Experiment No. 9

Title: Case study: Big data platform / analytics as business need)

Batch: B1 **Roll No.: 1714127** **Experiment No.:9**

Title: Case study: Coursera Review using Text Analysis

Resources needed: Microsoft Azure ML Studio, Kaggle Dataset, GitHub Account and Windows Operating System

---

Describe the following points with respect to the business under consideration,

Links –

1. Azure Studio - https://gallery.azure.ai/Experiment/Coursera-Review-Analysis-using-Text-Analysis

2. GitHub - https://github.com/Aagam1090/Coursera-Review-Analysis.git

Aim –
To Predict the Sentiment of Review posted on Coursera for a particular Course using Text Analysis

Need of Solution –
There are multiple review posted on Coursera for various different course in order to determine which course is better we need to go through all the various reviews. On average there are more than 5000+ reviews on a particular course and thus in order to determine if a review is positive or negative we need to read through all of them.

Proposed Solution –
Our Solution takes various Reviews from Coursera which are available on Kaggle Dataset and then based on Textual Analysis and Two Class Logistic Regression. Accuracy obtained using this method was about 90.9% when tested on 30000 reviews and the model was trained on 65000 reviews.

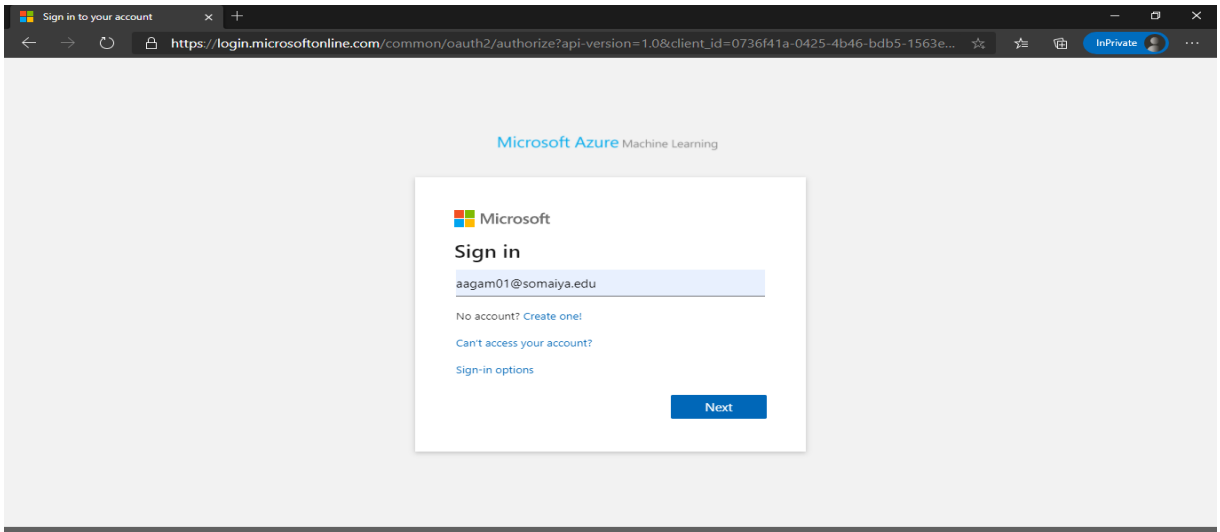
Dataset –
We have used "**100k coursera course reviews dataset"** from Kaggle Datasets (https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset).
The dataset has 3 different columns as

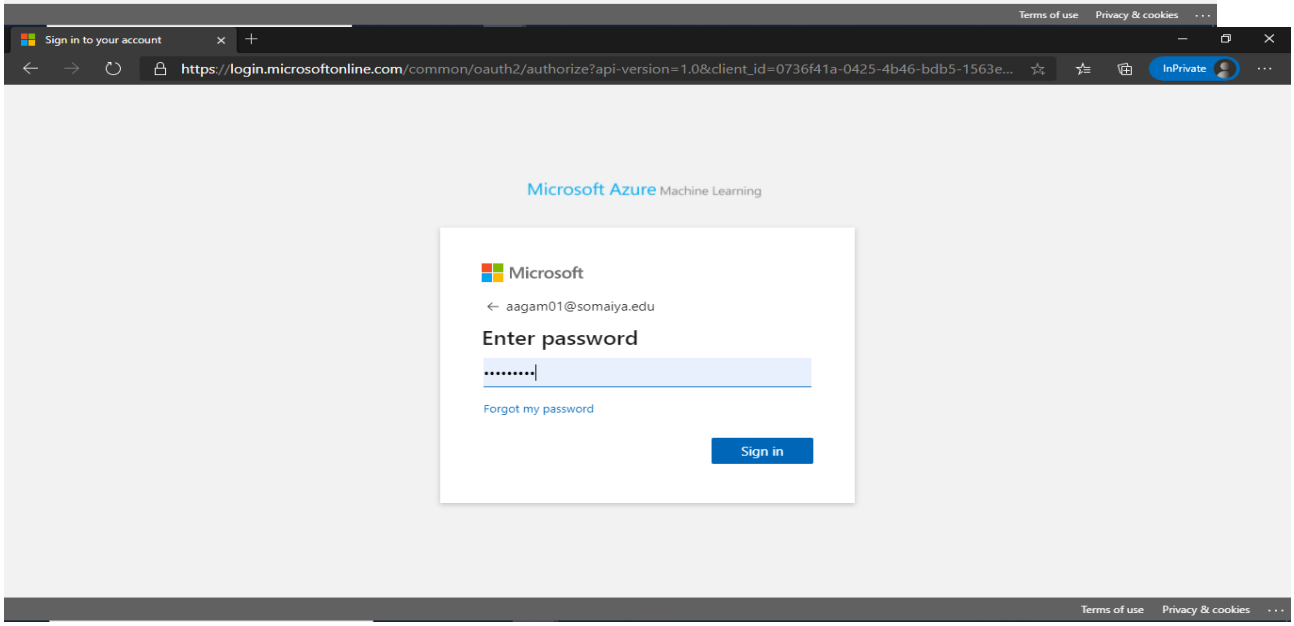
1. Id - The unique identifier for a review.
2. Review - The actual course review.
3. Label - The rating of the course review.

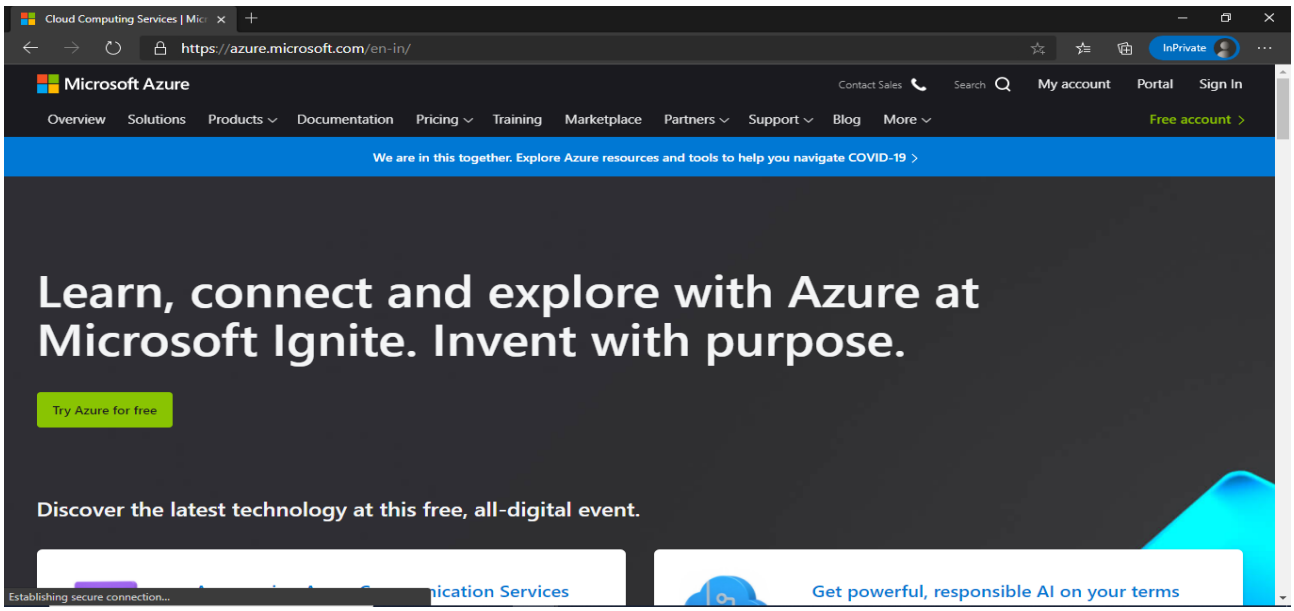
Azure Machine Learning Studio -
Azure Machine Learning Studio is web-based integrated development environment (IDE) for developing data experiments. It is closely knit with the rest of Azure's cloud services and that simplifies development and deployment of machine learning models and services.

