Dublin Business School Assessment Brief

Assessment Details

Unit Title	Data Mining
Unit Code	B9DA103
Unit Leader	Terri Hoare
Level:	9
Assessment Title	Application of Data Mining Tools&Techniques
Assessment Number	1
Assessment Type:	Group (2-3 students)
Assessment Weighting	30%
Issue Date:	Tuesday 12 June 2018
Hand in Date:	Sunday 29 July 2018 (23:55)
Mode of Submission:	On-line Moodle

Student No	**10374210**
Name	**Sanjay Kumar**

• Business Understanding

➢ Business Objectives

Every Business build on the customer requirements and demand which directly means that Business is a direct relationship with customer. To improve business growth, you always need to understand the customer feedback. So, I choose some relevant topic for data mining project named **Sentiment analysis on Europe Hotel Reviews.**

➢ Objectives

There are two main objectives are first is to Improve Business and earn more revenue and second is to find out areas which are highly needed to improve like Rooms, Staff, Services, Breakfast, cleaning and many more.

➢ Success Criteria

For success first, we need to understand the customer reviews and find out that areas or sectors of hotels which not perform well. Improving those areas is our success.

➢ Data mining Goals

Data mining goals are to find great insight from a data and deploy a model which can helps to understand weak areas of each Hotel on user review. Data mining with Text Data is also difficult to handle but try to make good out of it.

➢ Project planning

#No	Dates	Tasks
1	25-06-2018	Business objective and data mining goals
2	02-07-2018	Data understanding
3	09-07-2018	Data preparation
4	16-07-2018	Algorithm and modelling technique
5	06-08-2018	Evaluation
6	10-08-2018	Deployment

Group Member's	Roll Number
Sanjay kumar	10374210
Harsha Bidappa	

- # Data Understanding

➢ Initial Data Report

This Dataset is belonging to the most popular site booking.com which do the booking and get the online feedback from the customer, so they can improve in their suggestions and recommendation to the new customers. So, they publish their dataset on Kaggle to get more insights of their dataset. This dataset is available on Kaggle by name of 515k Europe Hotel Reviews. This dataset reviews collected between two years.

➢ Describe Data Report

Dataset is in a good form. Dataset is an unstructured data which means its without label. It consists of 17 features out of which are very useful for findings like Rating, Positive and Negative Review column. Dataset more of text type data which means we must do text mining with that.

➢ Explore Data

Dataset consist of 515k rows with 17 columns. Dataset have some missing values and duplicate values. Dataset is already pre-processed with removing punctuation, Unicode and white spaces. Dataset have not one column of Review which mostly present in the other datasets while versa in one row it has two columns one is Positive Review and second is Negative Review which make them different and unique from another dataset's that's why me and my team partner choose this Dataset to get a good knowledge of text mining and analysis.

➢ Challenges with Data and mining both

Big Data

The collected dataset was huge in size with 515K reviews. There are around 1500 hotels and each hotel has multiple number of reviews. The reviews had two different column, namely positive review and negative review. Analysing the data was challenging as it consumed more processing time.

- **Unlabelled data**

The available data was unlabelled, and the text containing in positive and negative reviews was not labelled, as few of reviews were randomly given people containing unnecessary junk texts. Manual labelling had to be done, and meaningful tags was given.

- **Stemming, Stop Words removal**

In the process of stemming, suffix of the words has been stemmed, keeping the word meaningful to understand. So, the data size can be reduced keeping the words short and easy to analyse it.
Along with that, stop words, such as 'a', 'an', 'are', 'the' has been removed. Because they are the common word which carries no meaning or value in the process.

- **Missing data and datatypes-**

There are few hotels which proper reviews does not have, which is mentioned as 'no reviews'. It had to be processed. And few reviews had special characters.

- **Use of text mining for TF-IDF Document (Term Frequency Inverse frequency Document)-**

As many important words was repeating, text mining is done for TF-IDF to give weights for most repeating words. As same words reoccur again and again, the frequency of the word increases. Along with that, the weight of the word also increases.

• Data Preparation

➢ Dataset Description

This dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe.

The csv file contains 17 fields. The description of each field is as below:

- Hotel_Address: Address of hotel.
- Review_Date: Date when reviewer posted the corresponding review.
- Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
- Hotel_Name: Name of Hotel
- Reviewer_Nationality: Nationality of Reviewer
- Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- Review_Total_Negative_Word_Counts: Total number of words in the negative review.
- Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- Review_Total_Positive_Word_Counts: Total number of words in the positive review.
- Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers has given in the past.
- Total_Number_of_Reviews: Total number of valid reviews the hotel has.
- Tags: Tags reviewer gave the hotel.
- days_since_review: Duration between the review date and scrape date.
- Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- lat: Latitude of the hotel
- lng: longtitude of the hotel

Note : Dataset link is provided in the appendix

➢ Select Data

we select all variables leaving these variables count of Positive, Negative and reviews. But our more focus on Positive and Negative review more than others. Like other variables Tags and Review Date is crucial as well to find the trend of customer's in particular time of year.

➢ Clean Data

In cleaning process, we first remove all those Neutral Reviews like No Positive, No Negative, Nothing, Nothing at all and Null from Review's. Then Remove the Duplicate row's Data.

➢ Construct Data

We create a label manually by breaking two columns Positive and Negative review in Two rows. what we did you can see in the below figures. As well as in Rapid miner we create unique ID for rows and a Target variable to solve SVM input problem.

Before

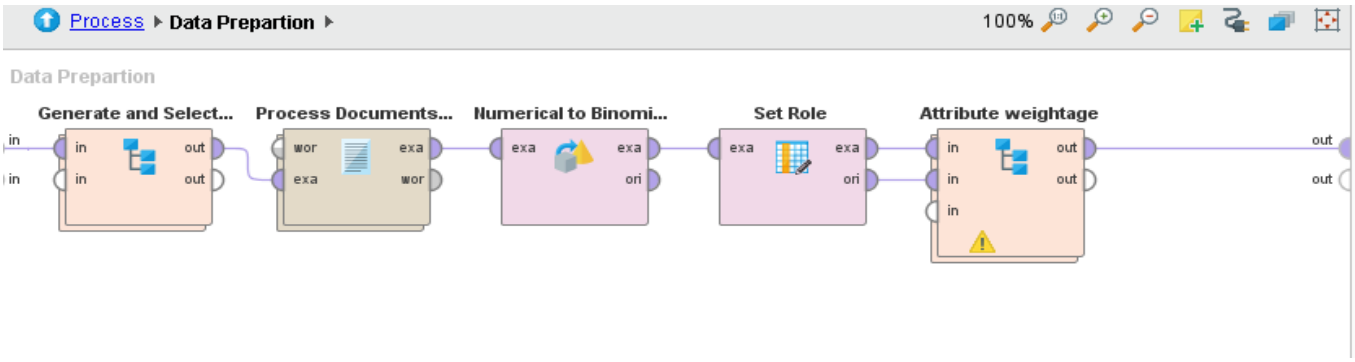
A	B	C	D	E	F	G	H	I	J	K	L	M	N
Hotel_Ad	Additiona	Review_D	Average_	Hotel_Na	Reviewer	Negative_Review	Review_T	Total_Nur	Positive_Review	Review_T	Total_Nur	Reviewer	Tags
s Gravesa	194	########	7.7	Hotel Are	Russia	I am so angry that i made t	397	1403	Only the park outside of the hotel was beautiful	11	7	2.9	[' Leisure t
s Gravesa	194	########	7.7	Hotel Are	Ireland	No Negative	0	1403	No real complaints the hotel was great great loca	105	7	7.5	[' Leisure t
s Gravesa	194	7/31/2017	7.7	Hotel Are	Australia	Rooms are nice but for eld	42	1403	Location was good and staff were ok It is cute hot	21	9	7.1	[' Leisure t
s Gravesa	194	7/31/2017	7.7	Hotel Are	United Ki	My room was dirty and I wa	210	1403	Great location in nice surroundings the bar and re	26	1	3.8	[' Leisure t
s Gravesa	194	7/24/2017	7.7	Hotel Are	New Zeal	You When I booked with y	140	1403	Amazing location and building Romantic setting	8	3	6.7	[' Leisure t
s Gravesa	194	7/24/2017	7.7	Hotel Are	Poland	Backyard of the hotel is tot	17	1403	Good restaurant with modern design great chill o	20	1	6.7	[' Leisure t

