

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

The first screenshot shows the Microsoft Azure login page. The browser address bar displays `https://login.microsoftonline.com/common/oauth2/authorize?api-version=1.0&client_id=0736f41a-0425-4b46-bdb5-1563e...`. The page title is "Sign in to your account". The main content area features the Microsoft Azure Machine Learning logo and a "Sign in" form. The form includes the Microsoft logo, the text "Sign in", an email input field containing "aagam01@somaiya.edu", and links for "No account? Create one!", "Can't access your account?", and "Sign-in options". A blue "Next" button is at the bottom right of the form.

The second screenshot shows the password entry step. The browser address bar is the same. The page title is "Sign in to your account". The main content area features the Microsoft Azure Machine Learning logo and an "Enter password" form. The form includes the Microsoft logo, a back arrow, the email "aagam01@somaiya.edu", the text "Enter password", a password input field with masked characters ".....", a link for "Forgot my password", and a blue "Sign in" button.

The third screenshot shows the Microsoft Azure portal homepage. The browser address bar displays `https://azure.microsoft.com/en-in/`. The page title is "Cloud Computing Services | Microsoft". The header includes the Microsoft Azure logo, navigation links (Overview, Solutions, Products, Documentation, Pricing, Training, Marketplace, Partners, Support, Blog, More), and user account links (Contact Sales, Search, My account, Portal, Sign In). A blue banner reads "We are in this together. Explore Azure resources and tools to help you navigate COVID-19 >". The main content area features the text "Learn, connect and explore with Azure at Microsoft Ignite. Invent with purpose." and a green "Try Azure for free" button. Below this, it says "Discover the latest technology at this free, all-digital event." and shows a partial view of a registration form with a "Communication Services" button.

Course: Big Data Analytics_O-20 | Experiments - Microsoft Azure ML | BDA EXP 9.docx - Google Drive | Microsoft Azure Machine Learning | +

https://studio.azureml.net

Microsoft Azure Machine Learning Studio (classic)

>Welcome back aagamshah109!

MY RECENT WORKSPACES:

- aagam shah-Free-Workspace

MY RECENT EXPERIMENTS:

- Coursera Review Analysis using Text Analysis
- Loan Prediction System
- Anomaly Detection: Credit Risk

[my experiments](#)

Announcements NEW!

Azure Machine Learning Studio R Runtime Upgrade
Aired on October 31, 2018

The R language engine in the Execute R Script module of Azure Machine Learning Studio has added a new R runtime version -- Microsoft R Open (MRO) 3.4.4. MRO 3.4.4 is based on open-source CRAN R 3.4.4 and is therefore compatible with packages that works with that

Mining Campaign Funds
Aired on August 03, 2017

Play with 2016 Presidential Campaign finance data while learning how to prepare a large dataset for machine learning by processing and engineering features. This sample experiment works on a 2.5 GB dataset and will take about 20 minutes to run in its entirety.

[Learn More](#)

Inside the Data Science VM
Aired on June 21, 2016

DSVM is a custom Azure Virtual Machine image that is published on the Azure marketplace and available on both Windows and Linux. It contains several popular data science and development tools both from Microsoft and from the open source community all pre-installed and pre-configured and ready to use.

2. Create a new experiment

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

Microsoft Azure Machine Learning Studio (classic)

experiments

MY EXPERIMENTS | SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED	PROJECT
<input checked="" type="checkbox"/>	Coursera Review Anal...	aagamshah109	Finished	10/15/2020 1:18:25 PM	None
<input type="checkbox"/>	Loan Prediction System	aagamshah109	Finished	10/14/2020 7:07:08 PM	None
<input type="checkbox"/>	Anomaly Detection: Cr...	AzureML Team	Finished	10/8/2020 1:35:29 PM	None

+ NEW | DELETE | ADD TO PROJECT

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>

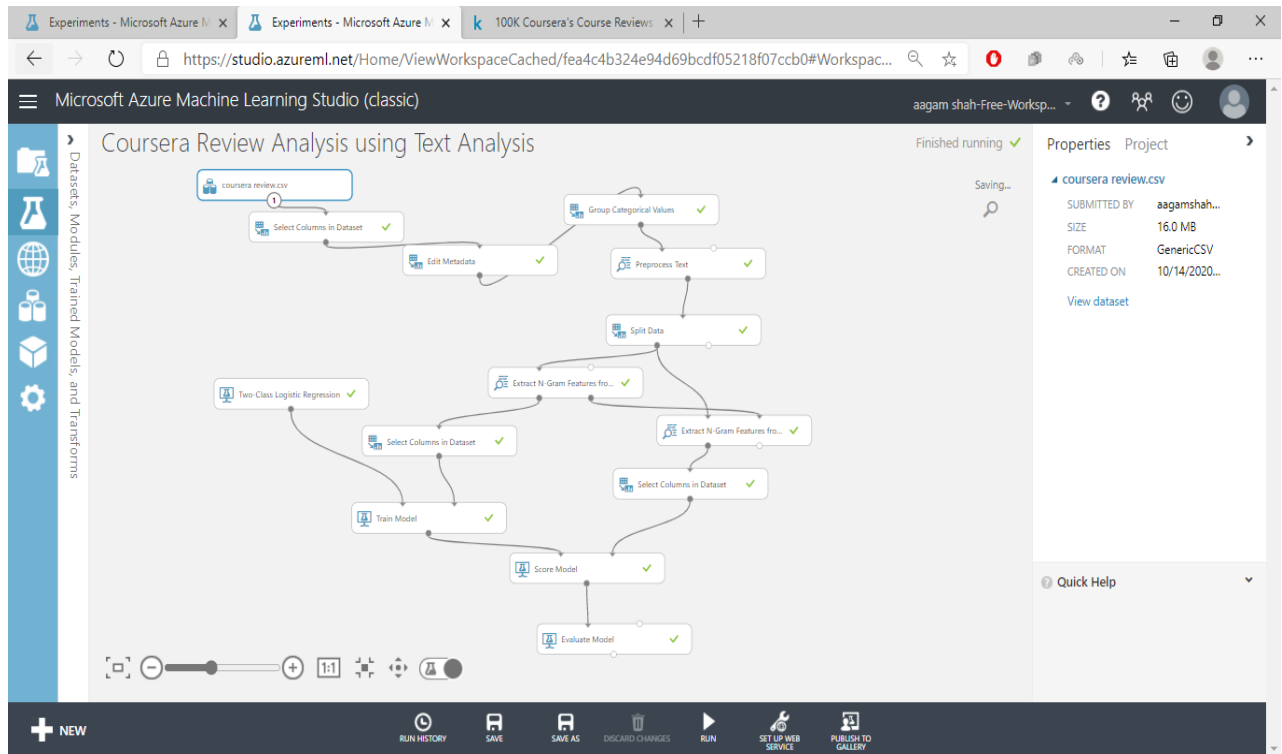
The screenshot shows the Kaggle dataset page for '100K Coursera's Course Reviews Dataset'. The page includes a search bar, a sidebar with navigation links (Home, Compete, Data, Notebooks, Discuss, Courses, Jobs, More), and a main content area. The dataset is by Jan Charles Maghirang Adona, updated 2 years ago (Version 2). It has a usability of 7.1 and is licensed under 'Database: Open Database, Contents: Database Contents'. The tags are 'education'. The description states: 'This dataset was used for my undergraduate research. The main problem in this dataset is its imbalanced nature.'