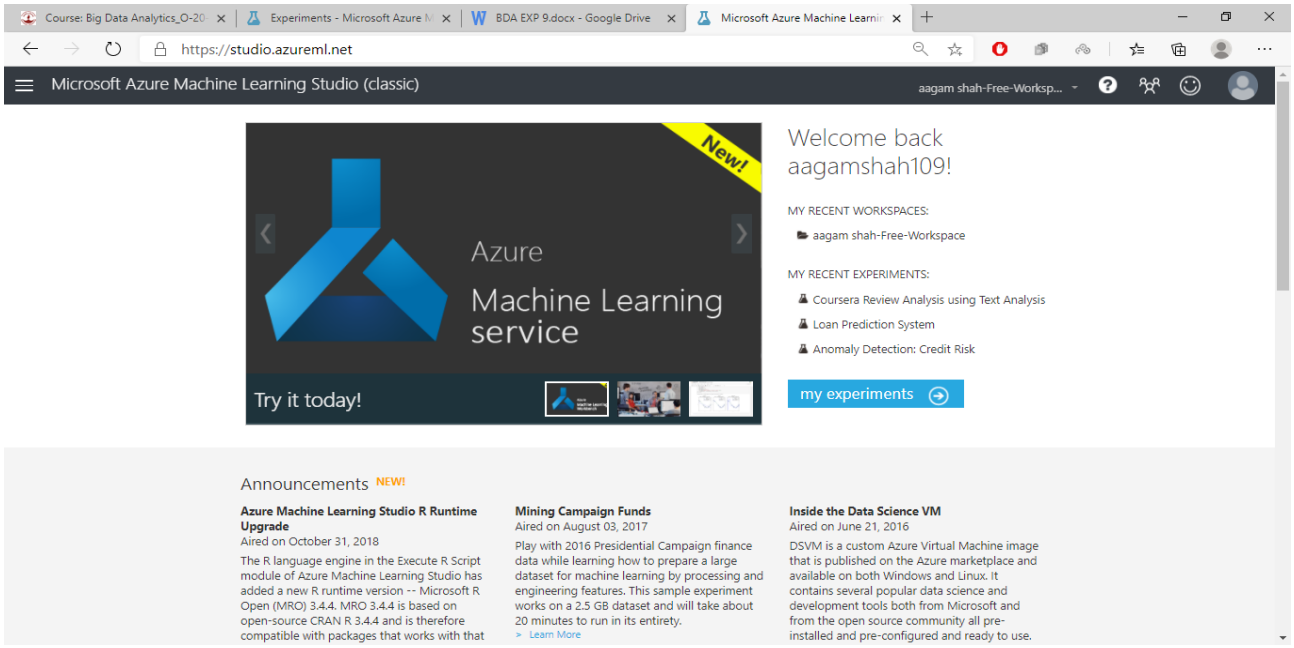
Implementation –

1. We setup our account for Microsoft Azure Studio


Sign in to your account
https://login.microsoftonline.com/common/oauth2/authorize?api-version=1.0&client_id=0736f41a-0425-4b46-bdb5-1563e...
InPrivate
Microsoft Azure Machine Learning
Microsoft
Sign in
aagam01@somaiya.edu
No account? Create one!
Can't access your account?
Sign-in options
Next
Terms of use
Privacy & cookies

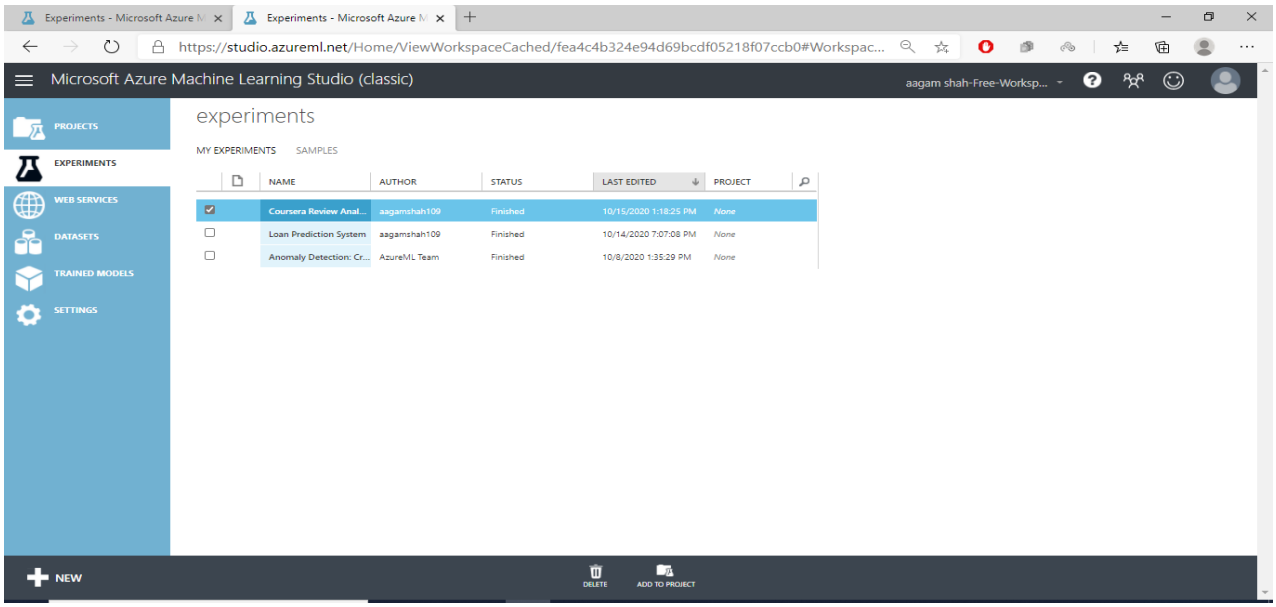

Sign in to your account
https://login.microsoftonline.com/common/oauth2/authorize?api-version=1.0&client_id=0736f41a-0425-4b46-bdb5-1563e...
InPrivate
Microsoft Azure Machine Learning
Microsoft
aagam01@somaiya.edu
Enter password
Forgot my password
Sign in
Terms of use
Privacy & cookies


Cloud Computing Services | Mic
https://azure.microsoft.com/en-in/
InPrivate
Microsoft Azure
Contact Sales
Search
My account
Portal
Sign In
Overview
Solutions
Products
Documentation
Pricing
Training
Marketplace
Partners
Support
Blog
More
Free account
We are in this together. Explore Azure resources and tools to help you navigate COVID-19
Learn, connect and explore with Azure at Microsoft Ignite. Invent with purpose.
Try Azure for free
Discover the latest technology at this free, all-digital event.
Get powerful, responsible AI on your terms
Establishing secure connection...



2. Create a new experiment

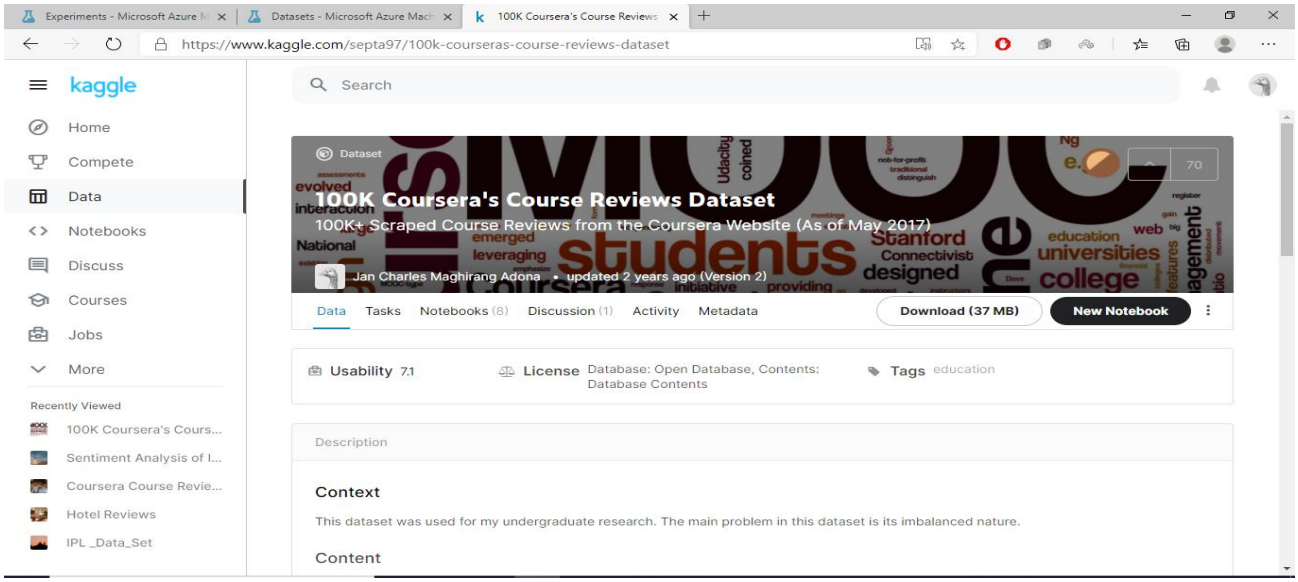
We create a new Experiment as blank experiment with name as Coursera Review Analysis.