After

Hotel_Ad	Review_D	Average_	Hotel_Na	Reviewer	Review	Reviewer	Tags	Label
s Gravesa	########	7.7	Hotel Are	Russia	I am so angry that i made this post available via all p	2.9	[' Leisure trip ', ' Cou	Negative
s Gravesa	########	7.7	Hotel Are	Russia	Only the park outside of the hotel was beautiful	2.9	[' Leisure trip ', ' Cou	Positive
s Gravesa	########	7.7	Hotel Are	Ireland	No real complaints the hotel was great great locatio	7.5	[' Leisure trip ', ' Cou	Positive
s Gravesa	7/31/2017	7.7	Hotel Are	Australia	Location was good and staff were ok It is cute hotel t	7.1	[' Leisure trip ', ' Fam	Positive
s Gravesa	7/31/2017	7.7	Hotel Are	Australia	Rooms are nice but for elderly a bit difficult as most	7.1	[' Leisure trip ', ' Fam	Negative
s Gravesa	7/31/2017	7.7	Hotel Are	United Ki	Great location in nice surroundings the bar and resta	3.8	[' Leisure trip ', ' Solc	Positive
s Gravesa	7/31/2017	7.7	Hotel Are	United Ki	My room was dirty and I was afraid to walk barefoot	3.8	[' Leisure trip ', ' Solc	Negative

➢ Integrate Data

This Part we did in Rapid by integrate data after getting TF-IDF (Term Frequency -Inverse Document Frequency). In the last stage to know about the results.



• Modelling

➢ Modelling Techniques

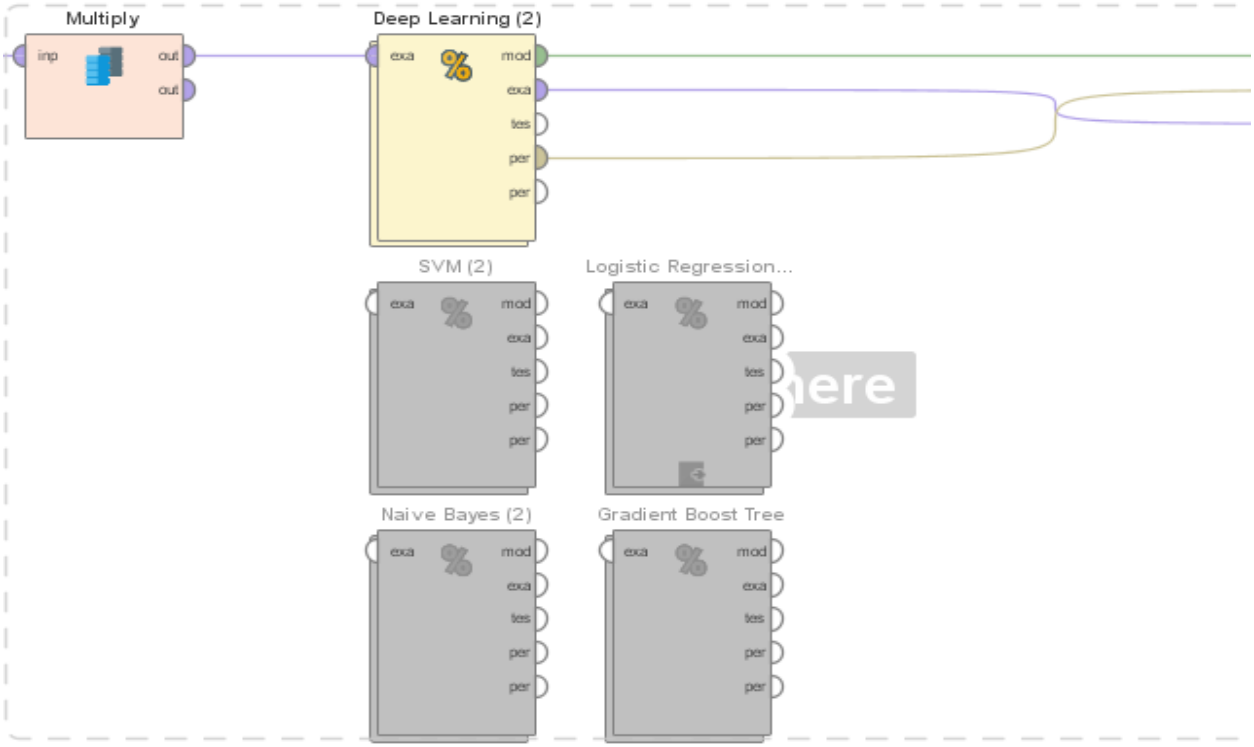
As our data we must classify the label into Positive and Negative. Its a classification Processes, so we choose serval Classification Algorithms to implement it. Reasons I choose these algorithms are: First they take binomial label, second, they classify most times correctly. Gradient Boost Tree help me out to find more weightage words and connected words with it also. Deep Learning is a black box but gives a good accuracy. SVM is used for high Dimensional space means with many numbers of features well. Naive Bayes probability Based Model which performs well on all type of data but not give such good insight.

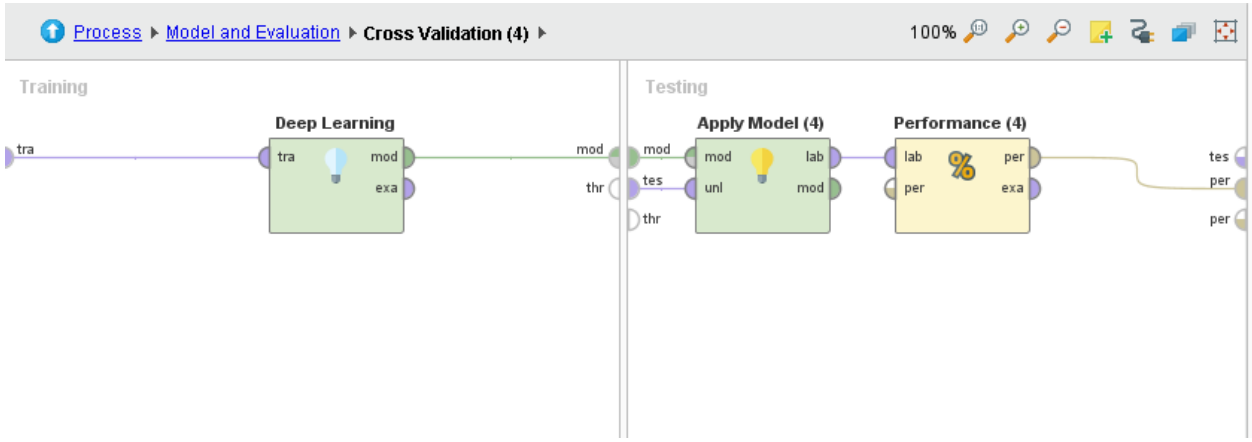
#No	**Algorithms**	**Cost and Memory**	**Time**
1	Gradient Boost Tree	High	High
2	Deep Learning	medium	High
3	Random Forest Tree	high	medium
4	Generalized Linear Model	medium	medium
5	Support Vector Machine (SVM)	medium	high
6	Naive Bayes	medium	high

➢ Build Model

We build Models using Rapid Miner and Python Language. Before building we must process through text mining and we need the output in TF-IDF (Term Frequency Inverse Document Frequency). So, then we can pass the TF-IDF of top 1000 word which selected based on chi-squared Error. Then we pass them into the above algorithms to know which model is best for our Business.

Rapid Miner


Multiply
inp
out
out
Deep Learning (2)
exa
mod
exa
tes
per
per
SVM (2)
Logistic Regression...
Naive Bayes (2)
Gradient Boost Tree


Process
Model and Evaluation
Cross Validation (4)
100%
Training
Testing
Deep Learning
tra
mod
exa
Apply Model (4)
mod
tes
thr
unl
lab
Performance (4)
lab
per
exa
tes
per

/

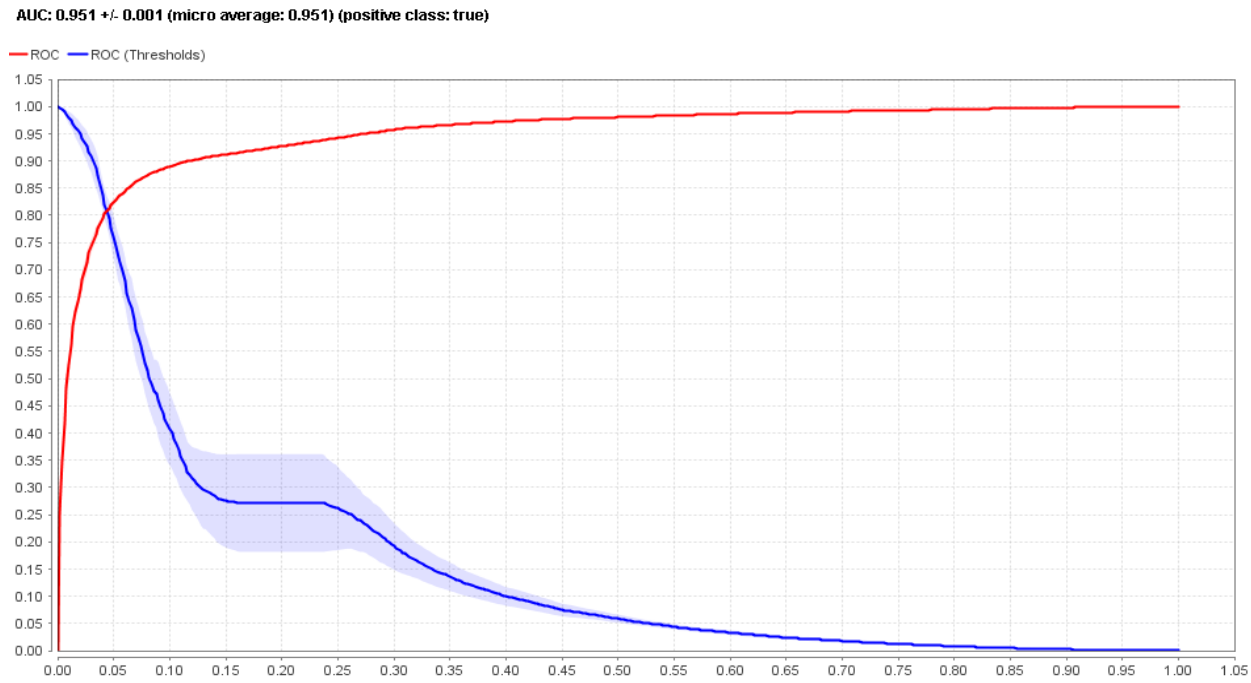
Python

```
In [136]: #testdata
          test=count_vect.transform(X_test.astype('U'))
          test.shape

Out[136]: (284516, 1000)

In [139]: #Navie Bayes Algorithim
          clf_fit=clf.fit(tfidf,Y_train)
          #just passing same data to check accuracy on same
          print(clf_fit.score(tfidf,Y_train))
          #Predict
          predict=clf_fit.predict(test)
          #Accuracy
          np.mean(predict==Y_test)

          0.8600126027435203

Out[139]: 0.8639268090371016

In [141]: #Random Forest
          clf_fit_1=clf1.fit(tfidf,Y_train)
          #just passing same data to check accuracy on same
          print(clf_fit_1.score(tfidf,Y_train))
          #Predict
          predict=clf_fit_1.predict(test)
          #Accuracy
          np.mean(predict==Y_test)

          0.8272350827141601

Out[141]: 0.8206005989118362
```

Asses Models

#No	Algorithms	**Accuracy**
1	Gradient Boost Tree	**81.48%**
2	Deep Learning	**89.34%**
3	Random Forest Tree	**82 %**
4	Generalized Linear Model	**57%**
5	Support Vector Machine (SVM)	**77.13%**
6	Naive Bayes	**79.52% & (86 % Python)**

All Models got good Accuracy and but compare to all Deep Learning win the race with good accuracy rate. Its good to try different algorithms so it helps us to get clear image deep learning is the best.



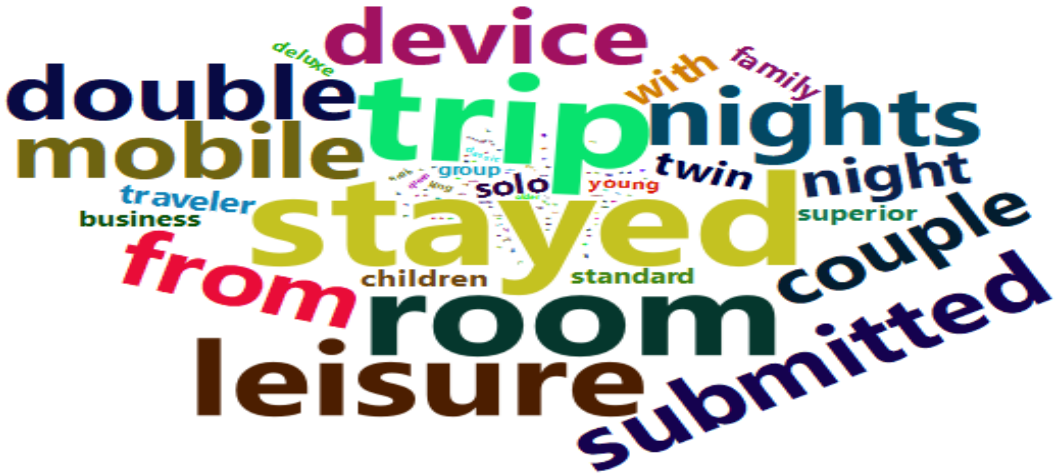
this Deep learning graph of Area under curve graph clearly explains the true positive and negative and also prove the result and the accuracy of Data.

- **Evaluation**
  - **Evaluate Results**

Results came very good in a way of identifying the Area's as well as the good accuracy with it. The Area's which perform bad in a negative Reviews are in the word cloud. We find Deep learning a suitable Model to tell them to improve business from previous stages and its meet's all requirements.

Negative Review

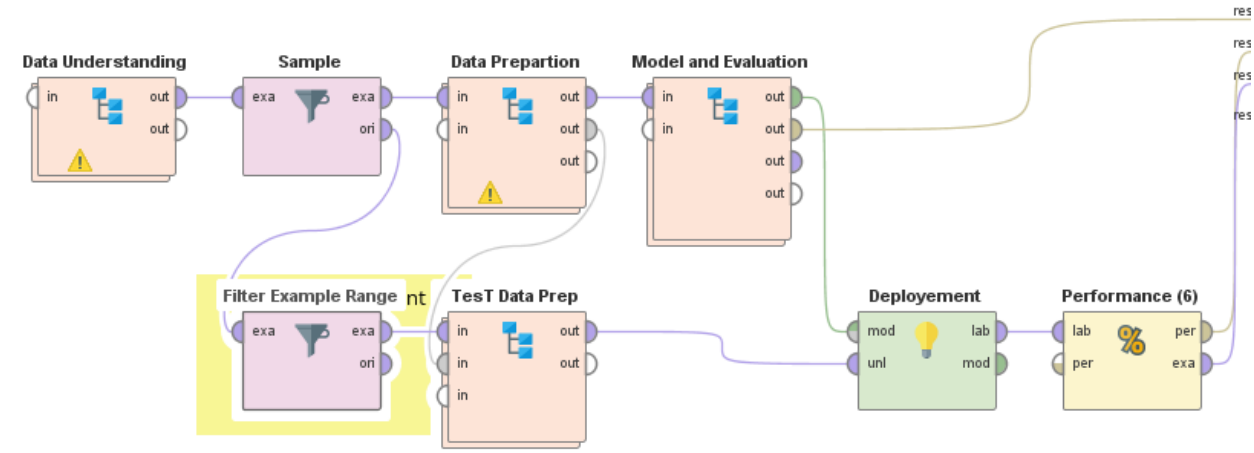
Hotel Booking Tags



- # Deployment

➢ Deployment Plan

We just deploy this project on test data and it performs very well with almost same accuracy. But this model is very useful for online booking websites. As the data is already belong to one famous site. For further process we can build recommender system also.



Deployment with Unseen Data

ExampleSet (20001 examples, 4 special attributes, 29610 regular attributes)

Filter (20,001 / 20,001 examples): all

Row No.	Target	prediction(T...	confidence(f...	confidence(t...	Hotel_Addre...	Review_Date	Average_Sc...	Hotel_Name	Reviewer_N...	Reviewer_S.
1	true	true	0.039	0.961	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
2	false	false	0.806	0.194	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
3	true	true	0.006	0.994	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	8.800
4	false	false	0.999	0.001	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	8.300
5	false	false	0.970	0.030	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
6	true	true	0.096	0.904	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
7	true	true	0.004	0.996	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	9.600
8	false	false	0.967	0.033	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	United Kingd...	9.600
9	true	true	0.024	0.976	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	Bosnia and ...	7.900
10	false	false	0.591	0.409	7 Pepys Stre...	Aug 2, 2017 1...	8.700	DoubleTree b...	Bosnia and ...	7.900
11	true	true	0.224	0.776	7 Pepys Stre...	Aug 1, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
12	false	false	0.744	0.256	7 Pepys Stre...	Aug 1, 2017 1...	8.700	DoubleTree b...	United Kingd...	10
13	true	true	0.031	0.969	7 Pepys Stre...	Jul 29, 2017 ...	8.700	DoubleTree b...	United Kingd...	7.500
14	false	true	0.157	0.843	7 Pepys Stre...	Jul 29, 2017 ...	8.700	DoubleTree b...	United Kingd...	7.500
15	true	true	0.002	0.998	7 Pepys Stre...	Jul 29, 2017 ...	8.700	DoubleTree b...	Australia	10
16	false	true	0.185	0.815	7 Pepys Stre...	Jul 28, 2017 ...	8.700	DoubleTree b...	United Kingd...	10

Test Data Result

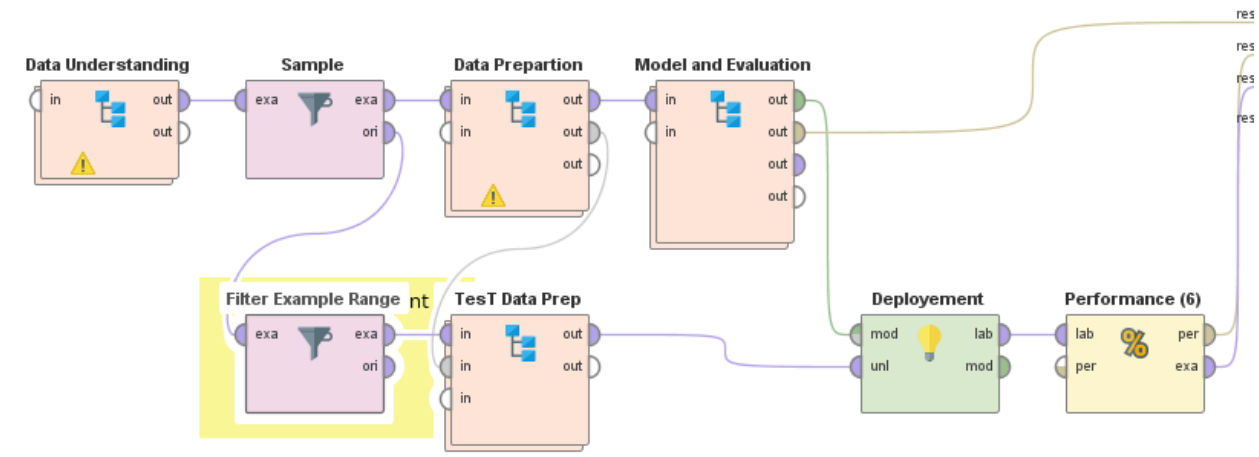
accuracy: 90.48%

	true false	true true	class precision
pred. false	8367	1312	86.44%
pred. true	593	9729	94.25%
class recall	93.38%	88.12%	

- **Appendix**

Dataset link

https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe



Deep Learning Result

accuracy: 89.34% +/- 0.10% (micro average: 89.34%)

	true false	true true	class precision
pred. false	19981	2934	87.20%
pred. true	2397	24688	91.15%
class recall	89.29%	89.38%	

Generalized Linear Model

accuracy: 57.01%

	true Negative	true Positive	class precision
pred. Negative	16966	14016	54.76%
pred. Positive	60121	81331	57.50%
class recall	22.01%	85.30%	

Naive Bayes

accuracy: 79.52% +/- 0.59% (micro average: 79.52%)

	true false	true true	class precision
pred. false	14149	2011	87.56%
pred. true	8229	25611	75.68%
class recall	63.23%	92.72%	

SVM

accuracy: 77.13% +/- 0.37% (micro average: 77.13%)

	true false	true true	class precision
pred. false	21934	10991	66.62%
pred. true	444	16631	97.40%
class recall	98.02%	60.21%	

Gradient Boost Tree

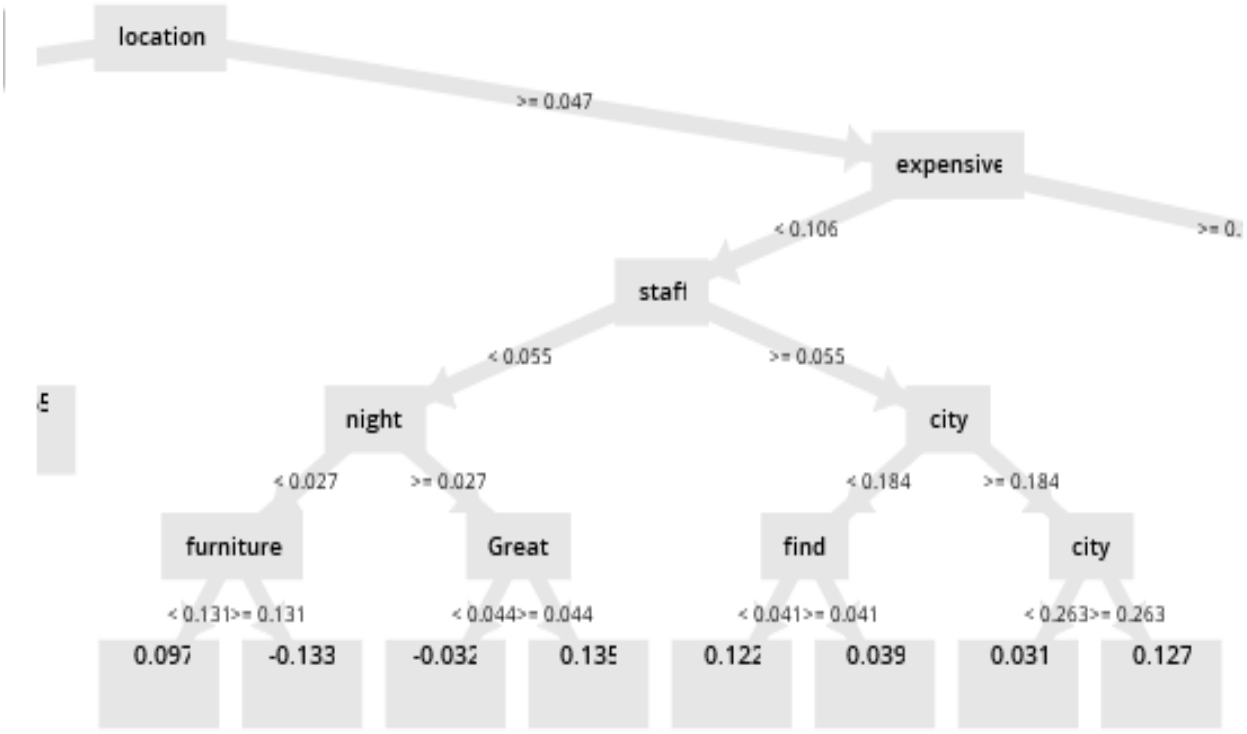


Table View Plot View

accuracy: 81.48% +/- 0.22% (micro average: 81.48%)

	true false	true true	class precision
pred. false	18557	5441	77.33%
pred. true	3821	22181	85.30%
class recall	82.93%	80.30%	