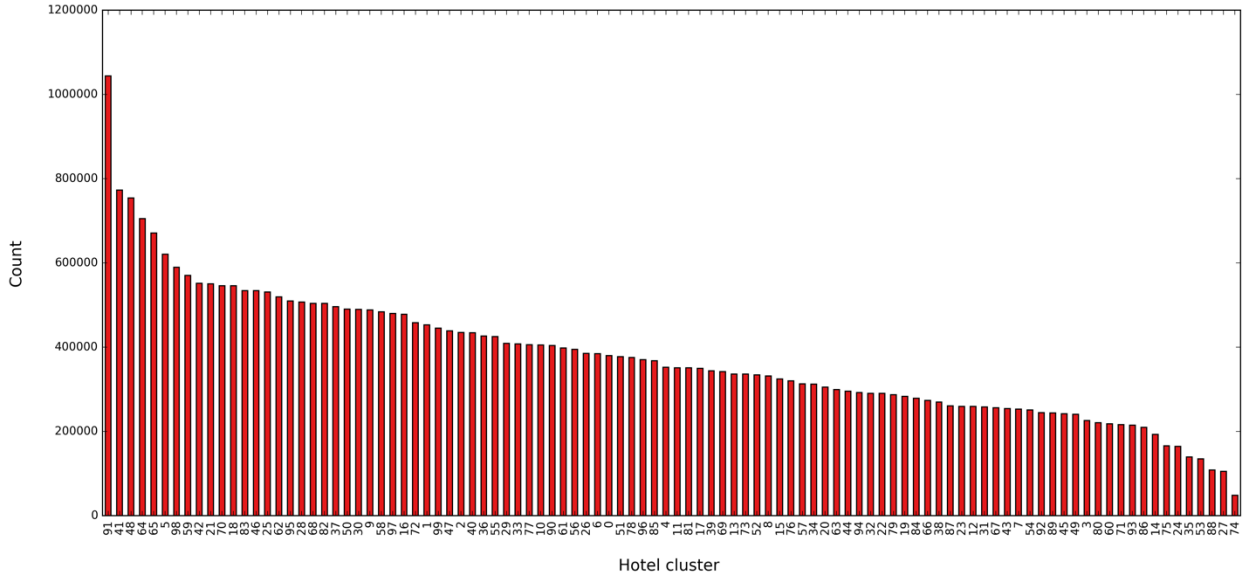


Figure 6: Frequency of hotel cluster

4.2 Approach 2

4.2.1 Data wrangling

A couple of pre-processing steps has been taken before training the model.

- Ten of most popular hotel clusters are chosen for analysis. The id of the clusters is 64, 65, 98, 59, 5, 41, 42, 48, 21 and 91. Due to selecting top ten clusters, the dataset has been reduced to 6.8 million rows from 37 million.
- Four more features such as year, month, day and day of week have been created from the timestamp to utilize this feature more efficient way. Thus the number of feature present in dataset increases to 27 from 23.
- As the given dataset is scaled down to a smaller set, it is now split into training and testing set in 70-30 ratio.

4.2.2 Method

- Earlier, processing time was excessively longer due to larger sample size even though only 3 million rows were considered. To handle this issue, linear classifiers with stochastic gradient descent (SGD) learning has been used.
- The logistic regression model with SGD learning has been trained with 17 features out of 27. Features which has missing data are not considered here.

- During training the logistic regression model, for each hotel cluster the training set was converted into binary classification set. For example, for hotel cluster id 64, the rows with hotel cluster 64 was considered as 1 and rest are 0. And in this way, 10 different classifiers were trained. Then each model was evaluated by metrics such as model accuracy, precision, f1 score, recall, confusion matrix, area under receiver operating characteristics (ROC) curve etc.

4.2.3 Limitation

The model accuracy for each classifier is 80% and higher. The precision and recall is as high as 91% and as low as 88%. The detailed metrics are summarized in table 3, appendix A. The precision-recall curve for all the hotel cluster is given below.

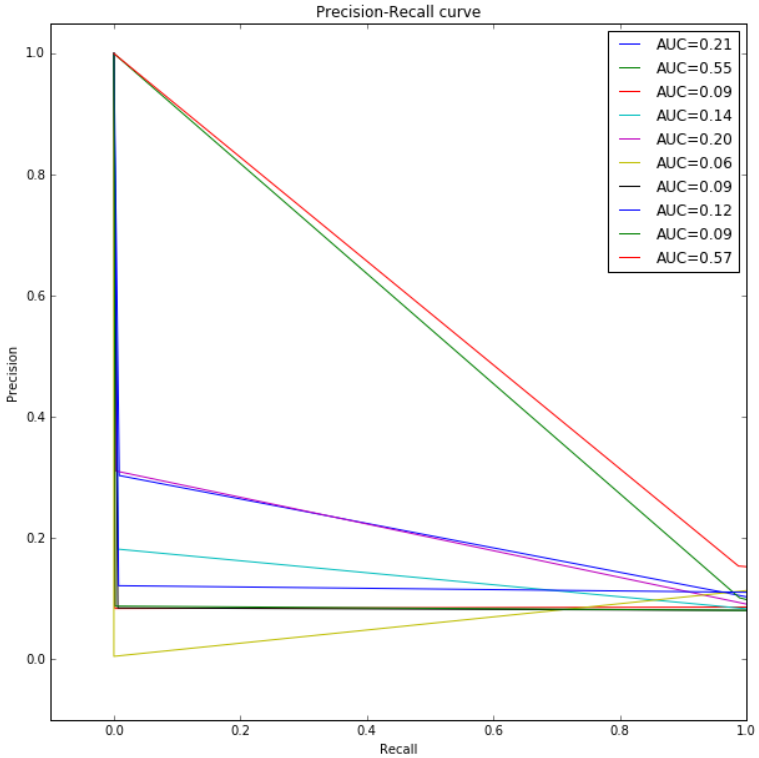


figure 7: Precision-recall curve for all hotel clusters

The area under the ROC curve for cluster 64 is 0.59, rest of the details are attached in appendix A, figure 3.

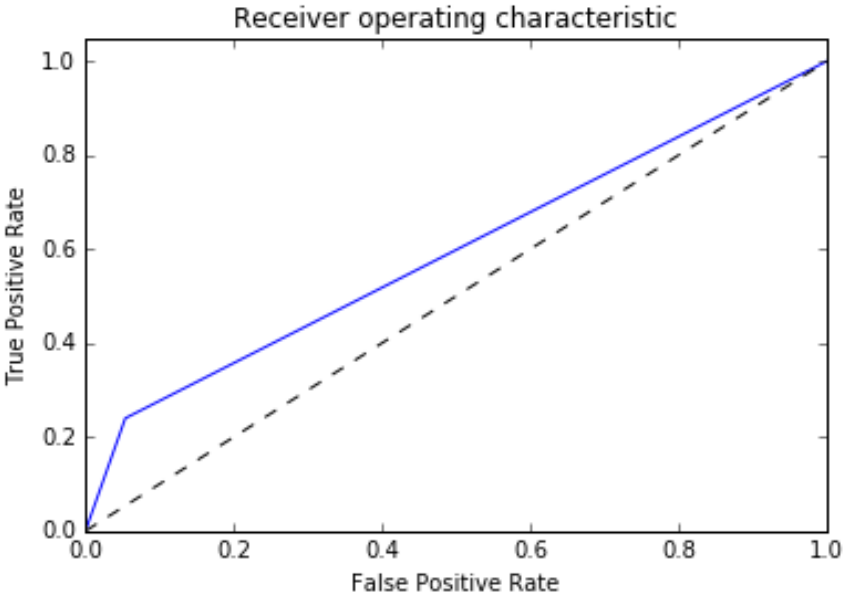


Figure 8: ROC curve for cluster 64

Confusion matrix has been constructed to evaluate how accurately the model can label the positive sample. For example, for cluster 64 the confusion matrix is as below.

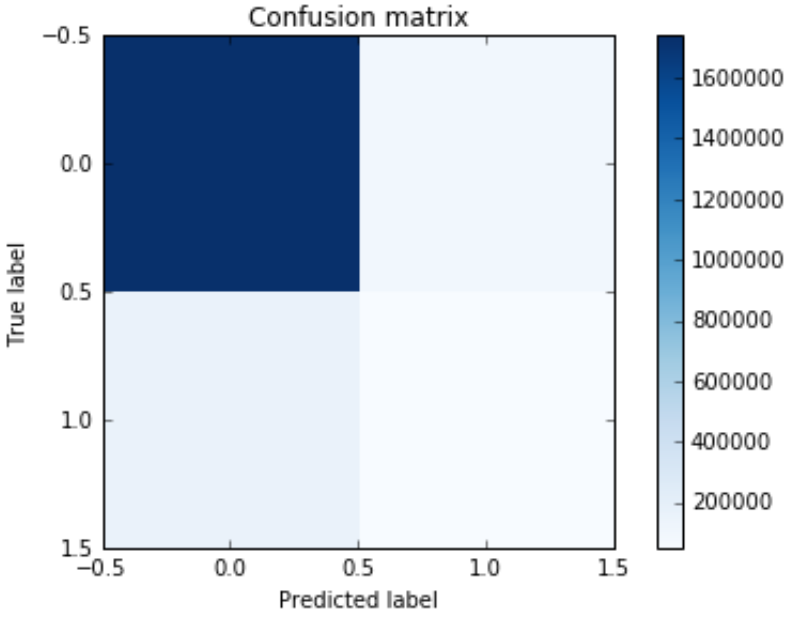


Figure 9: Confusion matrix for cluster 64

4.3 Approach 3

In approach 1 and 2, all the attribute features were considered and used to train the model. After taking a deep look into the features, it is seen that there is a mix of categorical and numerical attributes for which the classifier could not yield the optimum output. For this reason, in this approach categorical and numerical features are treated separately.

4.3.1 Data wrangling

First, categorical feature, numerical feature and features which are transformed from other features are identified. If categorical values are converted to numerical, this would not be able to capture true meaning. For this reason, for each categorical feature, dummy variables are created and then combined together into a singled data frame. But some levels of some categorical features from training set are not present in test set, and similar things have been noticed for test set too. To make both training and test sent consistent, the levels were dropped from training set, which are absent in test set.

Second, numerical features are added to the newly formed data frame containing categorical variables. Thus the training set is ready to train the SGD classifier and test the performance using test set.

4.3.2 Method

The training method in approach 2 is repeated here with the processed training set. Ten logistic regression model with SGD learning are trained for each hotel cluster and the predicted probabilities are estimated for each classifier model for each test row. For evaluation of the models, ROC curve, precision-recall curve and confusion matrix are created.

5 RESULT AND DISCUSSION

Due to treating the categorical and numerical attributes differently, the ROC curve came out than that of approach 2. The model could classify the hotel cluster with a good accuracy which is reflected from the measure of AUC, precision-recall curve and confusion matrix.

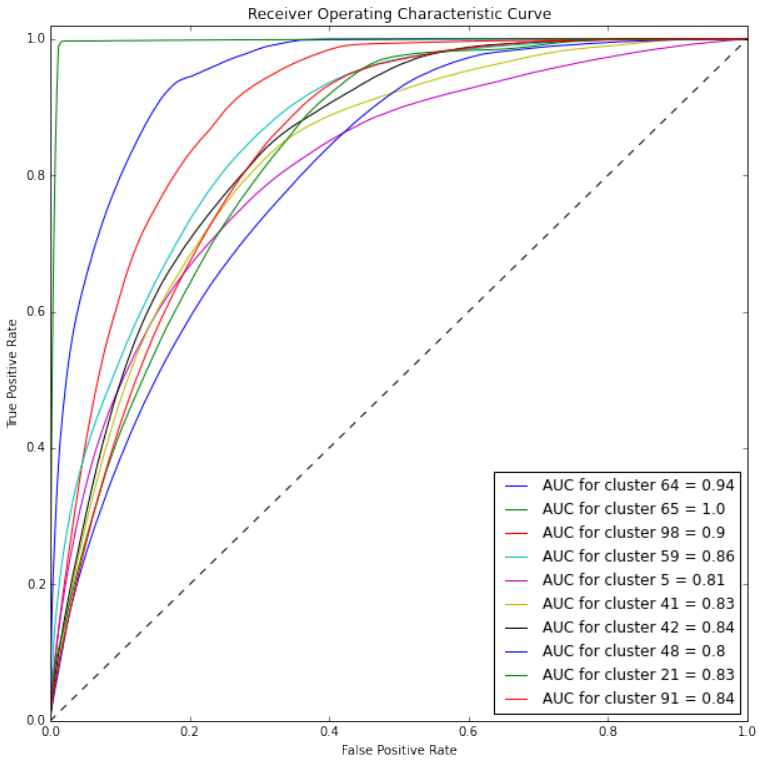


Figure 10: ROC curve for all hotel clusters

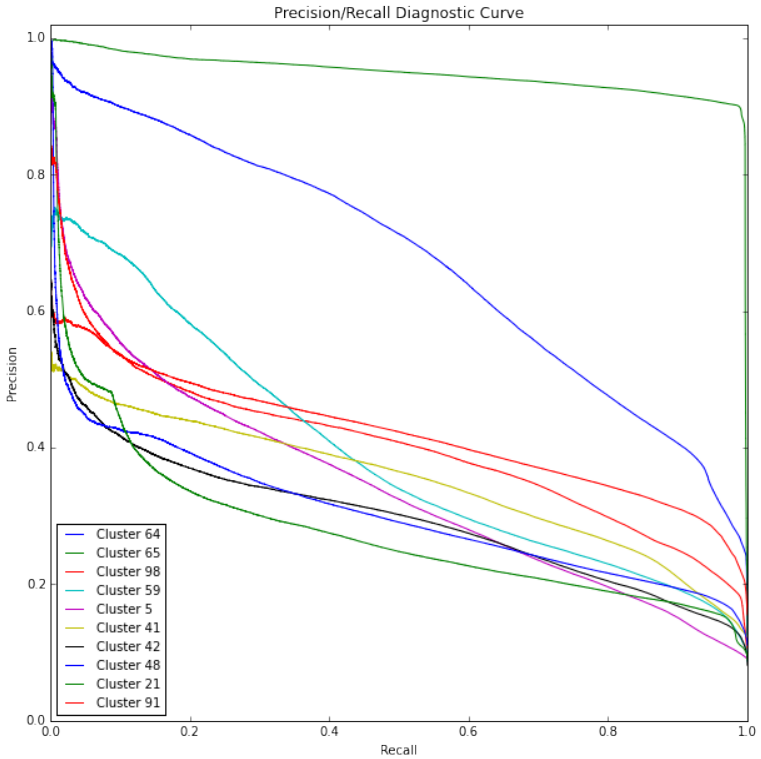


Figure 11: Precision-recall curve for all hotel cluster

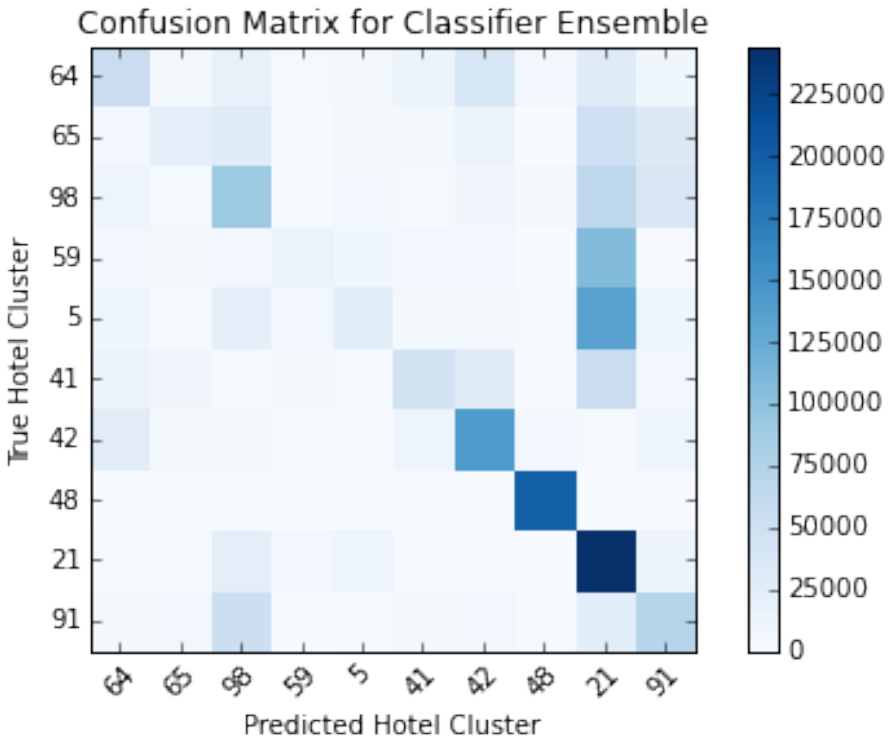


Figure 12: Confusion matrix for all hotel clusters

Appendix A

[1]The dataset can be downloaded from the following link:

(https://www.kaggle.com/c/expedia-hotel-recommendations/data).

Figure 1: The example of training set

	0	1	2
date_time	2014-11-22 22:00:24	2014-10-13 15:25:05	2014-07-10 23:26:18
site_name	30	2	25
posa_continent	4	3	2
user_location_country	195	66	23
user_location_region	991	174	48
user_location_city	47725	46432	4924
orig_destination_distance	NaN	110.51	NaN
user_id	1048	3313	3972
is_mobile	1	0	1
is_package	0	0	0
channel	9	1	9
srch_ci	2015-06-26	2014-10-24	2014-08-13
srch_co	2015-06-28	2014-10-26	2014-08-14
srch_adults_cnt	2	2	2
srch_children_cnt	0	0	1
srch_rm_cnt	1	1	1
srch_destination_id	8803	11835	8278
srch_destination_type_id	1	1	1
is_booking	0	0	0
cnt	1	1	1
hotel_continent	3	2	2
hotel_country	151	50	50
hotel_market	69	633	368
hotel_cluster	59	17	63

Table 1: Field name and description of train / test set

	Field name	Description
1	date_time	Timestamp
2	site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)
3	posa_continent	ID of continent associated with site_name
4	user_location_country	The ID of the country the customer is located
5	user_location_region	The ID of the region the customer is located
6	user_location_city	The ID of the city the customer is located
7	orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated
8	user_id	ID of user
9	is_mobile	1 when a user connected from a mobile device, 0 otherwise
10	is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise
11	channel	ID of a marketing channel
12	srch_ci	Check-in date
13	srch_co	Checkout date
14	srch_adults_cnt	The number of adults specified in the hotel room
15	srch_children_cnt	The number of (extra occupancy) children specified in the hotel room
16	srch_rm_cnt	The number of hotel rooms specified in the search
17	srch_destination_id	ID of the destination where the hotel search was performed
18	srch_destination_type_id	Type of destination
19	hotel_continent	Hotel continent
20	hotel_country	Hotel country
21	hotel_market	Hotel market
22	is_booking	1 if a booking, 0 if a click
23	cnt	Numer of similar events in the context of the same user session
24	hotel_cluster	ID of a hotel cluster

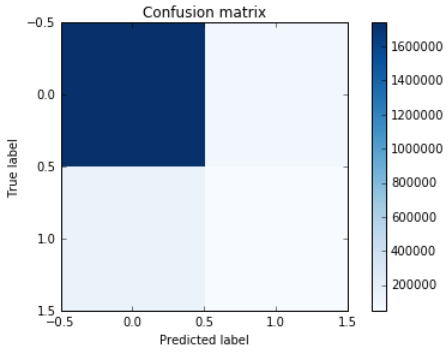
Table 2: correlation check

Field name	**Correlation coefficient**
site_name	-0.022408
posa_continent	0.014938
user_location_country	-0.010477
user_location_region	0.007453
user_location_city	0.000831
orig_destination_distance	0.007260
user_id	0.001052
is_mobile	0.008412
is_package	0.038733
channel	0.000707
srch_adults_cnt	0.012309
srch_children_cnt	0.016261
srch_rm_cnt	-0.005954
srch_destination_id	-0.011712
srch_destination_type_id	-0.032850
is_booking	-0.021548
cnt	0.002944
hotel_continent	-0.013963
hotel_country	-0.024289
hotel_market	0.034205

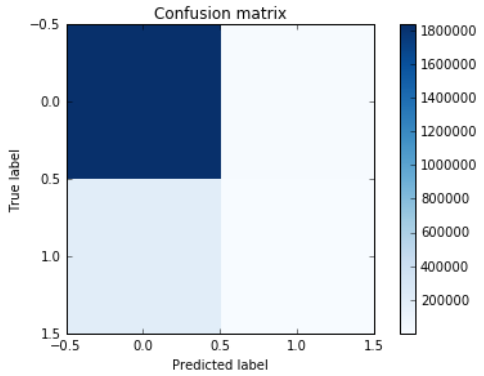
Table 3: Summary of evaluation metrics

Cluster ID	Accuracy in %	Precision	Recall	AUC of ROC
64	87.3	.91	.94	0.59
65	89.6	.90	.99	.50
98	91.2	.91	.98	.49
59	91.6	.91	.998	.50
5	90.8	.90	.998	.50
41	88.6	.88	.999	.49
42	90.8	.91	.986	.50
48	88.9	.88	.995	.50
21	91.9	.91	.999	.50
91	82.6	.84	.968	.50

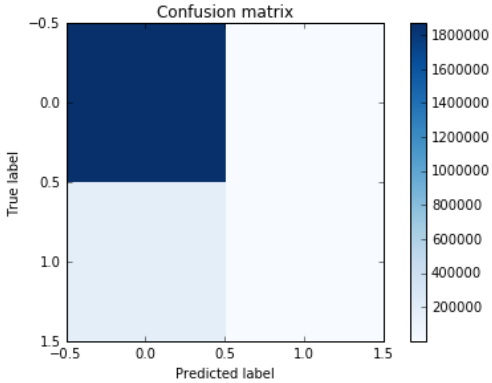
Figure 2: Confusion matrix for hotel cluster



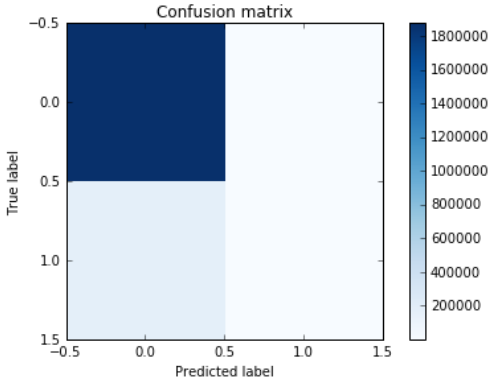
cluster 64



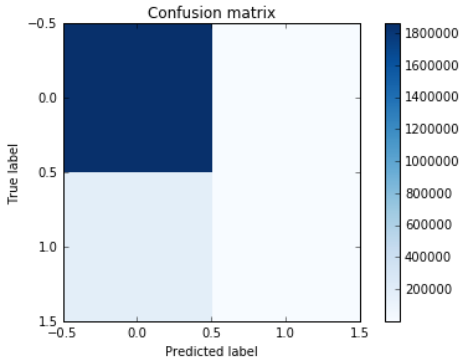
Cluster 65



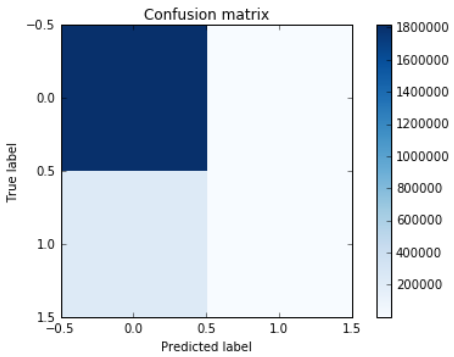
Cluster 98



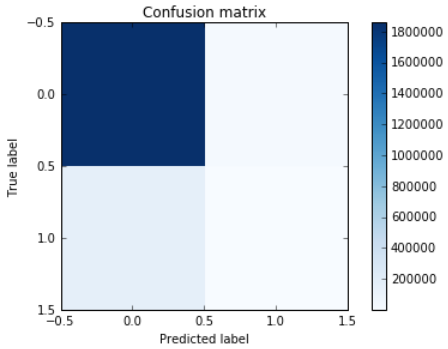
Cluster 59



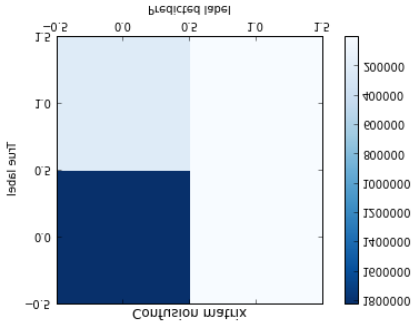
Cluster 5



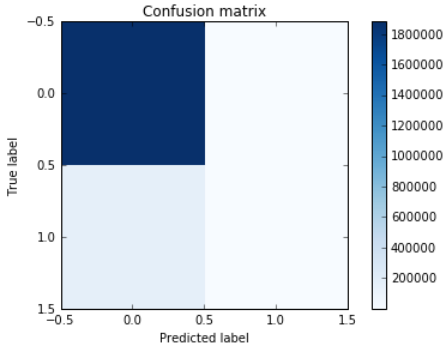
cluster 41



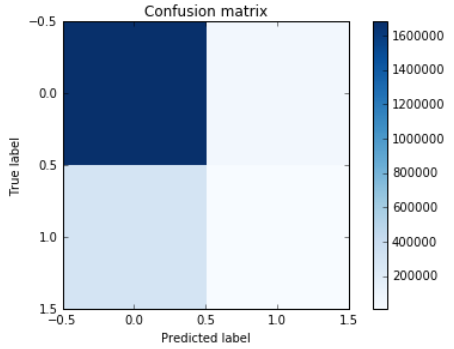
Cluster 42



Cluster 48

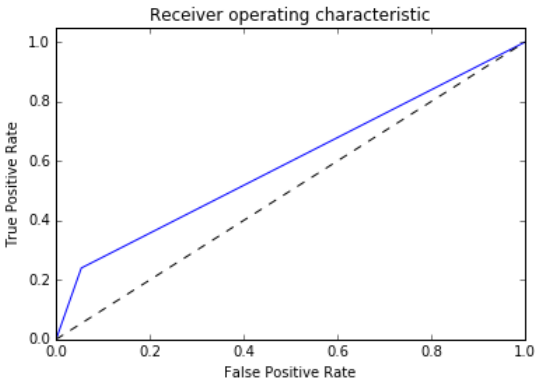


Cluster 21

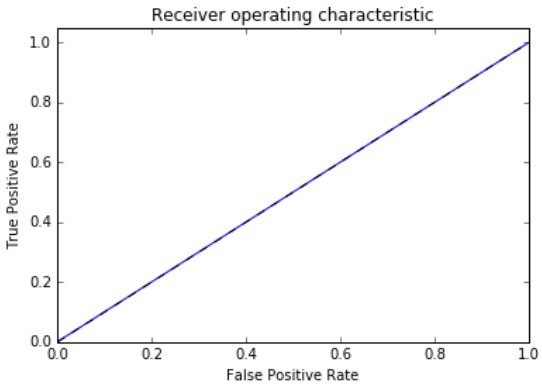


Cluster 91

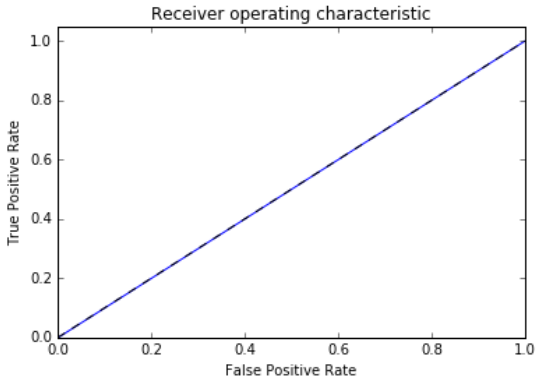
Figure 3: ROC curve for hotel cluster



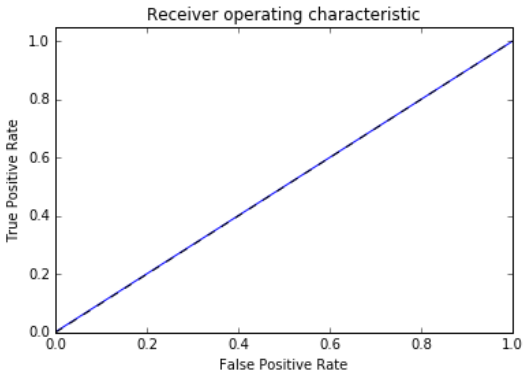
Cluster 64



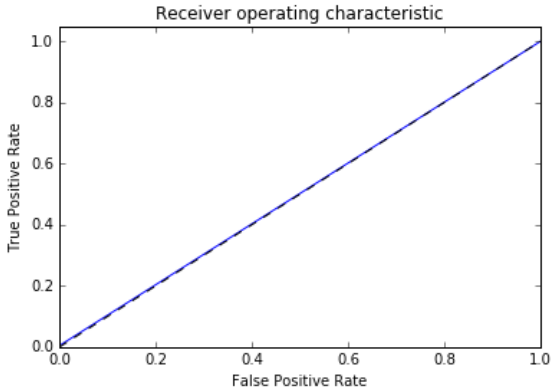
Cluster 65



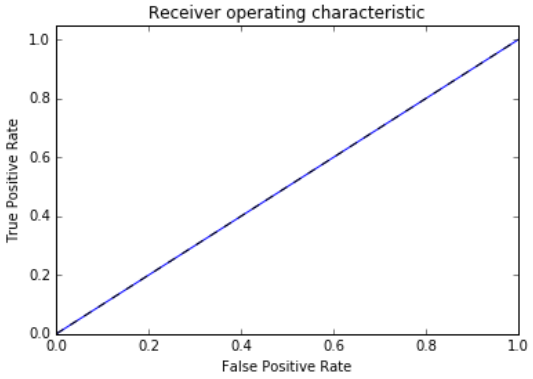
Cluster 98



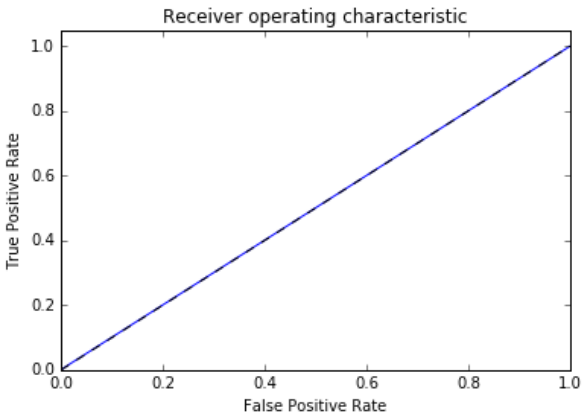
Cluster 59



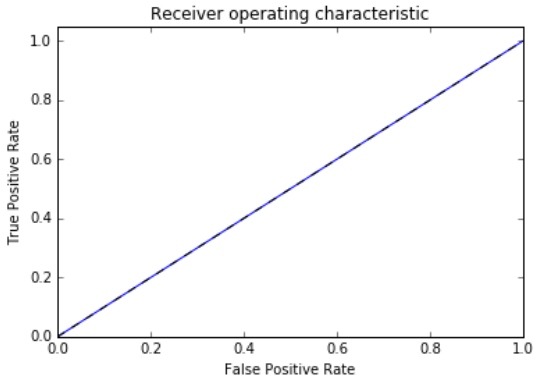
Cluster 5



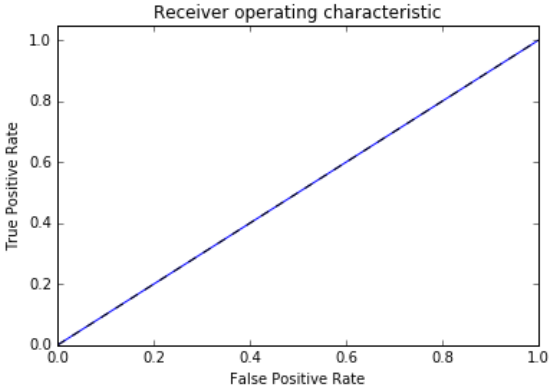
Cluster 41



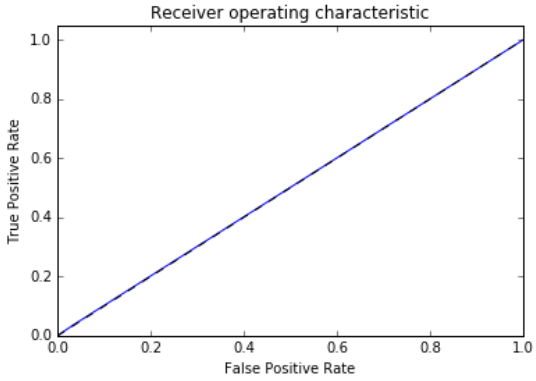
Cluster 42



cluster 48



Cluster 21



Cluster 91