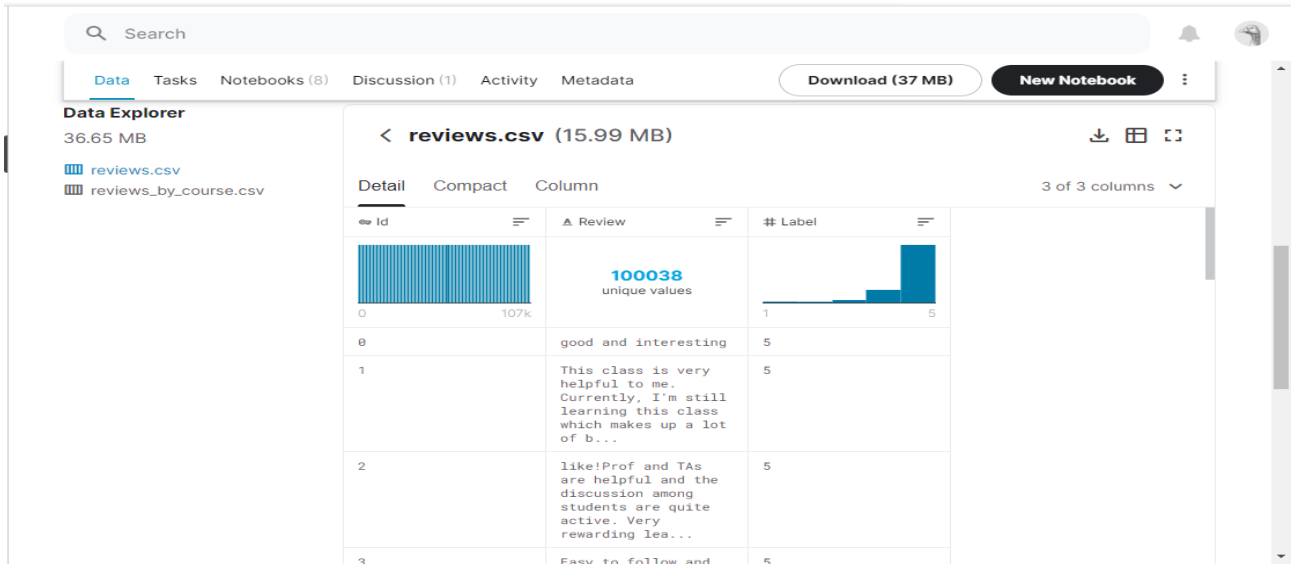


3. Dataset Downloading

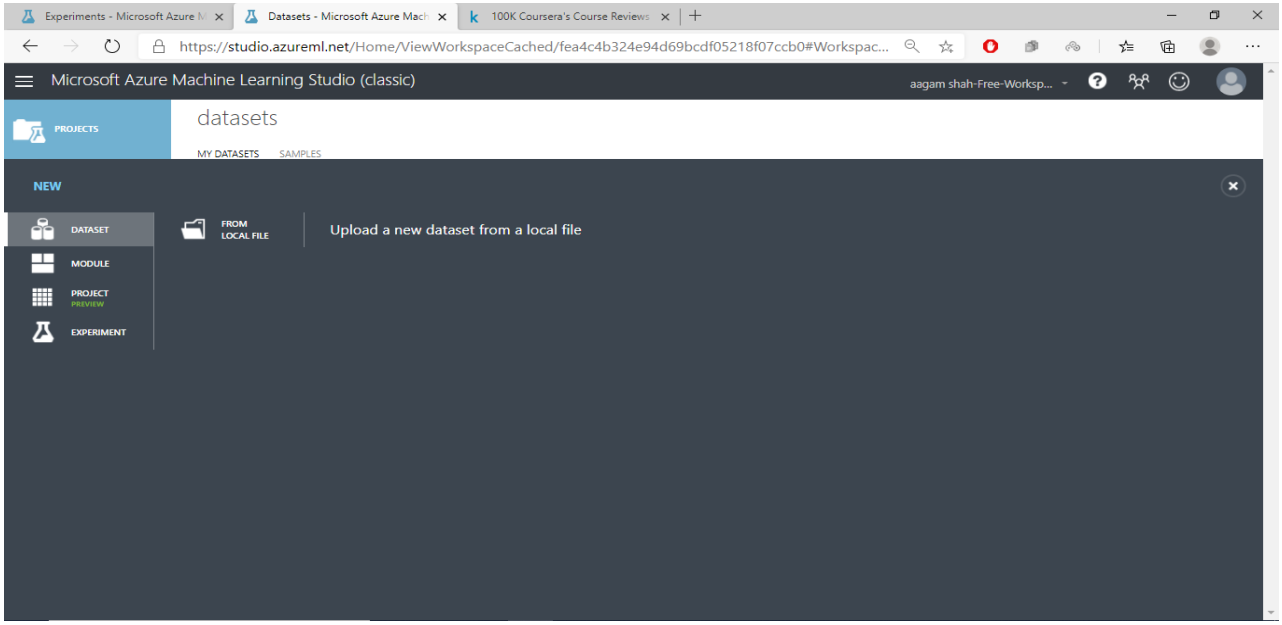
We download our dataset from Kaggle and export it as .csv file

https://www.kaggle.com/septa97/100k-courseras-course-reviews-dataset

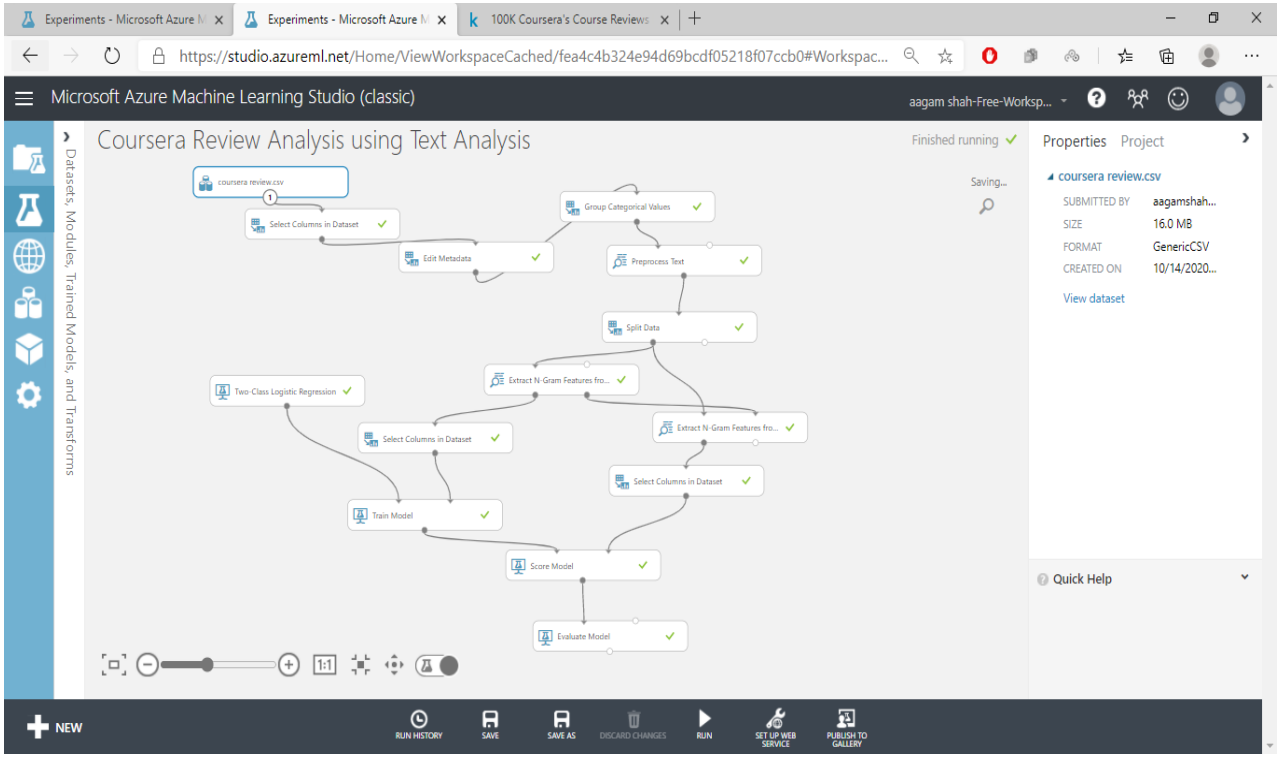




4. We Import our Dataset in Microsoft Azure ML Studio

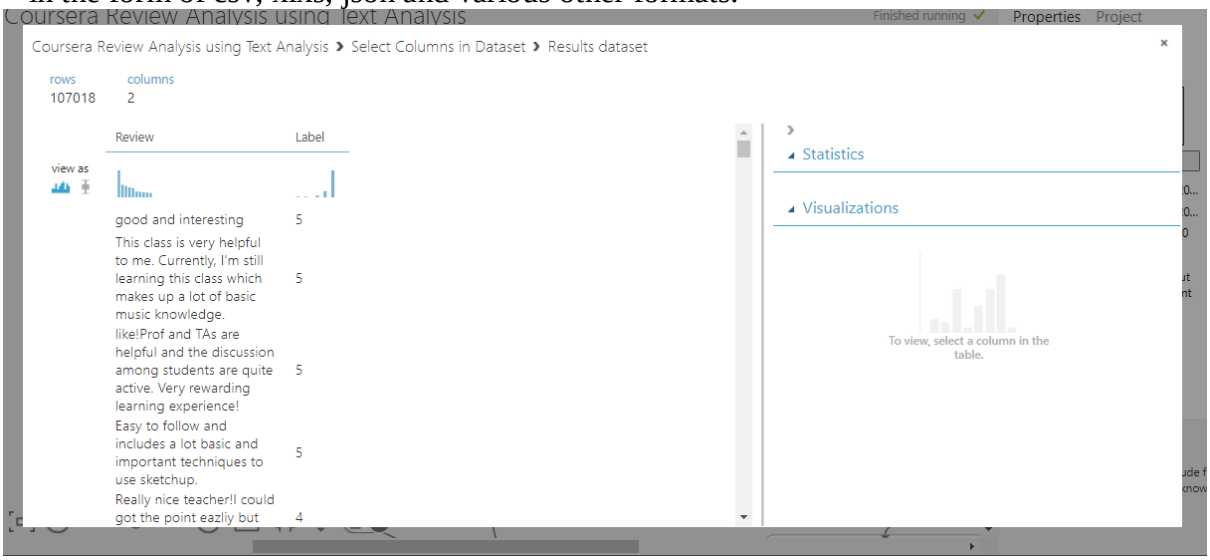


5. We create the our application structure using Drag and Drop UI in Microsoft Azure Studio

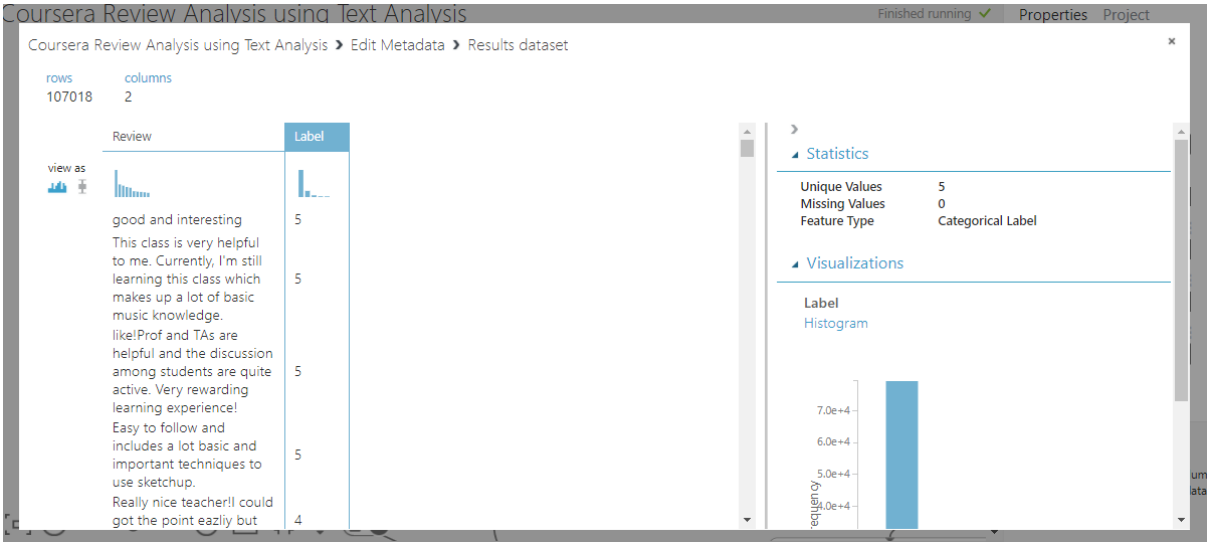


Module wise Explanation

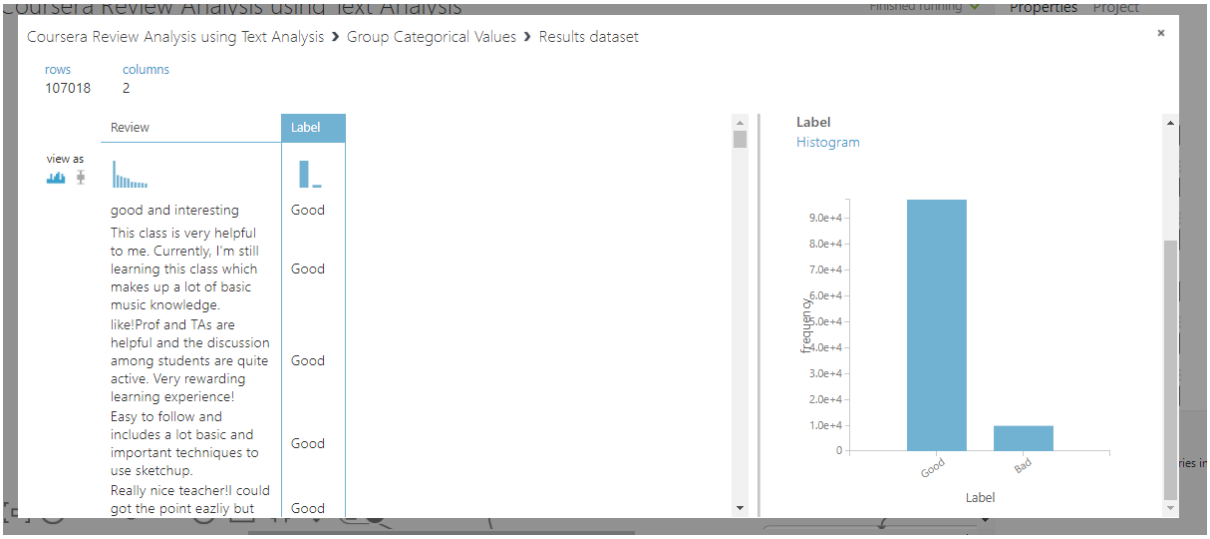
1. Select Columns in Dataset – We can select the dataset from this module. The dataset can be in the form of csv, xlxs, json and various other formats.



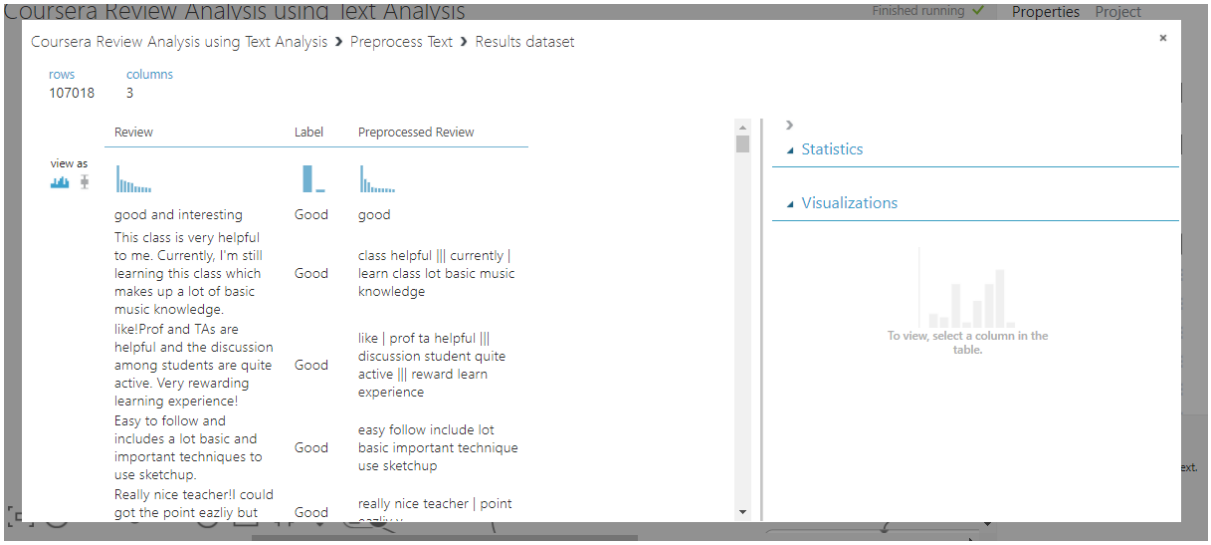
2. Edit Meta Data – We make the Label Column to Categorical.



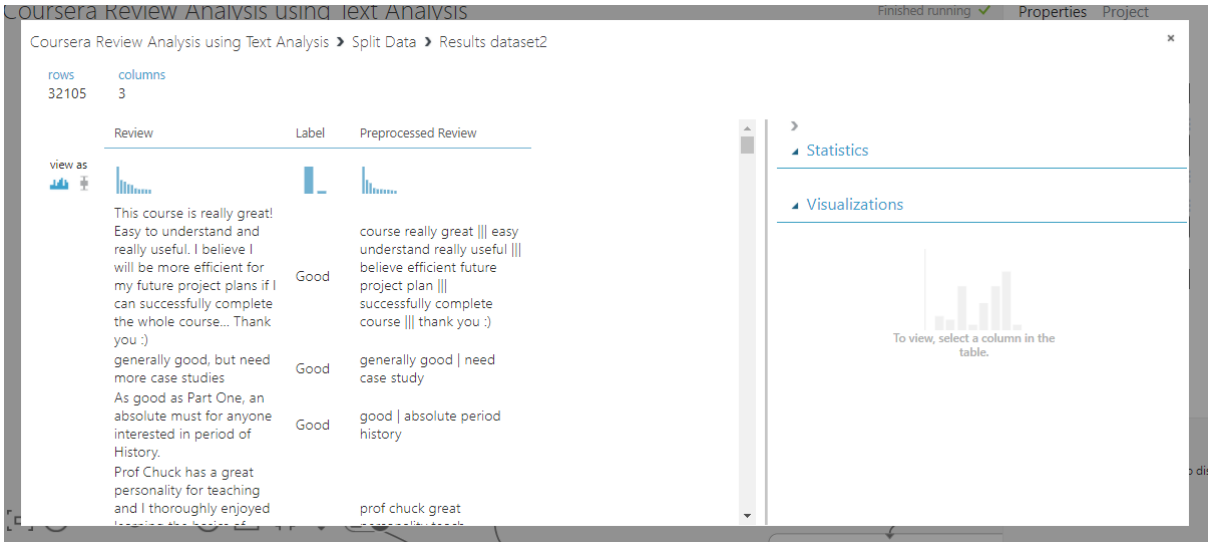
3. Group Categorical Values - In this module we set the value ranges from 0-2 to Bad and 3-5 to Good review



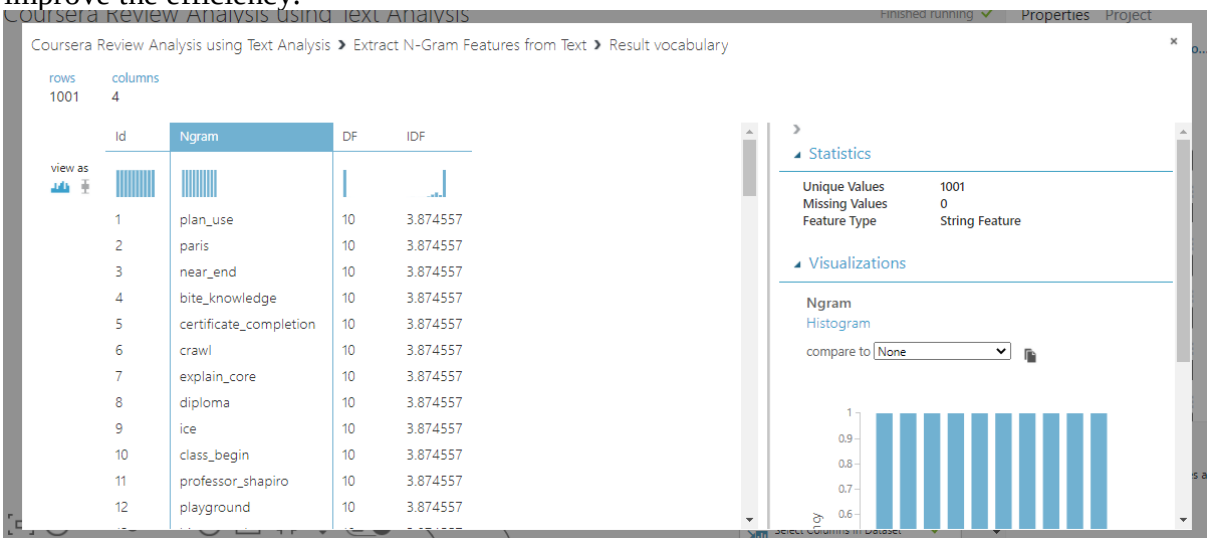
4. Pre-processing Text – In this Module we do the pre-processing of the text like removing the Stop word, lemmatization, Normalization of Case, removing url, special characters and other features

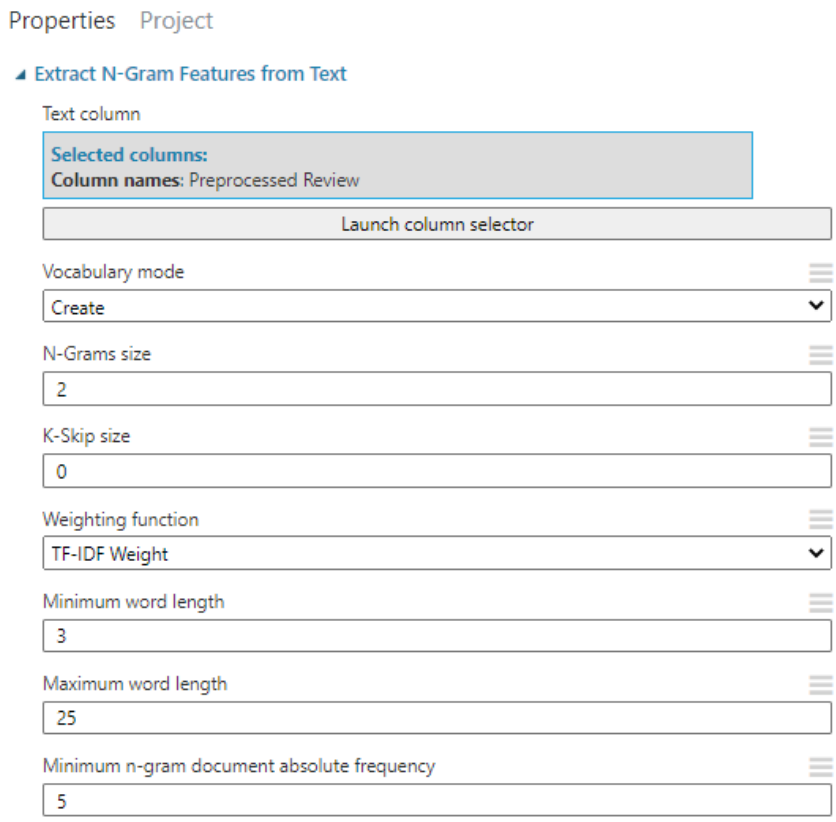


5. Split Data – Split Data

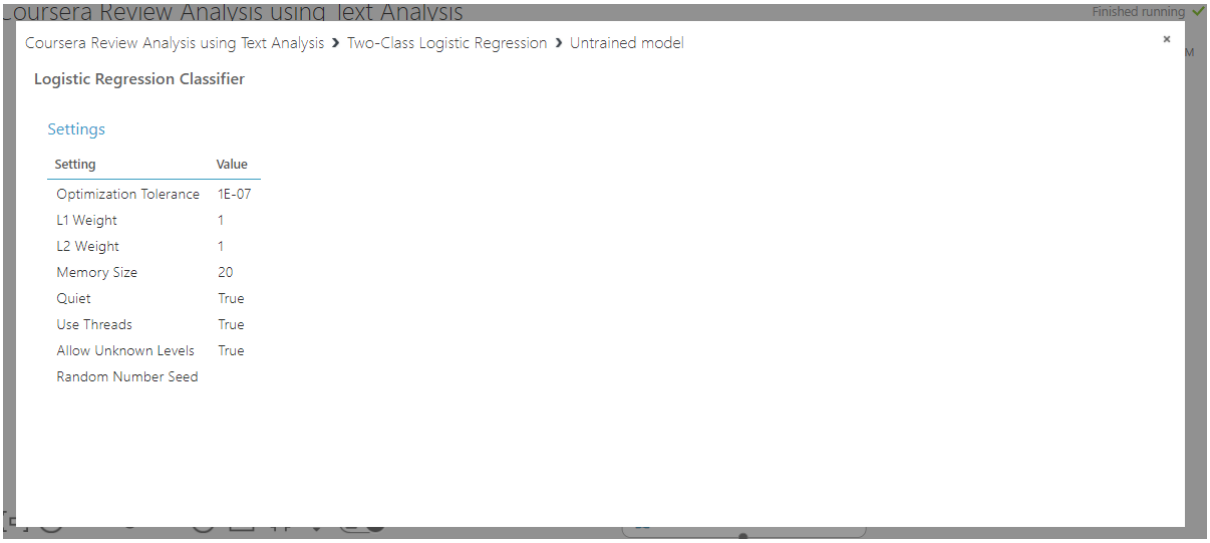


6. Extract N Grams – We extract all the different features from the pre-processed text here in our case we have set it to 1000 different features with N-Gram Size as 2 ie considering 2 words at a time for feature generation we validate the N-Gram by using Chi-Square test to improve the efficiency.

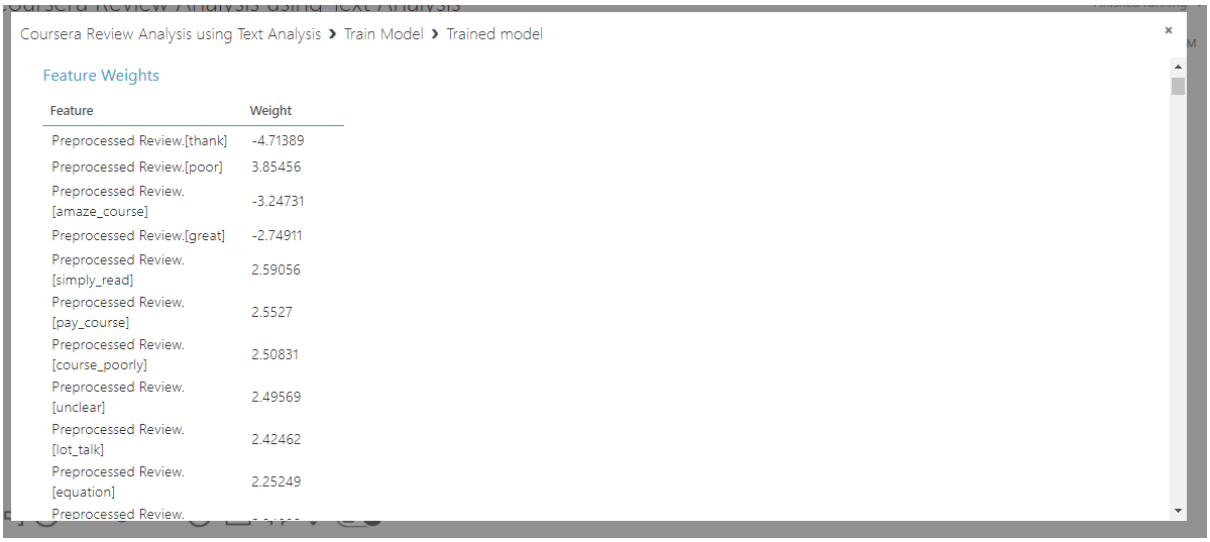




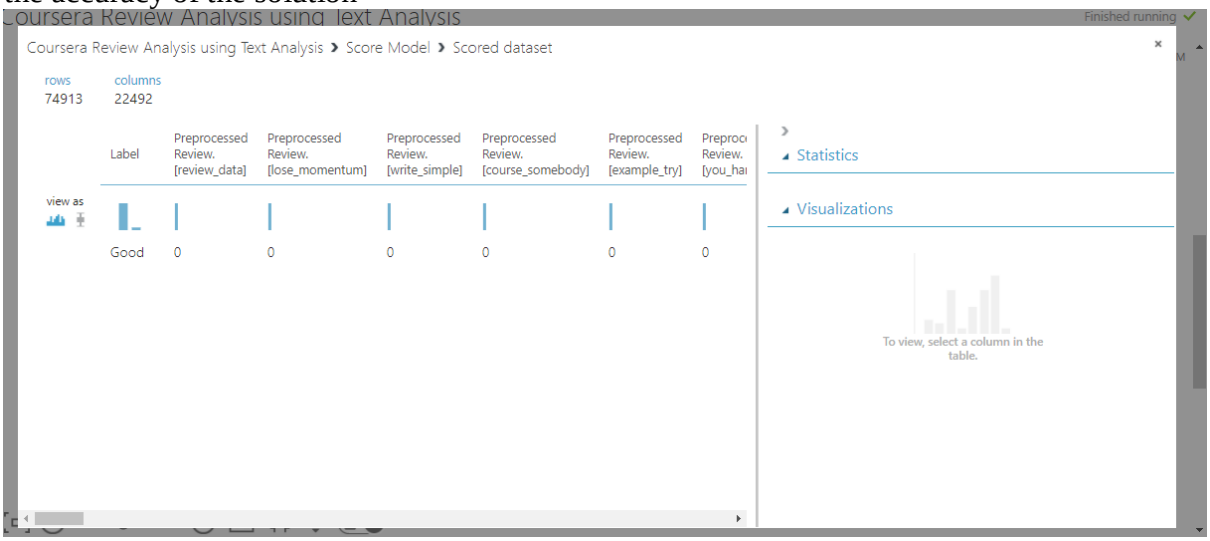
7. Two Class Logistic Regression – We used Two Class Logistic Regression in order to predict the Class Label of our Model



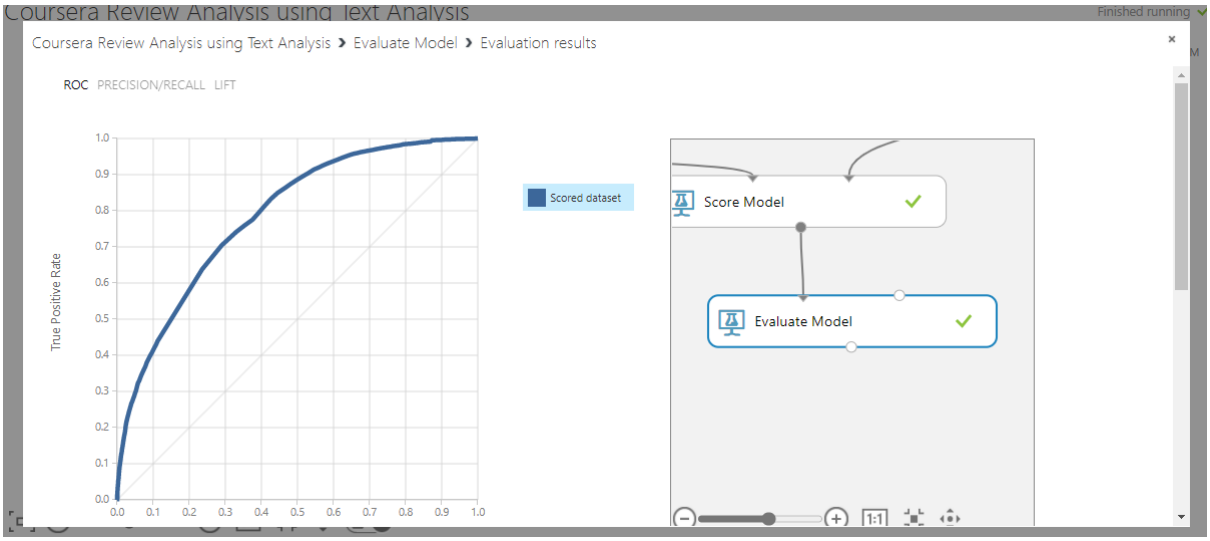
8. Train Model – This Module Trains our Model based on our dataset and the algorithm chosen.

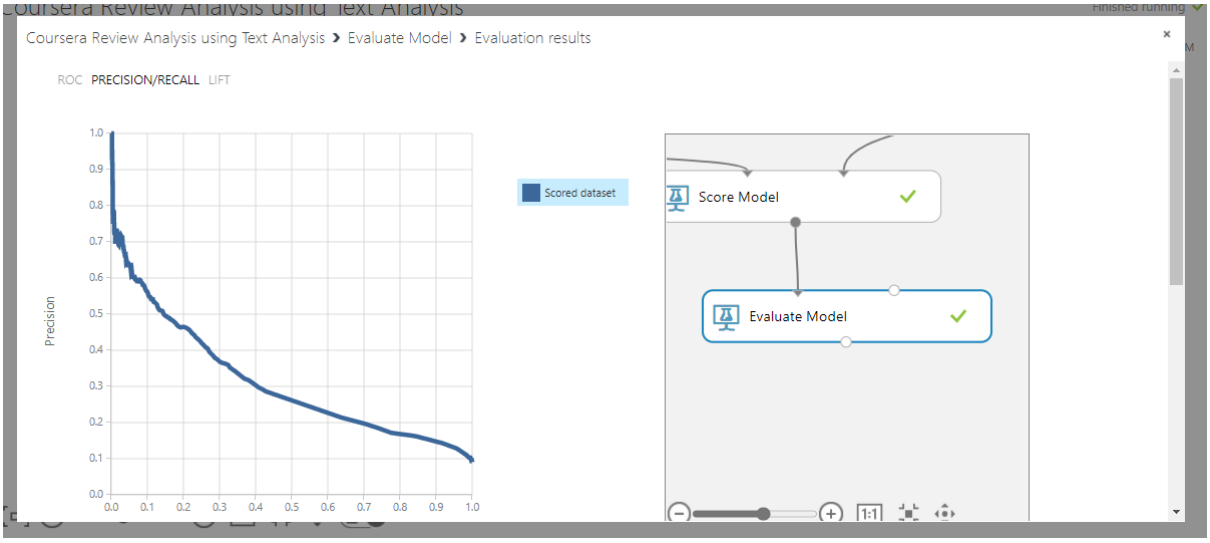


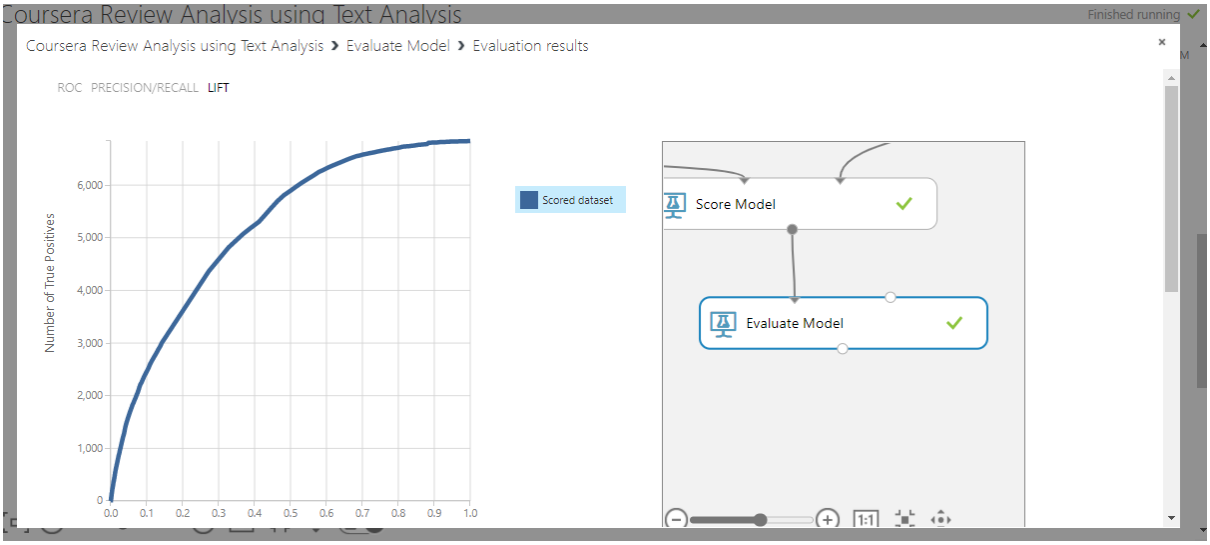
9. Score Model – Score model evaluated our model on new test dataset In order to determine the accuracy of the solution

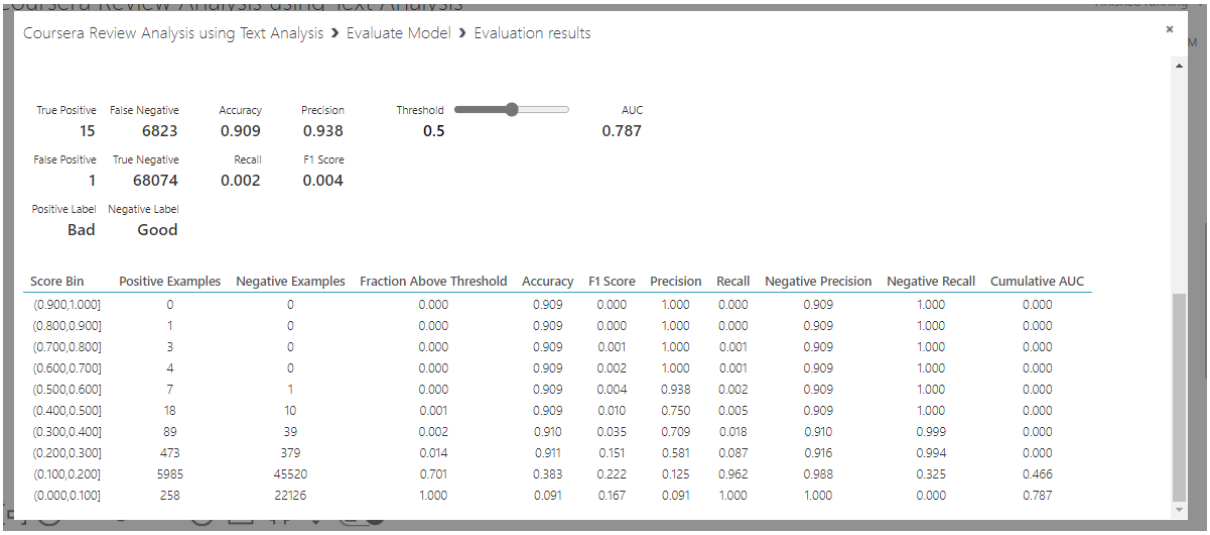


10. Evaluate Model – Gives us the Value of our Accuracy and confusion matrix







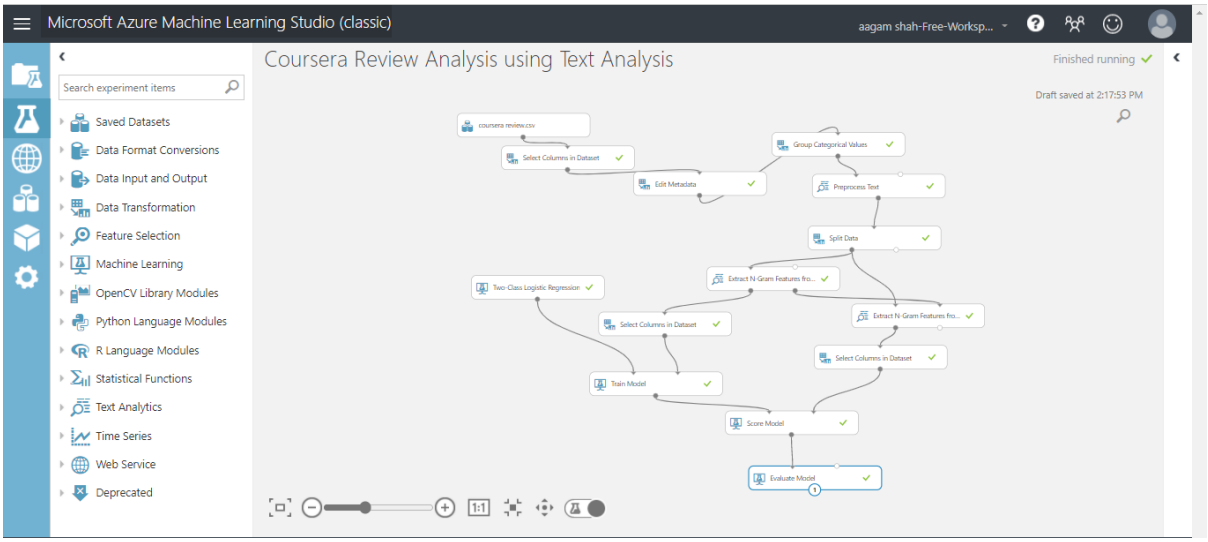


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
15	6823	0.909	0.938	0.5	0.787

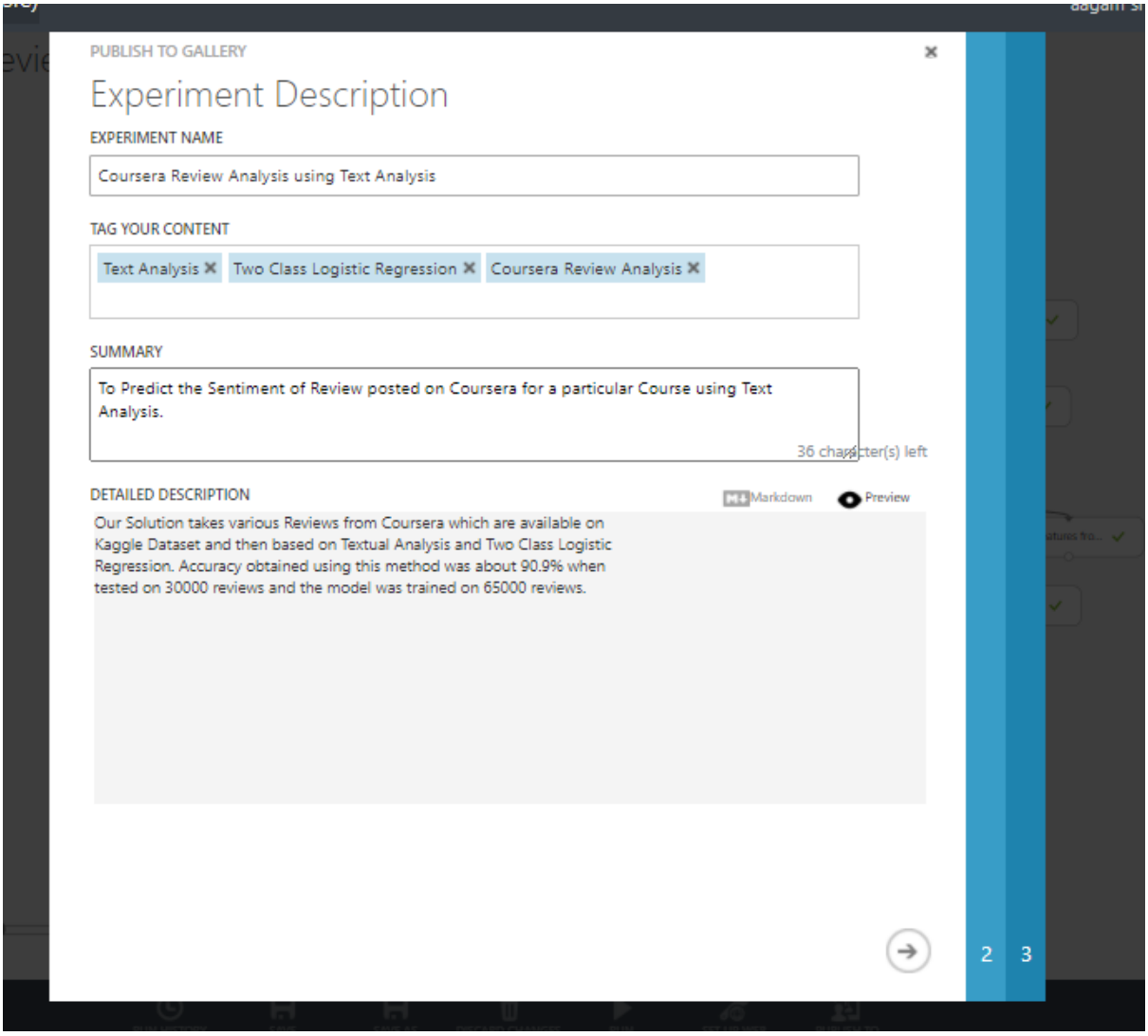
False Positive	True Negative	Recall	F1 Score
1	68074	0.002	0.004

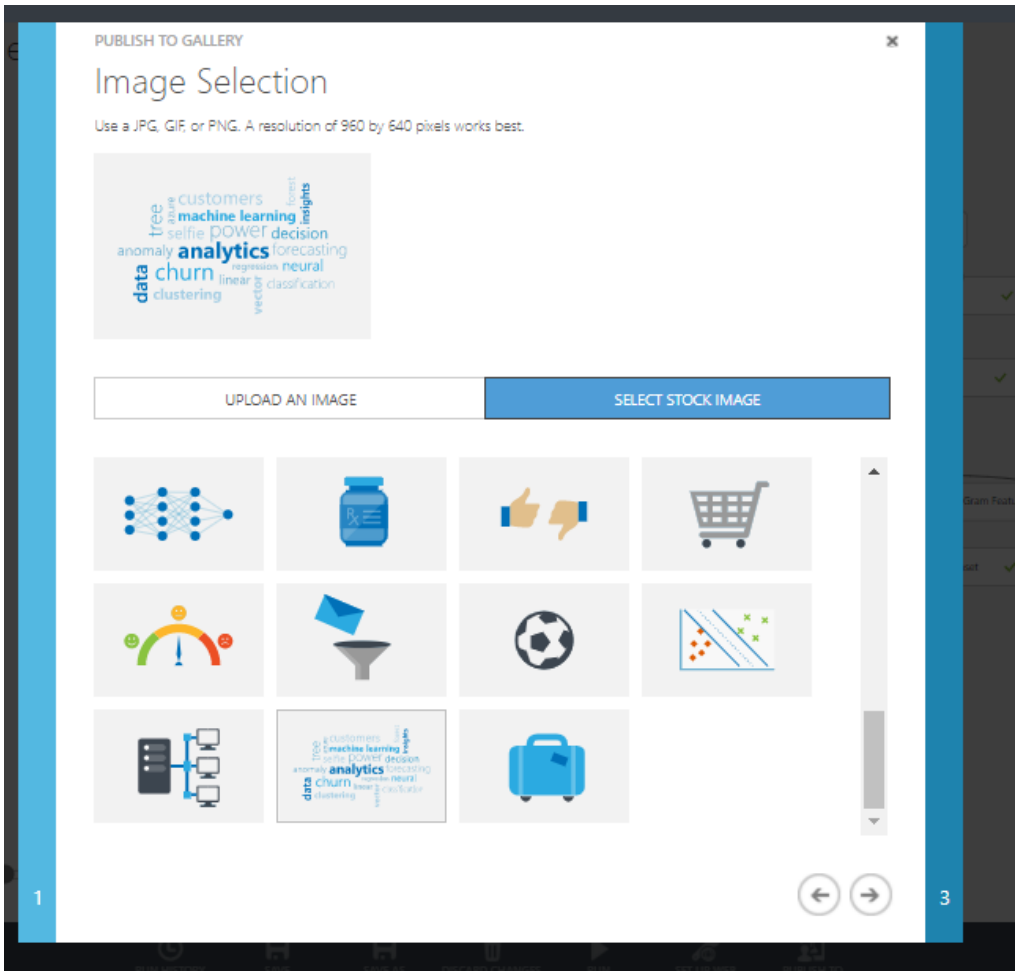
Positive Label	Negative Label
Bad	Good

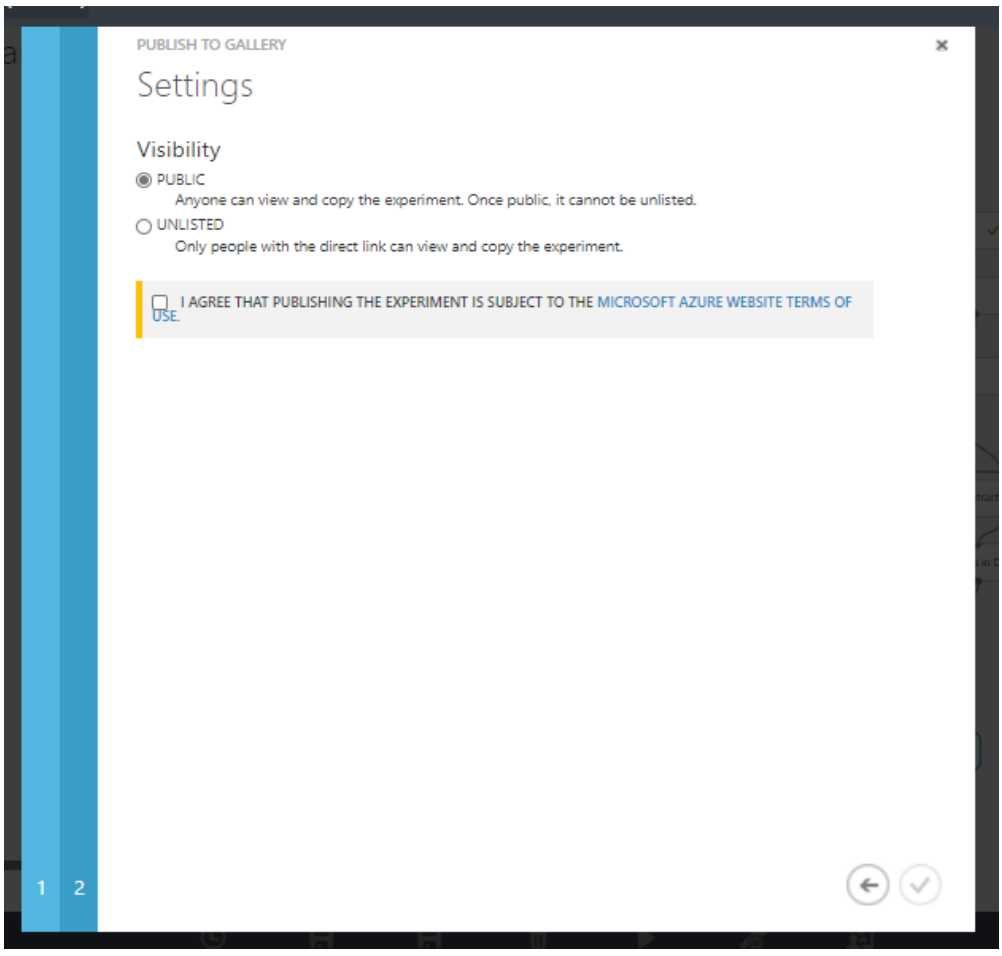
Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	0	0	0.000	0.909	0.000	1.000	0.000	0.909	1.000	0.000
(0.800,0.900]	1	0	0.000	0.909	0.000	1.000	0.000	0.909	1.000	0.000
(0.700,0.800]	3	0	0.000	0.909	0.001	1.000	0.001	0.909	1.000	0.000
(0.600,0.700]	4	0	0.000	0.909	0.002	1.000	0.001	0.909	1.000	0.000
(0.500,0.600]	7	1	0.000	0.909	0.004	0.938	0.002	0.909	1.000	0.000
(0.400,0.500]	18	10	0.001	0.909	0.010	0.750	0.005	0.909	1.000	0.000
(0.300,0.400]	89	39	0.002	0.910	0.035	0.709	0.018	0.910	0.999	0.000
(0.200,0.300]	473	379	0.014	0.911	0.151	0.581	0.087	0.916	0.994	0.000
(0.100,0.200]	5985	45520	0.701	0.383	0.222	0.125	0.962	0.988	0.325	0.466
(0.000,0.100]	258	22126	1.000	0.091	0.167	0.091	1.000	1.000	0.000	0.787

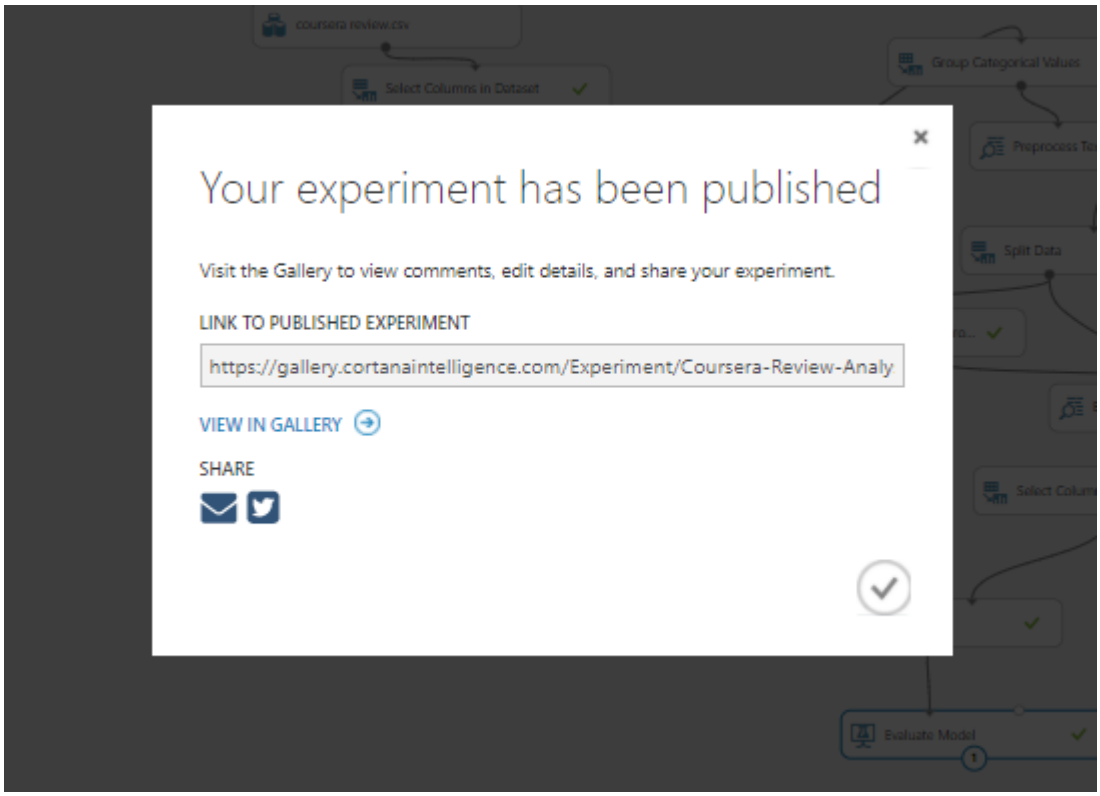


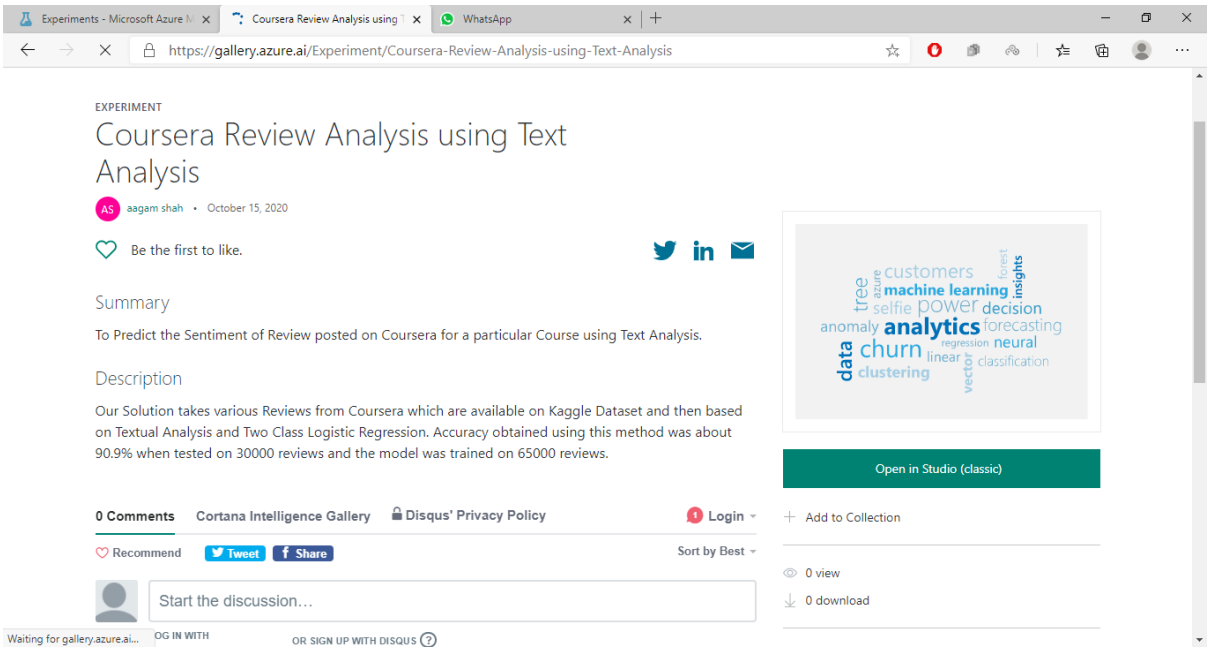
6. Publishing in Azure Gallery




PUBLISH TO GALLERY
Image Selection
Use a JPG, GIF or PNG. A resolution of 960 by 640 pixels works best.
UPLOAD AN IMAGE
SELECT STOCK IMAGE
1
3


PUBLISH TO GALLERY
Settings
Visibility
PUBLIC
Anyone can view and copy the experiment. Once public, it cannot be unlisted.
UNLISTED
Only people with the direct link can view and copy the experiment.
I AGREE THAT PUBLISHING THE EXPERIMENT IS SUBJECT TO THE MICROSOFT AZURE WEBSITE TERMS OF USE.
1
2





Outcomes: Realize adequate perspectives of big data analytics in various applications.

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Hence in this experiment we explored ML Azure Studio and implemented Coursera Review Analysis using Text Analysis using Two Class Logistic Regression Algorithm and predicted the sentiment of the Text Review found on coursera.

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites: