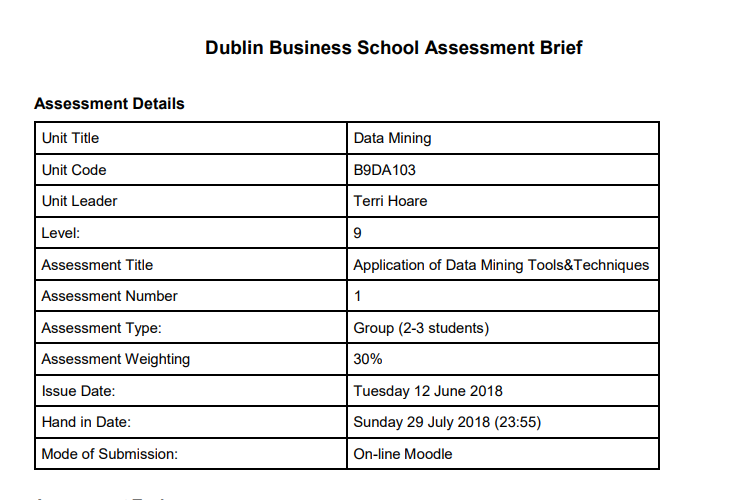
|  |  |
| --- | --- |
| Student No | **10374210** |
| Name | **Sanjay Kumar** |



* **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**

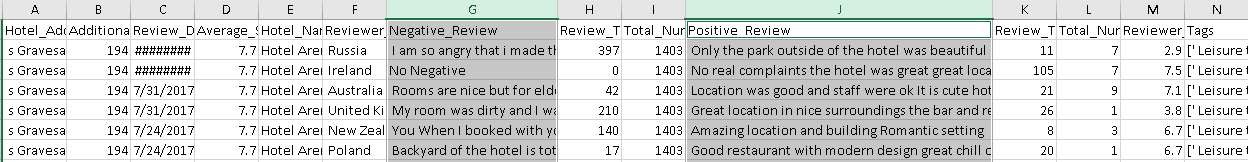
we select all variables leaving these variables count of Positive, Negative and reviews. But our more focus on Positive and Negative review more than others. Like other variables Tags and Review Date is crucial as well to find the trend of customer’s in particular time of year.

* + **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.

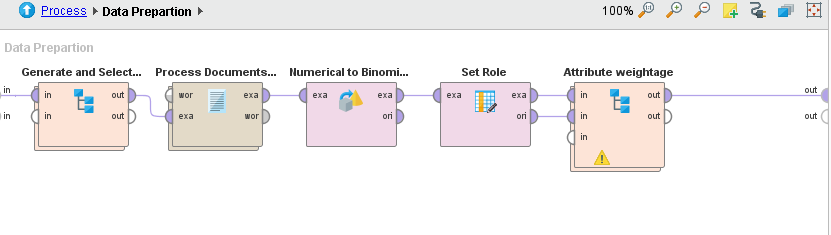
* + **Construct Data**

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**

**After**

* + **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.

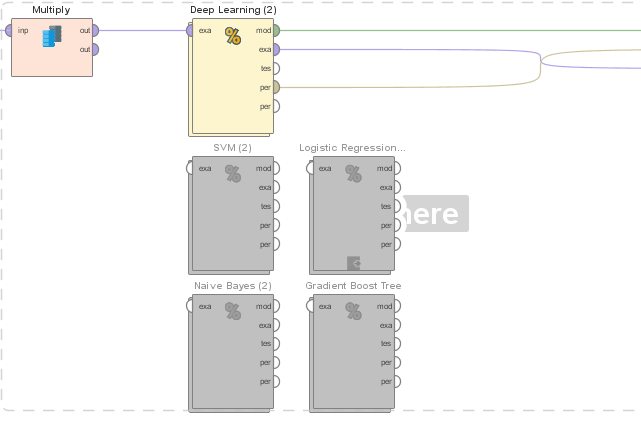
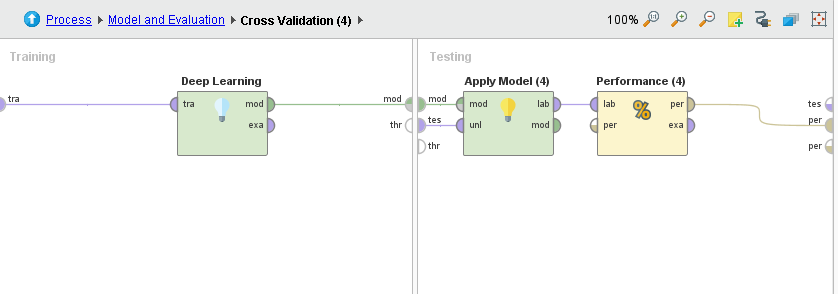
* **Modelling**
  + **Modelling Techniques**

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**

/

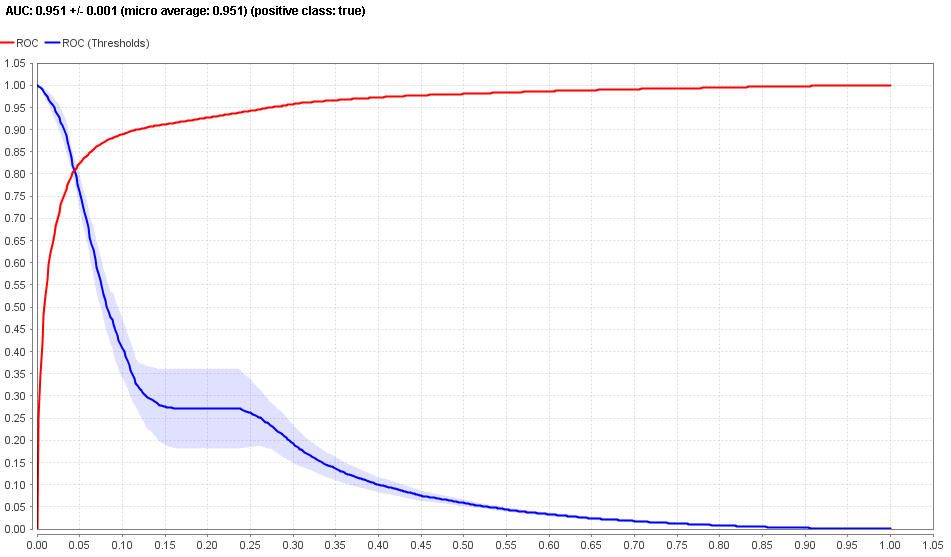
**Python**



**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.

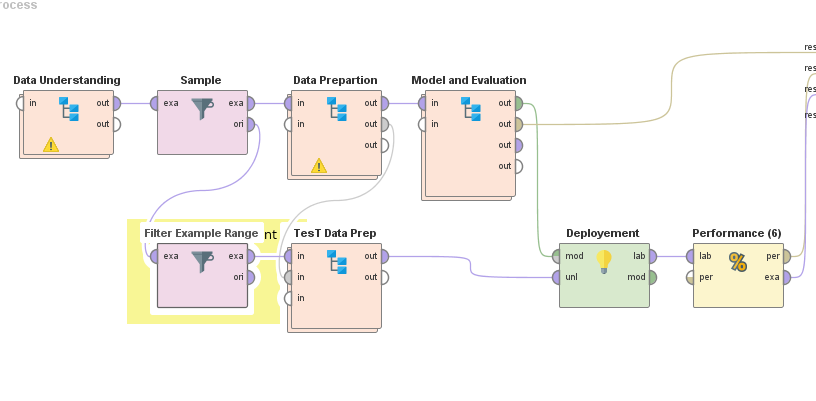
**Negative Review**



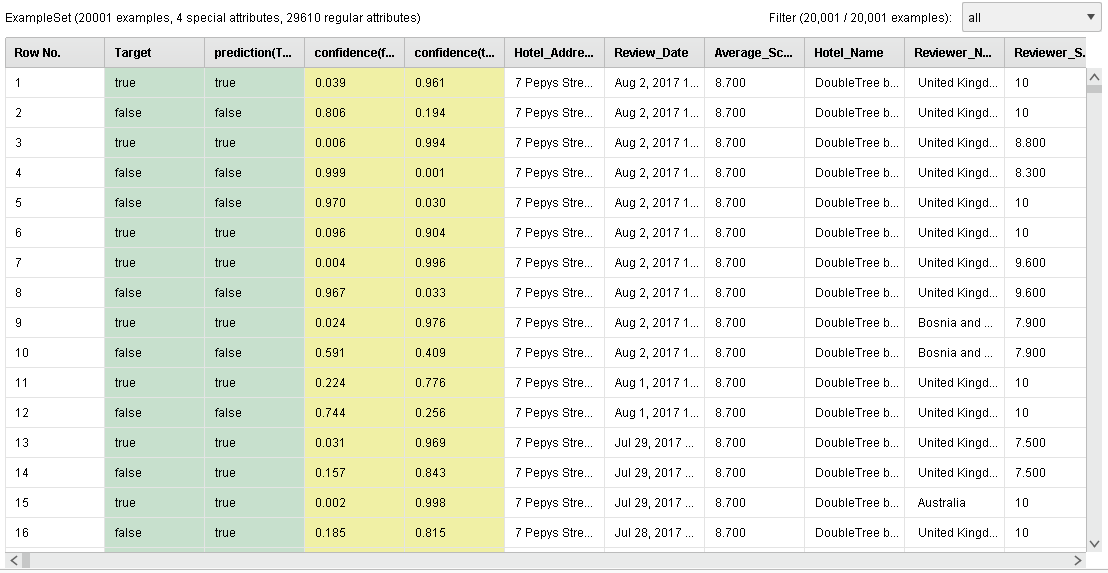
**Hotel Booking Tags**

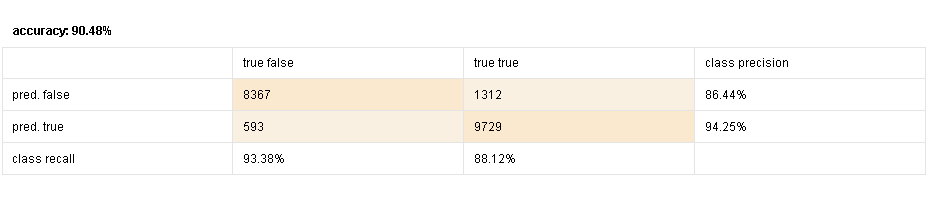
****

* **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**

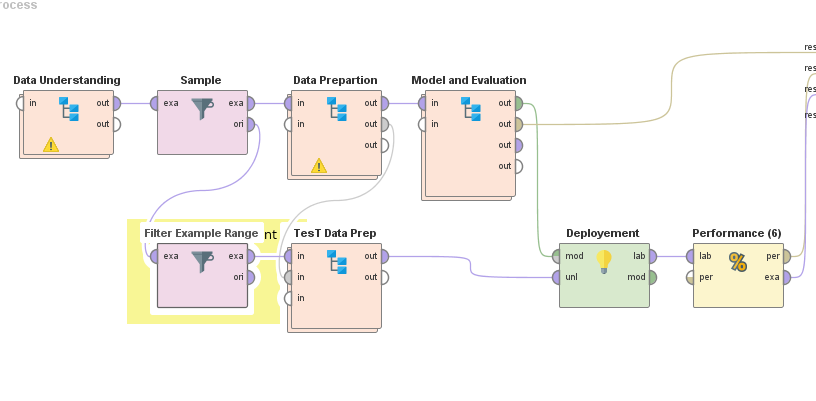
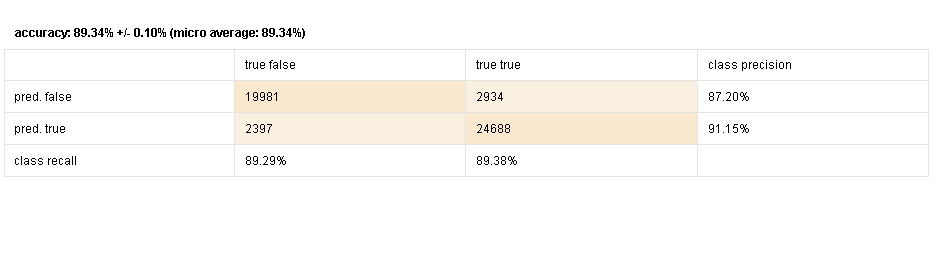
****

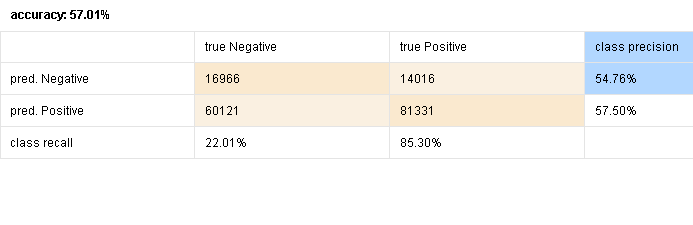
**Test Data Result**

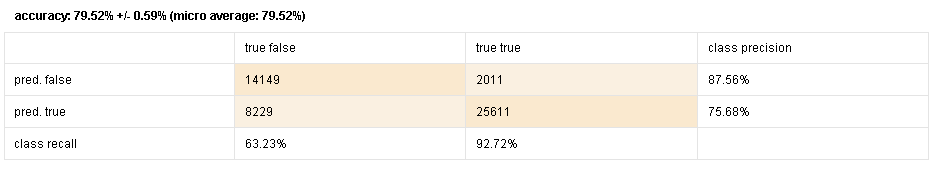
* **Appendix**

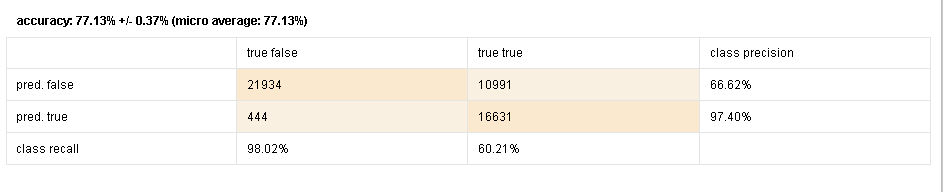
**Dataset link**

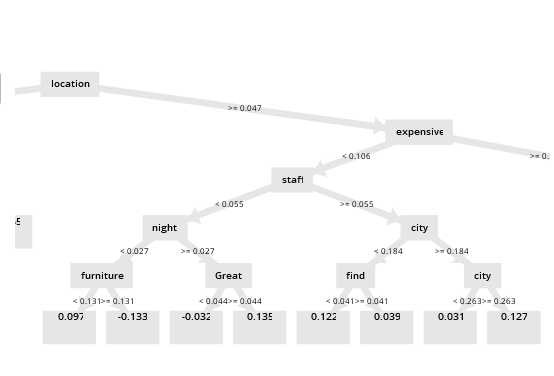
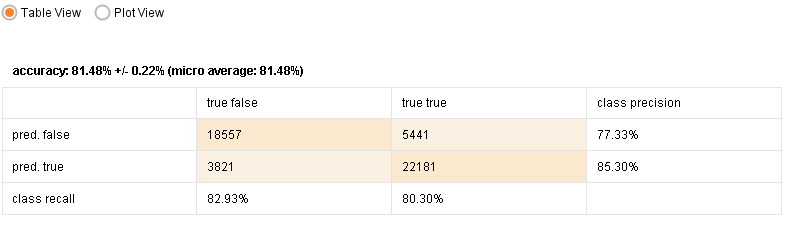
[**https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe**](https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe)

**Deep Learning Result**

**Generalized Linear Model**

**Naive Bayes**

**SVM**

**Gradient Boost Tree**