The screenshot shows the Microsoft Azure ML Studio Data Explorer interface. The dataset 'reviews.csv' (15.99 MB) is selected. The interface displays a table with columns: Id, Review, and # Label. The 'Review' column shows a preview of the dataset content, including phrases like 'good and interesting', 'This class is very helpful to me.', and 'like!Prof and TAs are helpful and the discussion among students are quite active. Very rewarding lea...'. The '# Label' column shows a distribution of labels, with a bar chart indicating 100038 unique values.

4. We Import our Dataset in Microsoft Azure ML Studio

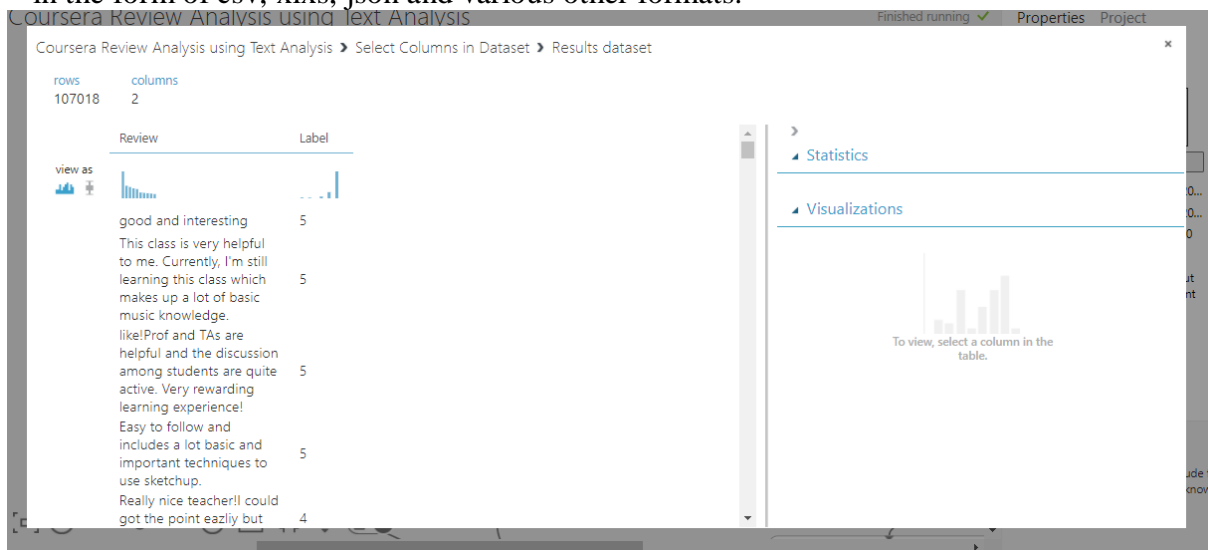
The screenshot shows the Microsoft Azure Machine Learning Studio (classic) interface. The 'datasets' section is active, displaying a 'NEW' button and a 'FROM LOCAL FILE' option. The main area prompts the user to 'Upload a new dataset from a local file'. The sidebar on the left shows navigation links for DATASET, MODULE, PROJECT, and EXPERIMENT.

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

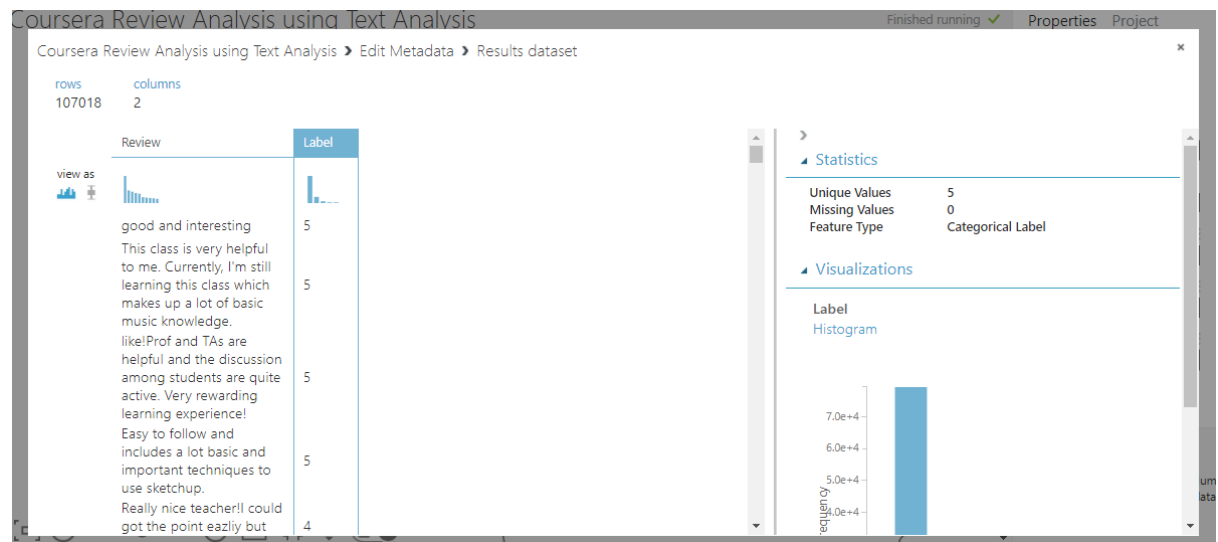


Module wise Explanation

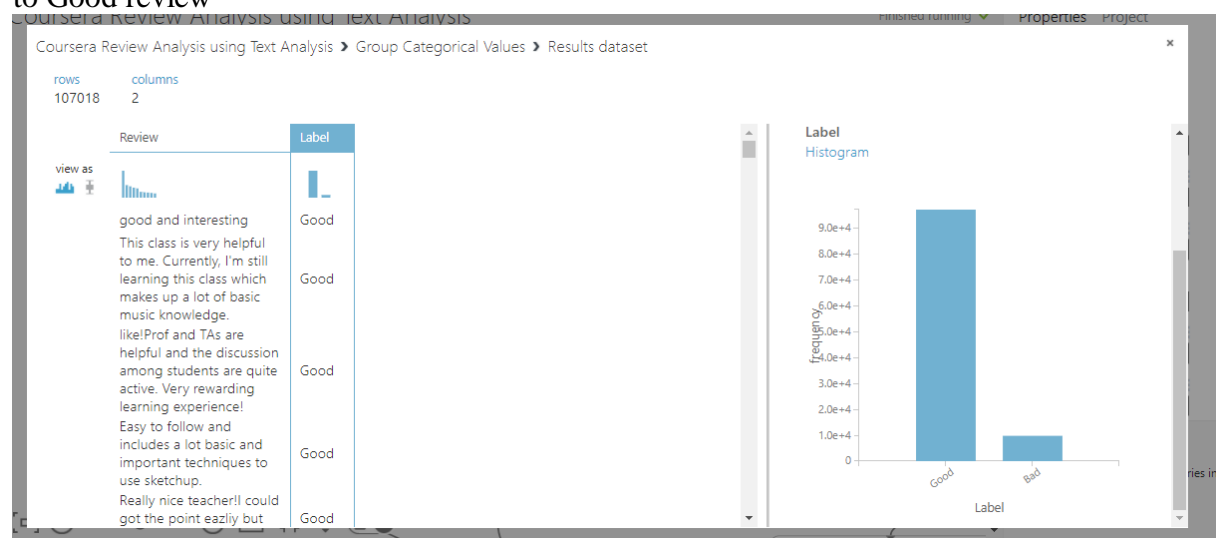
1. **Select Columns in Dataset** – We can select the dataset from this module. The dataset can be in the form of csv, xlsx, json and various other formats.



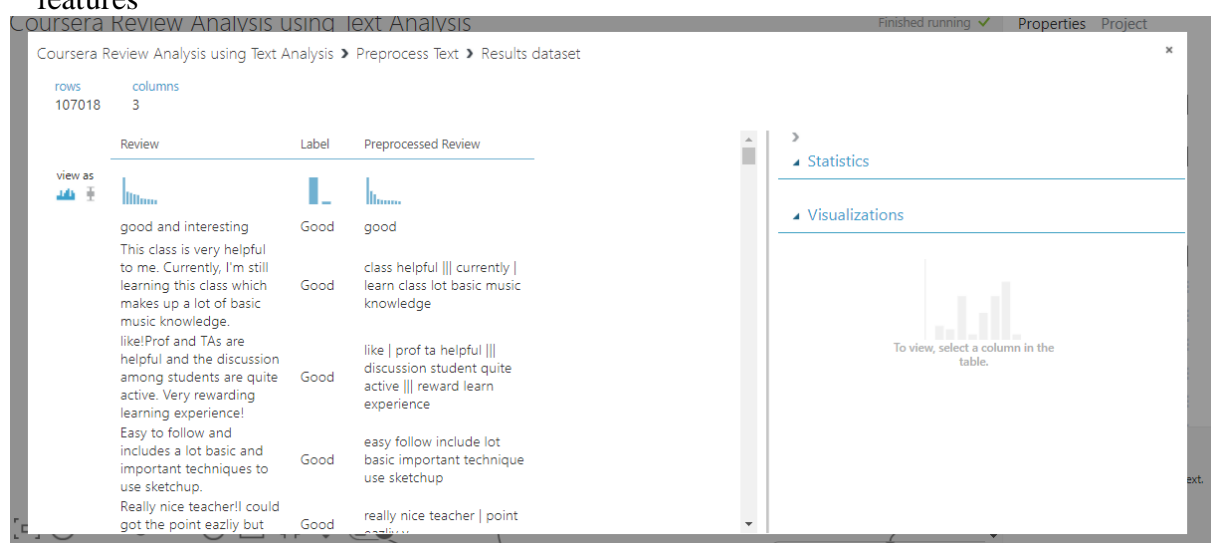
2. **Edit Meta Data** – We make the Label Column to Categorical.



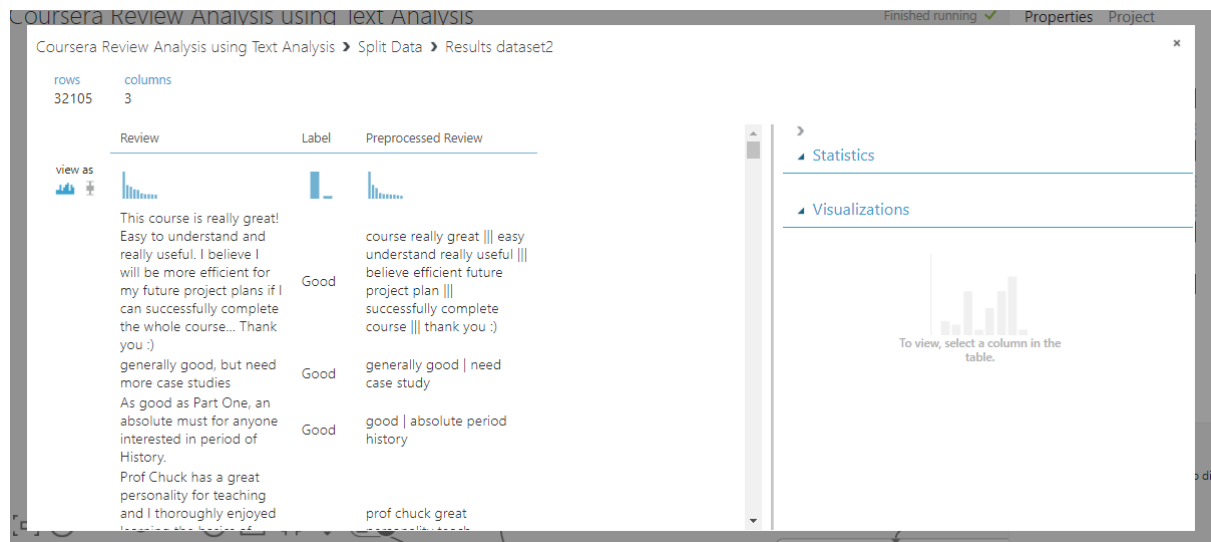
3. **Group Categorical Values** - In this module we set the value ranges from 0-2 to Bad and 3-5 to Good review



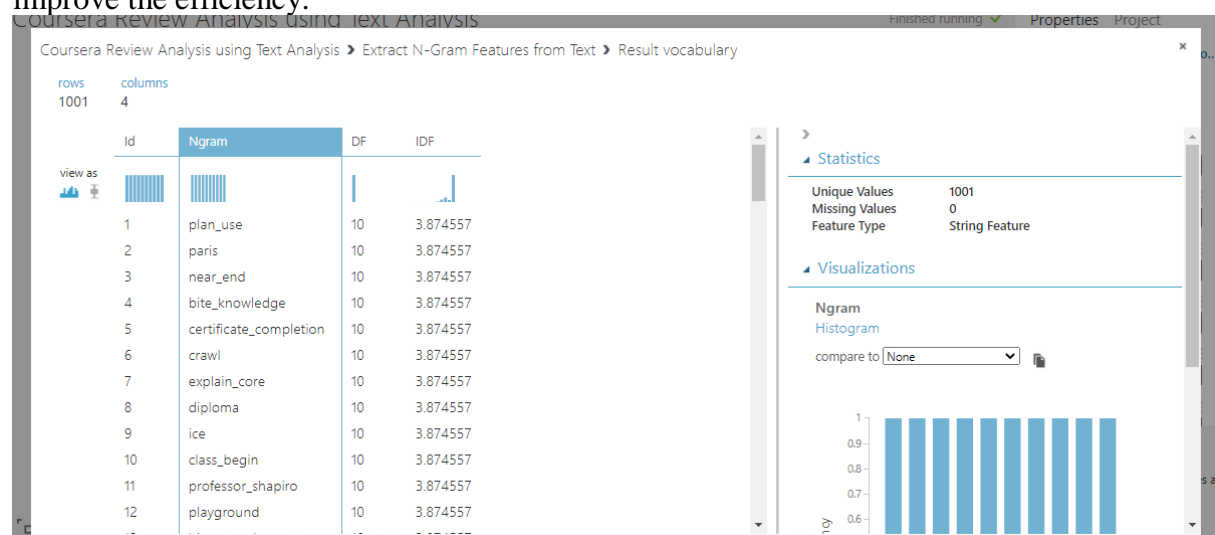
4. **Pre-processing Text** – In this Module we do the pre-processing of the text like removing the Stop word, lemmatization, Normalization of Case, removing url, special characters and other features



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.



Properties Project

Extract N-Gram Features from Text

Text column

Selected columns:

Column names: Preprocessed Review

Launch column selector

Vocabulary mode

Create

N-Grams size

2

K-Skip size

0

Weighting function

TF-IDF Weight

Minimum word length

3

Maximum word length

25

Minimum n-gram document absolute frequency

5

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

Courseera Review Analysis using Text Analysis

Finished running

Courseera Review Analysis using Text Analysis > Two-Class Logistic Regression > Untrained model

Logistic Regression Classifier

Settings

Setting	Value
Optimization Tolerance	1E-07
L1 Weight	1
L2 Weight	1
Memory Size	20
Quiet	True
Use Threads	True
Allow Unknown Levels	True
Random Number Seed	

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

Coursera Review Analysis using Text Analysis > Train Model > Trained model

Feature Weights

Feature	Weight
Preprocessed Review.[thank]	-4.71389
Preprocessed Review.[poor]	3.85456
Preprocessed Review.[amaze_course]	-3.24731
Preprocessed Review.[great]	-2.74911
Preprocessed Review.[simply_read]	2.59056
Preprocessed Review.[pay_course]	2.5527
Preprocessed Review.[course_poorly]	2.50831
Preprocessed Review.[unclear]	2.49569
Preprocessed Review.[lot_talk]	2.42462
Preprocessed Review.[equation]	2.25249
Preprocessed Review.	

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

Coursera Review Analysis using Text Analysis

Coursera Review Analysis using Text Analysis > Score Model > Scored dataset

rows: 74913, columns: 22492

Label	Preprocessed Review.[review_data]	Preprocessed Review.[lose_momentum]	Preprocessed Review.[write_simple]	Preprocessed Review.[course_somebody]	Preprocessed Review.[example_try]	Preprocessed Review.[you_hai]
Good	0	0	0	0	0	0

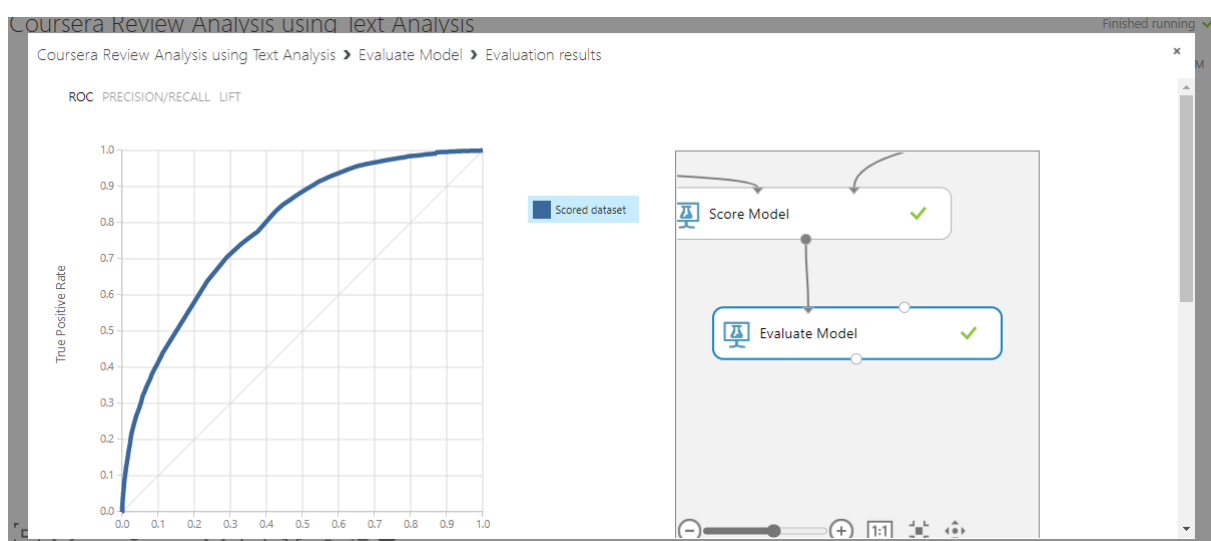
view as: [Bar chart icon]

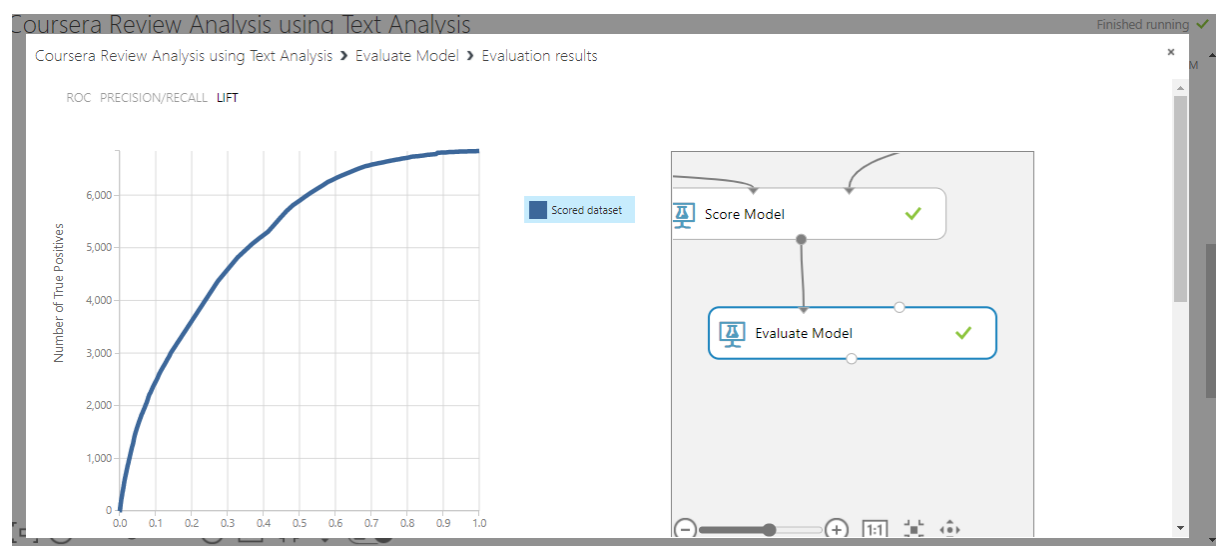
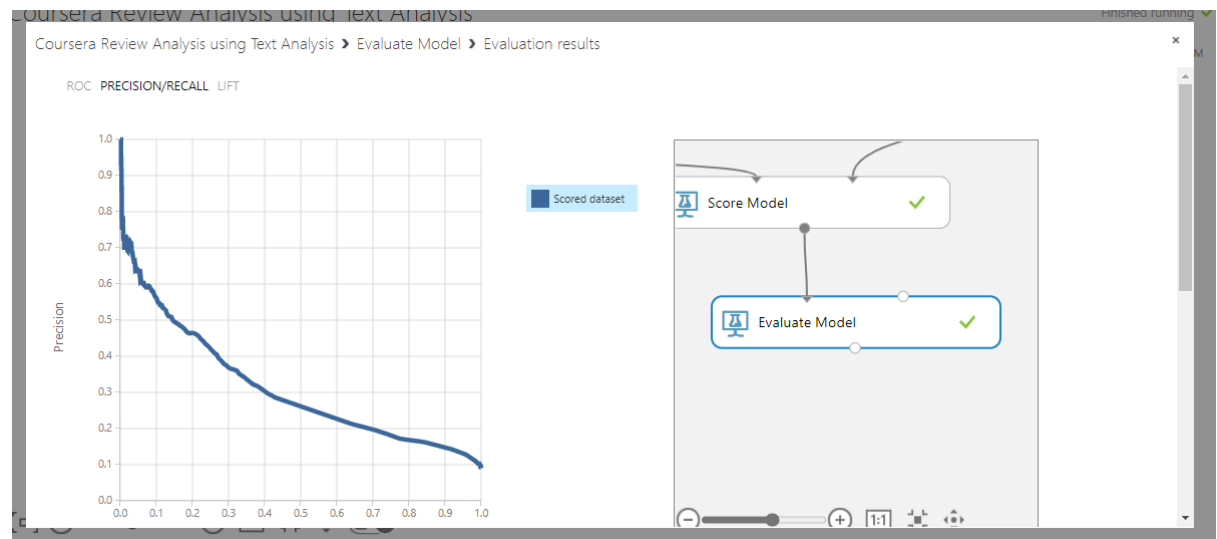
Statistics

Visualizations

To view, select a column in the table.

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



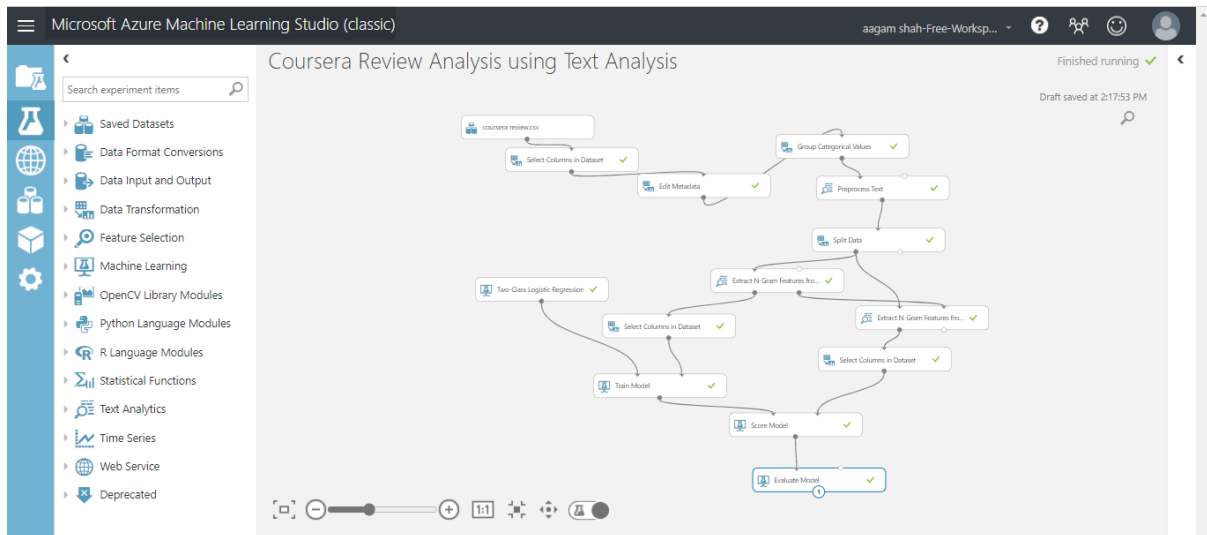


Courseera Review Analysis using Text Analysis

Courseera Review Analysis using Text Analysis > Evaluate Model > Evaluation results

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

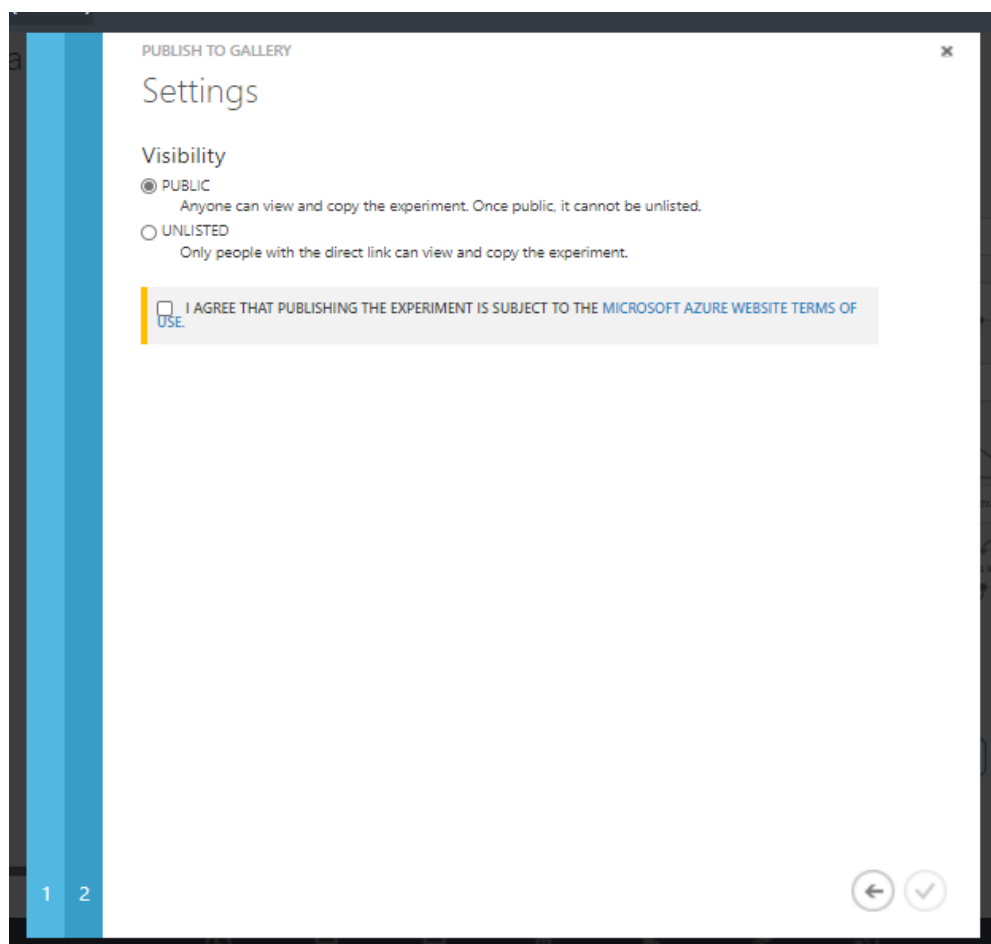
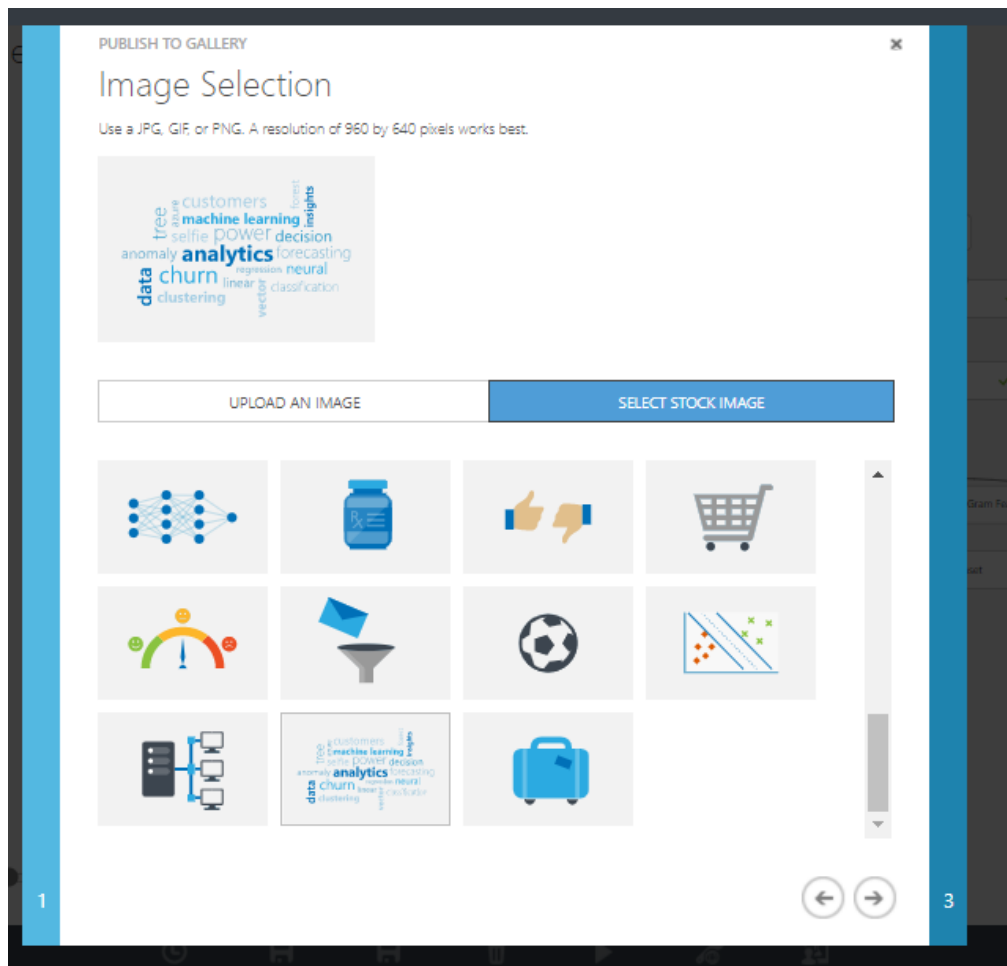


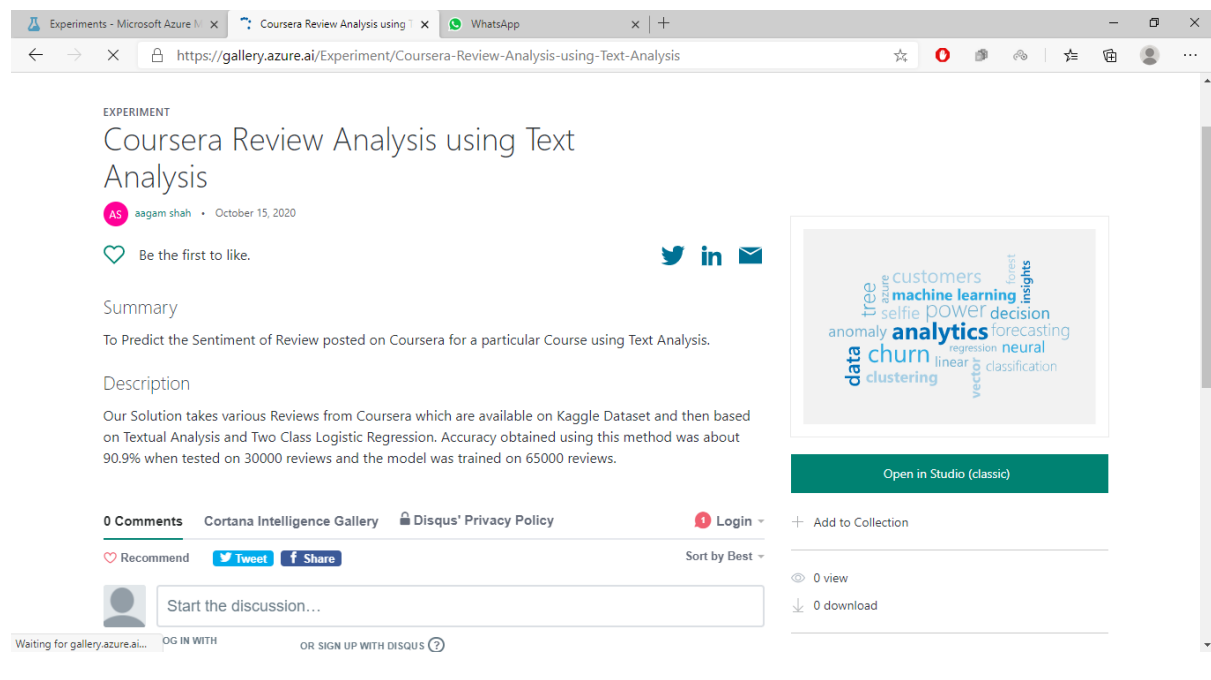
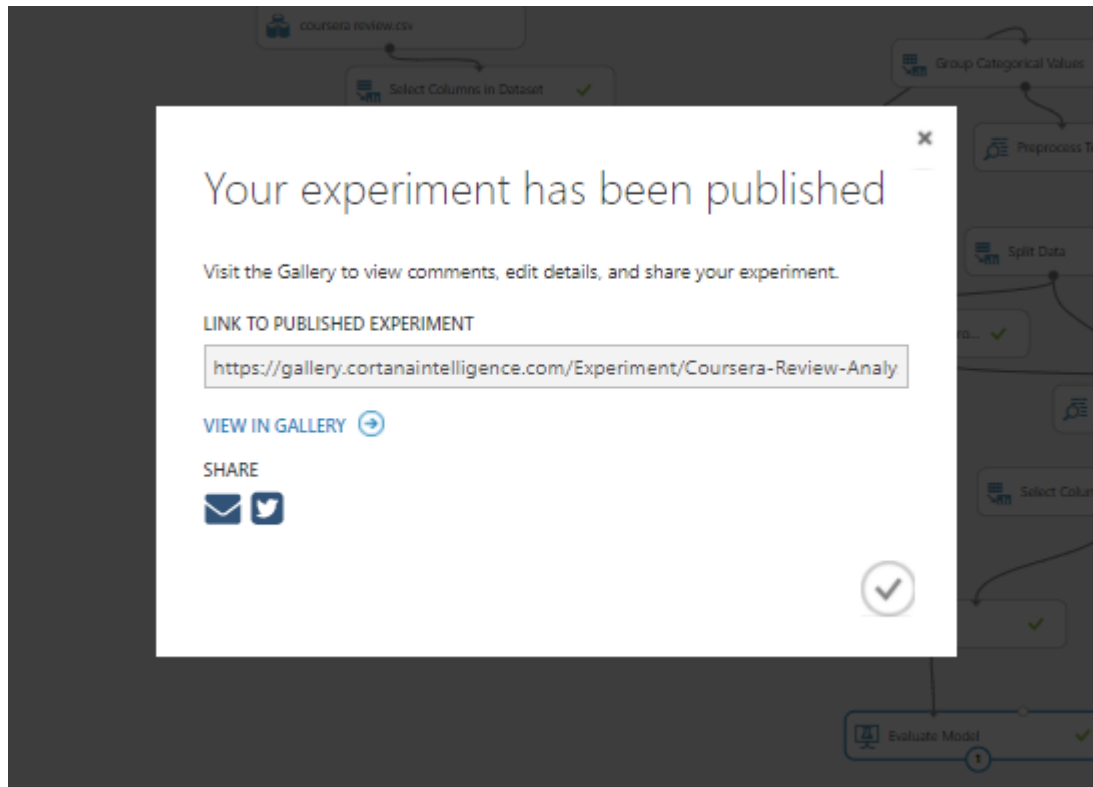
6. Publishing in Azure Gallery

The screenshot shows the 'PUBLISH TO GALLERY' dialog box in Azure Machine Learning Studio. The dialog has a title bar 'PUBLISH TO GALLERY' and a close button. The main content area is titled 'Experiment Description'. It contains the following sections:

- EXPERIMENT NAME:** A text box containing 'Coursera Review Analysis using Text Analysis'.
- TAG YOUR CONTENT:** A section with three tags: 'Text Analysis', 'Two Class Logistic Regression', and 'Coursera Review Analysis'.
- SUMMARY:** A text box containing 'To Predict the Sentiment of Review posted on Coursera for a particular Course using Text Analysis.'.
- DETAILED DESCRIPTION:** A section with a 'Markdown' icon and a 'Preview' icon. The text reads: '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.'

At the bottom right, there is a blue bar with the number '2' and a right arrow icon.





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:

