ANALYZING THE FACTORS AFFECTING COURSE RATINGS IN COURSERA

ANTARA SAHA, VERA FAN, ALFREDO FRANCISCO CARAFI, KRITIKA DWIVEDI

EXECUTIVE SUMMARY

The objective of this project was to analyze factors affecting Coursera user ratings, focusing on sentiment analysis techniques, the entity providing the course, and course category. Our aim was to identify user sentiments about courses, areas for improvement, and the factors affecting course ratings to inform Coursera's future decisions, thus leading to an increase in revenue, brand value and customer base.

Our primary data sources included publicly available datasets from Kaggle, containing Coursera course information, reviews, and world university rankings. We performed Exploratory Data Analysis (EDA), Aspect Based Sentiment Analysis (ABSA), Multivariate Regression and Topic Modeling on the reviews to gain insights into users' positive and negative sentiments regarding courses. We integrated data from different sources to analyze the relationships between course ratings, sentiment scores, course categories, and institutions.

Our key takeaways from this project are as following:

- Course ratings are significantly influenced by user sentiment, highlighting the importance of addressing factors affecting user experience, such as course difficulty, instructor capability, clarity of instructions, and affordability.
- Although there is a difference in average ratings between courses offered by universities and those offered by companies, the difference is not statistically significant, suggesting that both types of course providers can offer quality courses.
- The top three highest average rated course categories are Personal Development, Arts and Humanities, and Language Learning. These categories may present opportunities for Coursera to further invest in and expand their course offerings.
- Data Science and Business courses are the most frequently offered on Coursera, but their average ratings are below the median rating across all categories. This indicates a potential need for improvements in these popular categories to enhance user satisfaction.

1. PROBLEM FRAMING

Online learning platforms have gained popularity in the last two decades. They are a great way for students and professionals alike to expand their knowledge and skill sets at a convenient pace and less cost than the universities. Coursera is one such platform. It offers thousands of courses from more than 200 world-class universities and companies. But like any other business or product, online learning platforms and offered courses can get mixed reviews from their users. And like any other business, if they fail to address the negative feedback from their users, they will eventually run out of business. Hence, they must analyze and address user feedback. However, given the scale, for example, 5,400+ courses on Coursera with 100+ million users, it is impossible to analyze the user feedback without using sophisticated analytical techniques.

For our final project, we are analyzing the Coursera user reviews through sentiment analysis techniques to identify the user sentiments about their courses and the areas for further improvement. Further, we also aimed to identify additional factors that may be impacting the course ratings and enrollments, such as who is offering the course or the topic/category of it. Together, it will enable Coursera to identify ways to improve its existing courses and make data-driven decisions for the future.

DECISION TO BE IMPROVED:

- Help the Coursera curriculum development team to identify which courses are liked and not liked by the users and why.
- Help Coursera identify if there are positive or negative themes across courses that can help them grow their user base or help avoid user drop-offs from the platform.

WHO IS DECIDING?

Instructors and the Curriculum development team at Coursera can use this analysis to further improve the user experience on Coursera.

WHAT IS THE VALUE OF AN IMPROVED DECISION?

Quantitative:

- % Increased number of students
- % Increased student retention rate
- % Increased number of instructors or universities offering courses through Coursera
- Increased revenue for Coursera (25% - 30% revenue growth)

Qualitative:

- Improved course content
- Better student experience
- Increased reputational value of the Coursera brand.

Hypothesis:

The course rating is an important metric. Users giving a low rating to courses are more likely to drop out of it or even the Coursera platform. Further, users frequently use course ratings to decide whether to enroll in a particular course. So, it is of paramount importance for Coursera to improve on low ratings.

Our team believes that the course rating depends on various factors, such as the number of positive & negative sentiments of the reviews, the entity offering the course (university or company), and the topic or category of the course. Hence, by utilizing these factors, we can improve the decision on how to improve the ratings of the existing or invest in new courses that are likely to have higher ratings, thus improving user retention and revenue for Coursera. For example, assuming the course ratings and positive sentiment of the reviews are positively correlated, we can help Coursera increase the positive sentiment reviews by identifying and addressing the causes of negative feedback, thus, increasing the course rating. Similarly, if courses offered through universities or belonging to certain categories tend to have higher ratings, those could be possible avenues for Coursera to invest in future courses.

H_0: Course rating is independent of the number of positive & negative sentiments of the reviews, the entity offering the course (university or company), and the topic or category of the course.

H_a: Course rating depends on the number of positive & negative sentiments of the reviews, the entity offering the course (university or company), and the topic or category of the course.

Dependent Variable: Course Rating

Independent Variables: Sentiment score of a review, the entity offering the course (University or Company), and the topic or category of the course.

2. DATA COLLECTION

DATA SOURCES:

To conduct this analysis, we utilized data files such as Coursera course information and university rankings from Kaggle. The Coursera data is pertinent to this analysis because it includes information on Coursera course details with course names, an institute providing the course, and reviews. Additionally, we are utilizing the university ranking data file, which contains the academic rankings of universities worldwide.

The data is available at the URLs below:

- **Coursera_courses.csv**: This file contains course details, like course name, course id, course URL, and institution name offering the courses, for 623 courses. https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera
- **Coursera_reviews.csv**: This file contains course review details, like course id, review text, reviewer, review date, and rating. The original file has 1.45 million course reviews posted by students and participants between the years 2015 – 2020. https://www.kaggle.com/datasets/imuhammad/course-reviews-on-coursera
- **WORLD UNIVERSITY RANKINGS.csv**: This file contains the world ranking of universities. https://www.kaggle.com/datasets/whenamancodes/world-university-ranking-2022-2023
- **Coursera_courses_category.csv** : We manually created this file. In this file, we classified the 623 courses in the file Coursera_courses.csv into 11 subjects based on their course description.

DATA INTEGRATION:

After finalizing our dataset, we use Coursera_Courses.csv as the foundation and incorporate the additional information from the other files to create the Master dataset for our integration. Since we are integrating the data from multiple sources, there are mismatches in the University names with no specific pattern to the differences. Hence, we are matching the university names between the Coursera_courses.csv and WORLD UNIVERSITY RANKINGS.csv with the help of the Excel Vlookup function and manual updates. We joined files based on the course ID to gather information on course categories and reviews. As part of the cleaning steps, we drop the extra features and remove duplicate records and rows with blank reviews.

TEXT PREPROCESSING:

Raw text data is never clean. It could contain unwanted or unimportant text that could make it hard to understand and analyze the data and impact the accuracy of our model. So, in our dataset, we applied different preprocessing techniques to filter noisy and non-informative features from the reviews. In preprocessing, we changed the user reviews into lowercase. Also, we replaced "n't" with not, and removed non-alphabet characters and punctuation. We are also removing stopwords. For this, we created a custom stopword set that does not contain words like not, doesn't, etc. Besides this, we have also performed lemmatization on preprocessed reviews. Lemmatization converts every token or word to its meaningful base form, called Lemma.

At the end of the integration and preprocessing steps, the features from the master dataset we are using for our analysis are as follows:

Sno	Variable Type	Variable Name	Possible Values
1	Dependent Variable	CourseRating	1 to 5
2	Independent Variable	CleanReview	Text Data
3	Independent Variable	InstituteCategory	Company University
4	Independent Variable	CourseCategory	1. Data Science 2. Social Science 3. Personal Development 4. Information Technology 5. Business 6. Computer Science 7. Arts and Humanities 8. Physical Science and Engineering 9. Language Learning 10. Health 11. Math and Logic

Table 1: Dependent & Independent Variables used for the analysis.

TRADE-OFF:

When it comes to data source consideration, there are definitely some trade-offs to consider. On the one hand, you want access to as much data as possible to make well-informed decisions. However, you also need to be mindful of where that data comes from and whether it is reliable. It's important to balance these two factors to use the data available to you best. Ultimately, carefully considering data sources will help ensure your decisions are based on accurate and trustworthy information. Some data sourcing considerations include data quality, granularity, and cost aspects pertaining to our project are below.

DATA QUALITY

- Our data is public course content and reviews data without the reviewer's demographic and geographic details, which doesn't have privacy concerns or proprietary issues.
- The course reviews are only in the English language. We might lose a more significant number of reviewers in other languages.

We don't possess the most up-to-date review data since the dataset was from 2016 to 2020.

GRANULARITY

The dataset doesn't have the desired granularity in the course category to solve.

- The users' demand for each course and course category.
- The profitability of each course and course category.

COST

- We decided to label the 600~ courses by their category manually. The cost will increase as more courses are added to the platform.

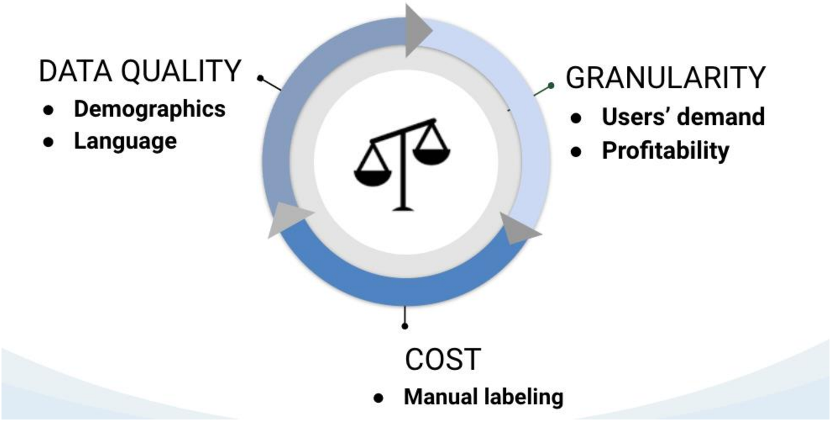


Figure 1: Data Source Trade-Offs

3. ANALYTIC PROBLEM TYPE AND TECHNIQUES

In our project, we are conducting a thorough analysis of the various factors that impact course ratings on the Coursera platform. The factors we are considering are:

- Student Review
- Institute Providing the course.
- Course Category

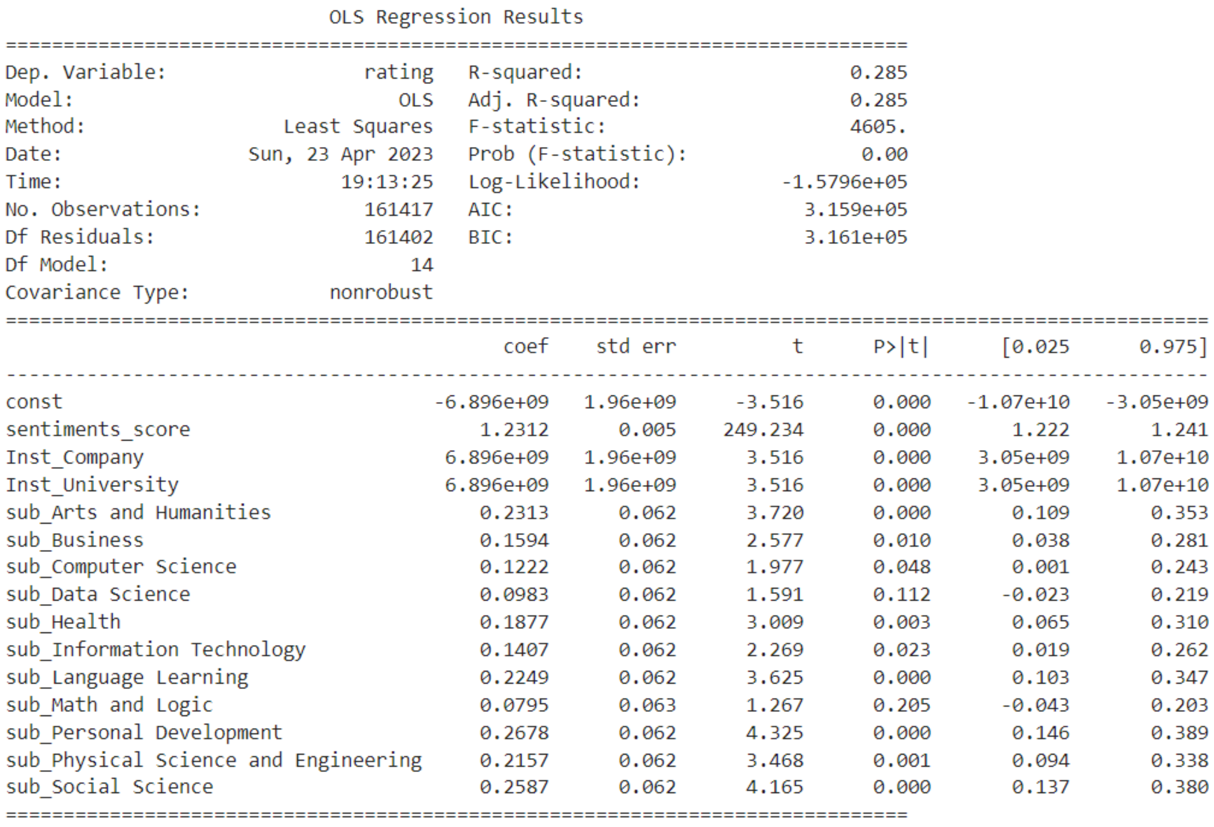
The analytics technique we are using on these factors are as follows:

Problem: To understand the impact of independent factors on course rating.

Analytic problem type: Regression

Analytic technique used: Multivariate Regression

We created a multivariate regression analysis model to understand the relationship between sentiment score of a review, the university offering the course and the course category with the course ratings. We determined the impact of these variables on the course ratings by analyzing the coefficient and significance level (p-value) associated with these variables in the model. It will help us identify universities or course categories having a higher chance of getting higher course ratings and use this data to optimize the courses offered by those universities or in those categories.The multivariate regression result is shown in Figure 2. The relationship between course rating and independent variables are significant with p <0.05 for most of the variables except for course category Data Science and Math& Logic.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rating   R-squared:                       0.285
Model:                            OLS   Adj. R-squared:                  0.285
Method:                 Least Squares   F-statistic:                     4605.
Date:                Sun, 23 Apr 2023   Prob (F-statistic):               0.00
Time:                        19:13:25   Log-Likelihood:            -1.5796e+05
No. Observations:              161417   AIC:                         3.159e+05
Df Residuals:                  161402   BIC:                         3.161e+05
Df Model:                          14
Covariance Type:            nonrobust
========================================================================================================
                                           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
const                                -6.896e+09   1.96e+09     -3.516      0.000   -1.07e+10   -3.05e+09
sentiments_score                         1.2312      0.005    249.234      0.000       1.222       1.241
Inst_Company                          6.896e+09   1.96e+09      3.516      0.000    3.05e+09    1.07e+10
Inst_University                       6.896e+09   1.96e+09      3.516      0.000    3.05e+09    1.07e+10
sub_Arts and Humanities                  0.2313      0.062      3.720      0.000       0.109       0.353
sub_Business                             0.1594      0.062      2.577      0.010       0.038       0.281
sub_Computer Science                     0.1222      0.062      1.977      0.048       0.001       0.243
sub_Data Science                         0.0983      0.062      1.591      0.112      -0.023       0.219
sub_Health                               0.1877      0.062      3.009      0.003       0.065       0.310
sub_Information Technology               0.1407      0.062      2.269      0.023       0.019       0.262
sub_Language Learning                    0.2249      0.062      3.625      0.000       0.103       0.347
sub_Math and Logic                       0.0795      0.063      1.267      0.205      -0.043       0.203
sub_Personal Development                 0.2678      0.062      4.325      0.000       0.146       0.389
sub_Physical Science and Engineering     0.2157      0.062      3.468      0.001       0.094       0.338
sub_Social Science                       0.2587      0.062      4.165      0.000       0.137       0.380
==============================================================================
```

Figure 2: Multivariate Regression Results

Problem: To understand the different positive and negative aspects of the Coursera courses.

Analytic Problem Type: Multiclass Classification

We performed multi-class classification to analyze courses and their associated sentiments across various aspects. It involves categorizing the courses into different sentiment categories, namely Negative, Positive, and Neutral, based on the sentiment expressed towards each aspect. Performing multi-class classification will enable us to categorize the courses into multiple classes, each representing a different sentiment towards a particular aspect. We can gain valuable insights into customer feedback and improve the quality of our courses accordingly.

Analytic Technique used:

- TOPIC MODELING (USING LDA WITH TFIDF VECTORIZER):

Our objective is to understand the sentiments (Positive / Negative / Neutral) of the users writing reviews. However, Machine Learning algorithms cannot comprehend classification rules from raw text alone and require numerical features to understand classification rules. Therefore, we are using feature engineering techniques to extract the features from user reviews and represent them in numerical form for modeling.

We used the TFIDF (Term Frequency - Inverse Document Frequency) function with n-grams as our feature engineering algorithm. Next, we used the topic modeling technique to classify the user reviews into key topics and determine the key aspects of each topic. These topics will provide crucial insights into the key areas of interest or concern related to the courses shared by the users. For instance, if a topic related to course materials consistently emerges as an area of concern in the reviews, Coursera can focus on improving the course material to avoid low-rated reviews in the future. To identify different aspects, we utilized the Latent Dirichlet Allocation (LDA) algorithm. Table 2 displays the top trigram sentences from each topic. Through analyzing these sentences, we identified patterns and themes that emerged from those topics. We then assigned a human-interpretable label to each topic and used that label as an aspect for Aspect Based Sentiment Analysis.

	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6
Top Trigrams seen under each topic.	▪ bad course ever ▪ not learn anything ▪ not like course	▪ quiz question not ▪ little bit tough ▪ little bit complicated	▪ instruction not clear ▪ not clearly explain ▪ assignment bit difficult	▪ video not good ▪ course material not ▪ poor audio quality	▪ waste time money ▪ not work properly	▪ peer grade assignment ▪ assignment not clear ▪ programming assignment not
ASPECTS	Course, content	Quiz	Assignment, Professor, Instructor	Material, Video, audio	Money	Assignment

Table 2: Identified Topics and Aspects

- ASPECT BASED SENTIMENT ANALYSIS USING VADER (VALENCE AWARE DICTIONARY FOR SENTIMENT REASONING)

Aspect-Based Sentiment analysis is an analytics technique that can help identify user sentiments (Positive, Negative, and Neutral) about a given aspect. For example: How does the user feel about the instructors on the course? The Aspect-based sentiment analysis technique can help determine that from the reviews.

We used the Aspect-based sentiment analysis technique to determine the user sentiments for each of the following aspects identified in the previous analysis: Course content, Instructor, Assignments or quizzes, Cost of the course, and Delivery method. The outcome of this analysis will help us determine the user sentiments about these specific aspects and provide a clear action plan for improvement. For example - If the user feedback is negative about the cost, it means Coursera needs to investigate the pricing for those courses.

We used the VADER algorithm for Aspect Based Sentiment Analysis. VADER is a rule-based sentiment analysis tool that can provide a polarity score for each sentence or phrase, as well as a sentiment intensity score for each aspect.

We selected these algorithms because they are straightforward, easy to understand, and do not require extensive training data or computational resources. Additionally, they are highly effective at processing large amounts of text data.

4. VISUALIZATION

It's important to keep track of course reviews, both positive and negative, to ensure that students are receiving the best online courses on Coursera. One way to do this is through visualization tools like box plots and bar graphs.

Box Plots can show the ratings distribution for a particular course, with the median, quartiles, and outliers clearly displayed. This can give educators a better understanding of how students perceive the course overall.

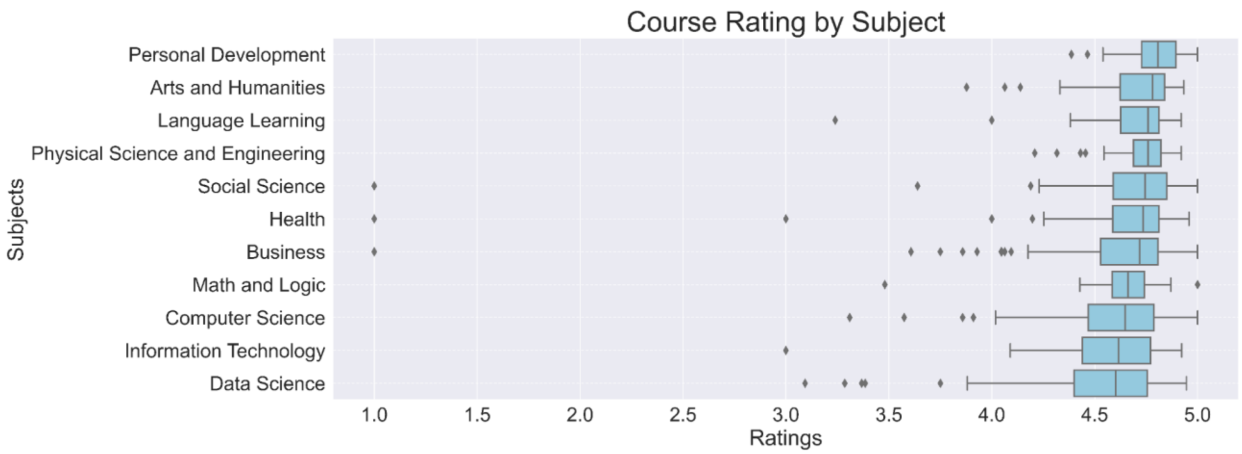


Figure 3: Course Rating by Course Subjects

In Figure 3, University Course Rating by Subjects, we used boxplots to visualize the course score by subject category. The box represents the middle 50% of the ratings, with the left being the first quartile (Q1) and the right of the box being the third quartile (Q3). The box length indicates the variability of the ratings in this range. Computer science, information technology, and data science have the longest boxes. Thus, there is a larger range of ratings in the middle ye 50%. Physical Science, Math/Logic, and Personal Development boxes are the shortest. Thus, the ranges are smaller. The whiskers extend from the box to the minimum and maximum values. Data science has the lowest rating among all the courses, whereas computer science has the top rating. Outliers are shown as individual diamond dots. Business, data science, and computer science are the categories with the most outliers. They may represent unusual or extreme values worth investigating, while Math/Logic has the fewest outliers. A line within the box represents the median. It indicates the midpoint of the dataset, with half the data above and half below. Course categories are sorted by median ratings.

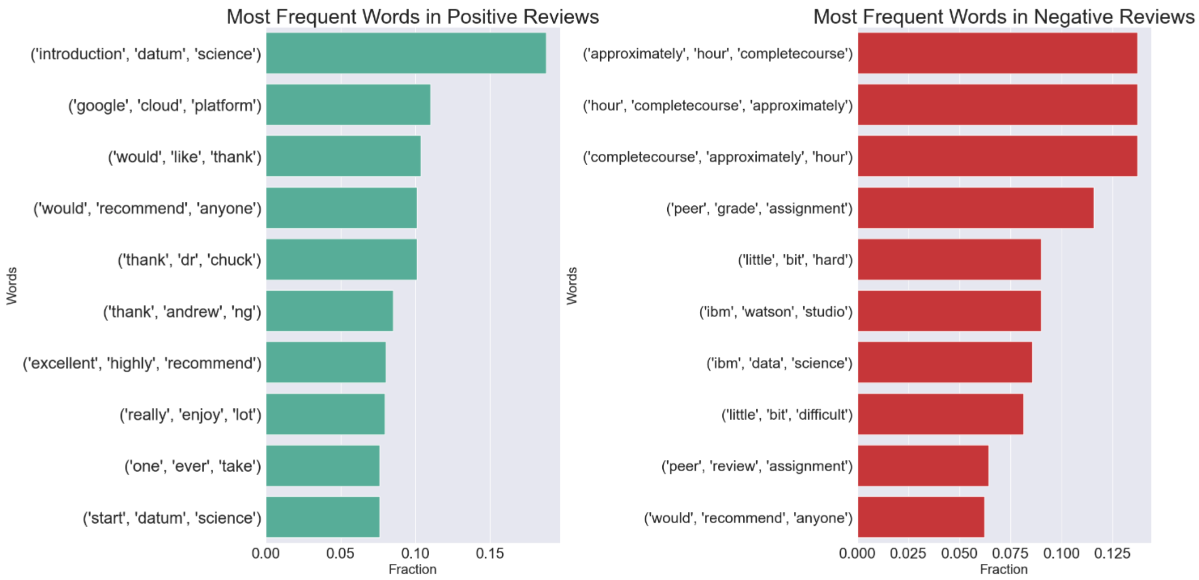


Figure 4: Trigrams of Top 20 Most Frequent Words in Positive and Negative Reviews

Bar graphs can be used to frequent word pairs in the positive and negative reviews, highlighting any trends or patterns that may emerge. This can help identify areas where improvements can be made and where resources should be allocated.

In Figure 4, Most Frequent Word Pairs by Sentiment, we used two Bar Plots to compare the fraction of the most frequently occurring trigrams in the positive and negative course reviews. The horizontal bar represents the fraction of the most frequent bigrams. The vertical axis represents the top 10 most occurring trigram pairs. Both bar charts follow a decreasing logarithmic decreasing trend. The feedback from the positive reviews is that the courses are a great introduction and students enjoy the class. The negative feedback ck is that the assignments are challenging and take hours to complete.

For various aspects, we are calculating a mean sentiment score for all the courses and classifying them into different sentiment categories. We can see the following results from the aspect-based sentiment analysis:

In Figure 5, we can see some good news and bad news. It is good that ~70% of the Coursera courses have no problematic aspects, i.e., on all the aspects that we analyzed above, the mean user sentiments were Positive. However, ~30% of the courses have at least one issue to address.

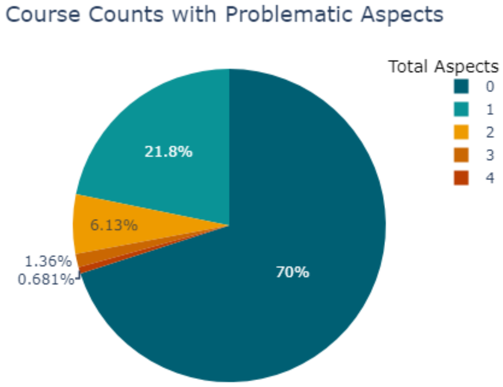


Figure 5: Percent of courses with Problematic aspects

In Figure 6, we can see that Money (or cost) is the most problematic aspect, i.e., for 70 out of 623 courses, users feel courses are priced higher than their value. Similarly, Audio / Video (or delivery) of courses is another area of feedback for Coursera to address. Lastly, improving the Quiz / Assignment experience is third on that list.

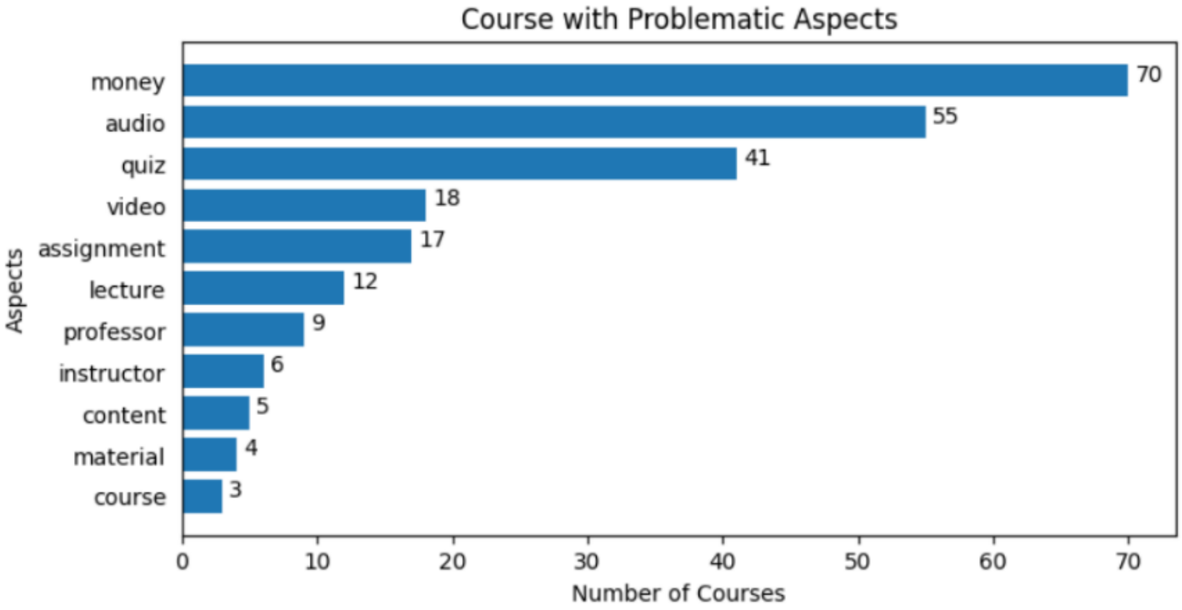


Figure 6: # of courses with Mean Negative sentiment for each aspect

In Figure 7, breaking down this feedback by the Course category, we can see that the majority of the Negative feedback is related to the Data Science courses (Money, Audio), Computer Science courses (Money, Audio), and Information Technology courses (Money, Quiz, Audio and Video). This also explains why these courses have the lowest median rating as shown in Figure 2.

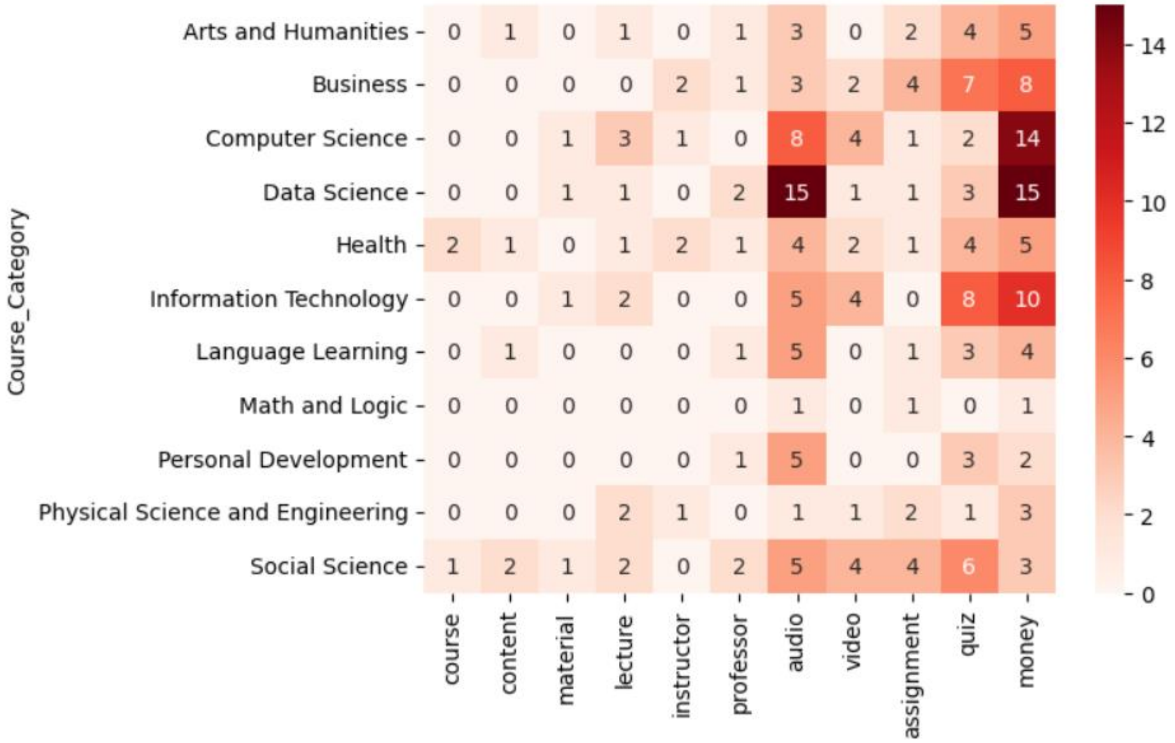


Figure 7: # of courses with Mean Negative sentiment for each aspect and course category

By using these visualization tools, educators can gain valuable insights into student feedback and take action to improve the quality of their courses continuously.

5. RESULTS AND RECOMMENDATION

Results:

Regression analysis

The results reject the null hypothesis and sustain the alternative hypothesis, confirming that the course ratings are dependent on the user sentiment mapped from the reviews. However, the Ordinary Least Squares Regression results show that rating is explained by the positive and negative sentiment scores (excluding neutral) only by 28.5% to 29.9%, depending on if we include categories and institutions as explanatory variables. The p-value is in all cases below 0.05 which indicates that the sentiment score relation with the rating is significant.

Course Ratings by Institution

Other results show that there is a difference of +0.08 points in the rating between courses offered by universities compared to courses offered by companies such as IBM. Also, the number of courses offered by universities exceeds five times the ones offered by companies.

Course Ratings by Categories

When analyzing the ratings by course category, the top five most offered categories are Business, Data Science, Social Science, Computer Science, and Health. However, the top five highest average rating categories are Personal Development, Arts and Humanities, Language Learning, Physical Science and Engineering, and Social Science.

Bigrams and Trigrams

Bigram frequency analysis shows that users associate bad experiences with courses being “difficult, understand”, “programming, assignments”, “waste, time”, “hard, understand” and “peer reviews”. At the same time, positive sentiments over courses are associated with “learn, lot”, “easy, understand”, “easy, follow”, “good, introduction”, “good, experience”, and “good, beginners”.

“Data Science” was identified as both positive and negative, so a Trigram was developed for further understanding.

Trigram frequency analysis (TFA) showed that the negative perception of “Data Science” is also associated with “IBM”. Also, “IBM Watson studio” is negatively perceived, so it may be possible that there is a course on IBM Watson studio that is not fulfilling the user's expectations. The TFA also showed that “little bit hard”, “peer grade assignment”, “peer review assignment” is paired with bad user experience.

On the other hand, “Introduction Data Science” and “google cloud platform” were identified as positive.

Topic Modeling

Six topics were mapped in the topic modeling. The analysis showed that courses were considered by users as: 1) irrelevant and disliked. 2) Complicated or difficult. 3) Unclear or badly explained. 4) W/ poor audio or video. 5) Waste of time. 6) Bad assignments. Each topic referred to different aspects related to, 1) money 2) video and audio 3) quizzes 4) lectures and professors 5) content & material.

The most important aspects were “money”, “audio” and “quiz”, and , ~30% of the courses have at least one issue to address.

Recommendations

By analyzing the presented results from a business perspective, we determined a series of recommendations that we consider could help Coursera to improve its value proposition, by improving the user experience which ultimately impacts their business opportunity.

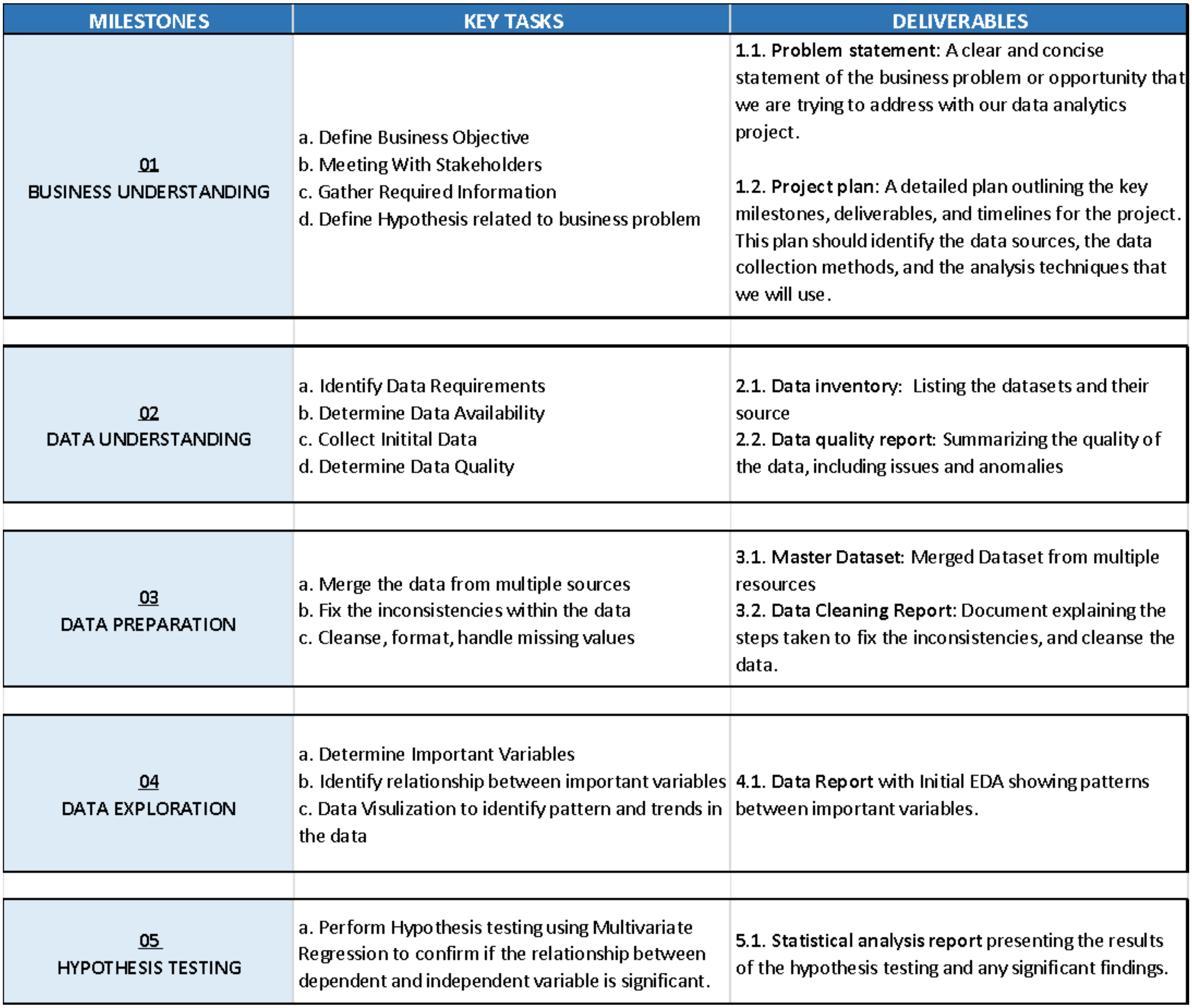
Based on the results we recommend to:

1. Continue focusing the courses on partnership with universities.
2. Explore opportunities to offer more courses related to Personal Development, Arts and Humanities, and Language Learning, as these are highly rated by users. We would also suggest exploring opportunities for introductory courses in Data Science, as it was positively perceived in the trigram analysis, and investigate why the IBM Watson course is associated with a negative sentiment.
3. A qualitative user research analysis could help understand the user experience in more detail and help identify specific insights that could complement this quantitative analysis.
4. Make courses easy to onboard, easy to follow and to understand.
5. Make courses shorter in time.
6. Reduce the workload on assignments.
7. It would be recommendable to have several short consecutive courses rather than one long course covering many topics. Disaggregating courses into different difficulty levels could help users to specifically enroll in what interests them. This would help prevent bad experiences from users who enroll in courses that exceed their level of knowledge. This also brings an opportunity to offer more courses and explore ways of doing price segmentation adjusted to users' proficiency levels.
8. Improvements should be made to the audio quality of lectures in Data Science and Computer Science courses.
9. The most problematic aspect is Money, indicating dissatisfaction with the cost or value of the product or service. Coursera can tackle this problem by finding ways to lower costs or providing discounts. Additionally, improving course features or quality may help increase its perceived value and justify its price.
10. Train Instructors and professors on courses where these terms are negatively perceived.

6. VESRION 2

6.1. MILESTONE AND DELIVERABLES

There is a Gantt chart in Figure 8 that shows a timeframe for executing the project plan. We outline the major project phases for version 2 along with the milestones and deliverables that go along with them in table 3.

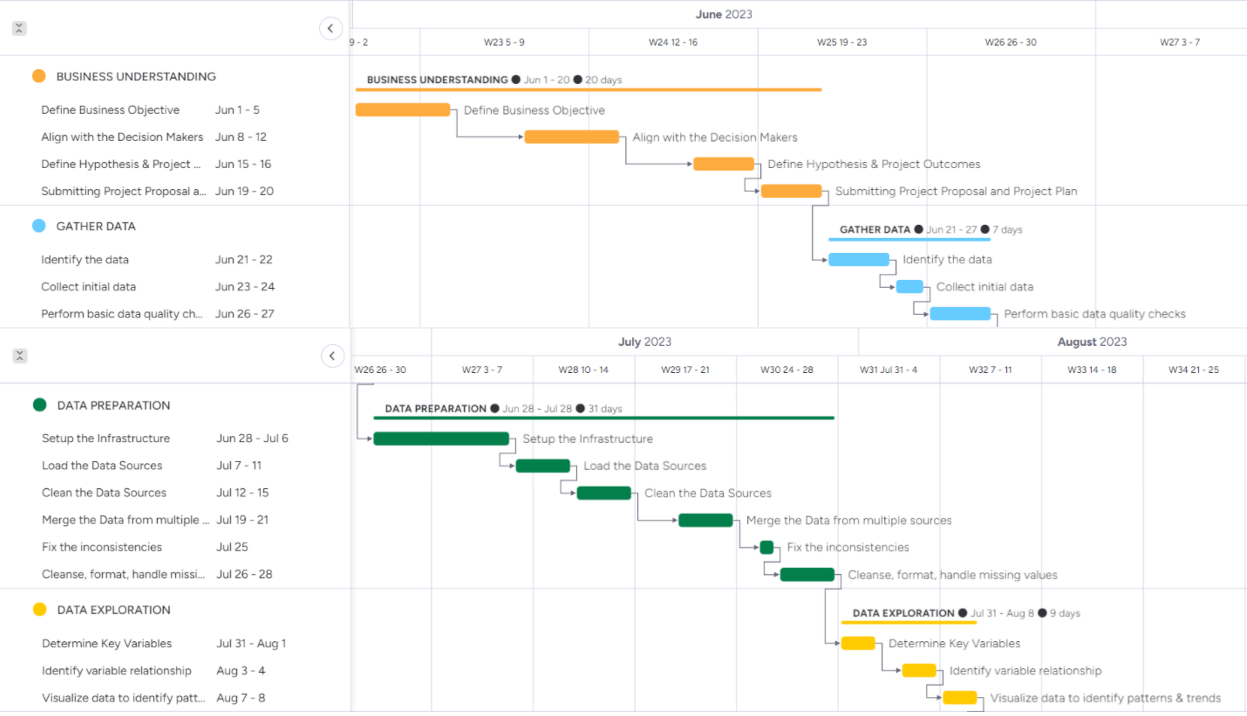


MILESTONES	KEY TASKS	DELIVERABLES
01 BUSINESS UNDERSTANDING	a. Define Business Objective b. Meeting With Stakeholders c. Gather Required Information d. Define Hypothesis related to business problem	**1.1. Problem statement**: A clear and concise statement of the business problem or opportunity that we are trying to address with our data analytics project. **1.2. Project plan**: A detailed plan outlining the key milestones, deliverables, and timelines for the project. This plan should identify the data sources, the data collection methods, and the analysis techniques that we will use.
02 DATA UNDERSTANDING	a. Identify Data Requirements b. Determine Data Availability c. Collect Initital Data d. Determine Data Quality	**2.1. Data inventory**: Listing the datasets and their source **2.2. Data quality report**: Summarizing the quality of the data, including issues and anomalies
03 DATA PREPARATION	a. Merge the data from multiple sources b. Fix the inconsistencies within the data c. Cleanse, format, handle missing values	**3.1. Master Dataset**: Merged Dataset from multiple resources **3.2. Data Cleaning Report**: Document explaining the steps taken to fix the inconsistencies, and cleanse the data.
04 DATA EXPLORATION	a. Determine Important Variables b. Identify relationship between important variables c. Data Visulization to identify pattern and trends in the data	**4.1. Data Report** with Initial EDA showing patterns between important variables.
05 HYPOTHESIS TESTING	a. Perform Hypothesis testing using Multivariate Regression to confirm if the relationship between dependent and independent variable is significant.	**5.1. Statistical analysis report** presenting the results of the hypothesis testing and any significant findings.

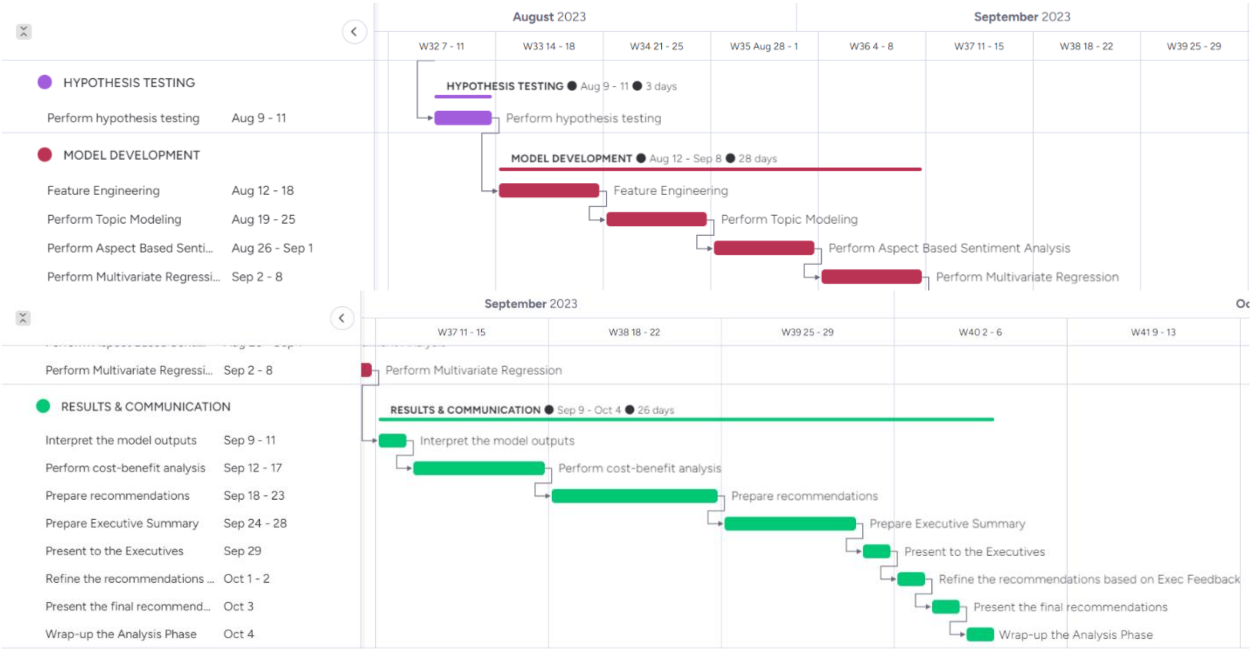
MILESTONES	KEY TASKS	DELIVERABLES
06 MODEL DEVELOPMENT	a. Feature engineering to convert raw text to important features for modeling. b. Use Analytic Technique to get the sentiment polarity from review for various aspect: ▪ Topic Modeling (LDA) ▪ Aspect Based Sentiment Analysis(VADER) c. Use analytic technique to understand the impact of various factors on course rating ▪ Multivariate Regression	**6.1. Model development report:** This report contain details of the entire model development process, including the feature engineering techniques used, the sentiment polarity and impact of factors reports, and the modeling techniques employed.
07 COMMUNICATING FINAL RESULTS	a. Interpret the model outputs b. Perform cost-benefit analysis c. Prepare recommendations d. Prepare Executive Summary e. Present to the Executives f. Refine the recommendations based on Exec Feedback g. Present the final recommendations Prepare a final report or presentation h. Wrap-up the Analysis Phase	**7.1. Final report or presentation** summarizing the project results and insights

Table 3: Version 2 Milestone and Deliverables

Gantt Chart for Project Plan:



● BUSINESS UNDERSTANDING	
Define Business Objective	Jun 1 - 5
Align with the Decision Makers	Jun 8 - 12
Define Hypothesis & Project ...	Jun 15 - 16
Submitting Project Proposal a...	Jun 19 - 20
● GATHER DATA	
Identify the data	Jun 21 - 22
Collect initial data	Jun 23 - 24
Perform basic data quality ch...	Jun 26 - 27
● DATA PREPARATION	
Setup the Infrastructure	Jun 28 - Jul 6
Load the Data Sources	Jul 7 - 11
Clean the Data Sources	Jul 12 - 15
Merge the Data from multiple ...	Jul 19 - 21
Fix the inconsistencies	Jul 25
Cleanse, format, handle missi...	Jul 26 - 28
● DATA EXPLORATION	
Determine Key Variables	Jul 31 - Aug 1
Identify variable relationship	Aug 3 - 4
Visualize data to identify patt...	Aug 7 - 8



HYPOTHESIS TESTING	
Perform hypothesis testing	Aug 9 - 11
MODEL DEVELOPMENT	
Feature Engineering	Aug 12 - 18
Perform Topic Modeling	Aug 19 - 25
Perform Aspect Based Senti...	Aug 26 - Sep 1
Perform Multivariate Regressi...	Sep 2 - 8
Perform Multivariate Regressi...	Sep 2 - 8
RESULTS & COMMUNICATION	
Interpret the model outputs	Sep 9 - 11
Perform cost-benefit analysis	Sep 12 - 17
Prepare recommendations	Sep 18 - 23
Prepare Executive Summary	Sep 24 - 28
Present to the Executives	Sep 29
Refine the recommendations ...	Oct 1 - 2
Present the final recommend...	Oct 3
Wrap-up the Analysis Phase	Oct 4

August 2023
September 2023
W32 7 - 11
W33 14 - 18
W34 21 - 25
W35 Aug 28 - 1
W36 4 - 8
W37 11 - 15
W38 18 - 22
W39 25 - 29
HYPOTHESIS TESTING ● Aug 9 - 11 ● 3 days
Perform hypothesis testing
MODEL DEVELOPMENT ● Aug 12 - Sep 8 ● 28 days
Feature Engineering
Perform Topic Modeling
Perform Aspect Based Sentiment Analysis
Perform Multivariate Regression
September 2023
Oc
W40 2 - 6
W41 9 - 13
RESULTS & COMMUNICATION ● Sep 9 - Oct 4 ● 26 days
Interpret the model outputs
Perform cost-benefit analysis
Prepare recommendations
Prepare Executive Summary
Present to the Executives
Refine the recommendations based on Exec Feedback
Present the final recommendations
Wrap-up the Analysis Phase

Figure 8: Gantt Chart for Project Plan

6.2. RESOURCES

Staffing Requirements

Estimated staffing requirements for the duration of the **4 months project**:

- Sr. Manager
- Sr. Data scientist
- Jr. Data scientist
- Sr. Functional Business Analyst
- Jr. Business Data Analyst

Staffing Requirement	Task Description	Experience Required	Hours per Week	Number of weeks
Sr. Manager	Project leader and supervisor. Overseeing the fulfillment of project goals, coordinating tasks between the staff members, aligning expectations with stakeholders. Planning and presenting Upselling opportunities.	10 years	9	12 weeks
Sr. Data Scientist	Lead technical Consultant. With significant experience on leading analytics projects and communicating with clients.	> 5 years	40	12 weeks

	Responsible for designing the analytics models with clear understanding of the business goals. Leads and works on the implementation and deliverables with the Semi Senior Data Scientist.			
Jr. Data Scientist	Individual contributor. With experience in developing analytics models. Doesn't need to interact much with the client. Needs to have a good understanding of models and techno cal aspects.	3-5 years	40	12 weeks
Sr. Functional Business Analyst.	Lead Business Consultant and Subject Matter expert. Industry Specialist. Responsible for mapping the client requirements and having a clear understanding of the business needs and decisions to be made. Is in periodic contact with the client and with the Sr. Data Scientist to ensure that the solution developed is aligned with the client expectations and the problem to be solved.	> 5 years	40	12 weeks
Jr. Business Data Analyst	Individual contributor. Works under the guidelines of the Functional Business Analyst. Has a technical background but with a good understanding of business analysis. Analyses data looking for insights and business opportunities.	3-5 years	40	12 weeks

6.3. RISKS AND MITIGATION

Likelihood: refers to the probability of a specific risk event occurring. It helps in understanding the chances of a risk event happening and informs the decision-making process regarding risk mitigation.

Description of Risk: provides a clear and detailed explanation of the specific risk being assessed. It includes identifying the potential hazards, vulnerabilities, or uncertainties that could lead to negative consequences or impacts.

Severity: Severity refers to the potential impact or harm that may result from a risky event. Understanding the severity of risk allows organizations to prioritize their efforts and allocate appropriate resources to manage the risk effectively.

Response: refers to the actions or measures taken to address an identified risk. It involves developing and implementing strategies to mitigate, transfer, accept, or avoid the risk.

Role: involves identifying the responsibilities and tasks assigned to individuals or teams responsible for managing and reducing risks. Effective roles and responsibilities ensure that risk mitigation efforts are well-coordinated and aligned with the organization's objectives.

Likelihood	Description of Risk	Severity	Response	Role
High	**Project Selection**: There is a risk that the selected topic may not align with the current market demand or users' needs.	**Low**	We have established mechanisms for continuous improvement based on users' reviews. We will regularly access the platform's performance.	Project Manager, Business Analyst, Data Scientist
Medium	**Time Constraint**: We faced a tight schedule, squeezed between completing Homework 3 and preparing for the final presentation. This time constraint posed a significant challenge as we needed ample time to implement changes based on the feedback, we received for Homework 3.	**High**	Recognizing the significance of stakeholder engagement and the need for timely feedback, we proactively reached out to key stakeholders. Their input and insights were invaluable in shaping our project's direction and ensuring its alignment with the desired outcomes.	Project Manager, Data Scientist
High	**Data Quality**: We encountered an unexpected issue with data quality- discovering duplicated rows within our dataset, which not only compromised the integrity of our findings but also significantly reduced the size of our dataset from approximately 1.2 million entries to just 300,000. Additionally, we discovered foreign language reviews within the dataset, further complicating our analysis.	**Low**	We meticulously reviewed the dataset and identified duplicated rows that could potentially skew the results. Removing these duplicates allowed us to work with a cleaner and more reliable dataset, enabling us to make accurate assessments of sentiment. Furthermore, we made the decision to simplify the model by removing foreign languages.	Business Analyst
High	**Technical**: We encountered a technical challenge when multiple team members simultaneously worked on the same Google PowerPoint presentation in the shared drive. This resulted in a revision history issue, making it difficult to track and consolidate everyone's contributions efficiently.	**Low**	We emphasized clear and effective communication among team members to ensure everyone was aware of ongoing revisions and updates. Additionally, we retrieved previous versions and history copies of the document, allowing us to restore any unintentional modifications and preserve the project's progress.	Data Scientist, Business Analyst, Project manager

Medium	**Modeling**: We encountered difficulties in accurately modeling this linear relationship between sentiment reviews and ratings.	**High**	We showed the statistical findings In Section 3 of this paper, shedding light on the patterns and trends observed in the sentiment reviews. However, we recognized that relying solely on this approach might not provide us with a comprehensive understanding of the data. To address this limitation, we embraced other modeling approaches, such as topic modeling, to gain deeper insights into the sentiments expressed.	Data Scientist
Low	**Human Resource**: One of our dedicated team members had to take a temporary leave from the project due to unforeseen personal circumstances. This unexpected absence left a void in our team and created additional pressure on the remaining members to fulfill the responsibilities and maintain the project's momentum.	**Low**	We had to adapt quickly, redistribute tasks, and provide support to ensure that the project continued to progress despite the temporary setback.	Project Manager

6.4. COMMUNICATING PROJECT PROGRESS

To ensure good internal communication within a project team, roles and responsibilities for each team member will be clearly defined. It is also important to create a communication plan that outlines the frequency, channels, and content of internal communication. This should include regular team meetings (weekly or more frequently), progress updates, peer reviews and brainstorming sessions. The pace of communication will be monitored and adjusted as needed, with also an implemented open-door policy to promote trust and transparency within the team.

By following the below guidelines, we can ensure that stakeholders receive detailed and relevant updates throughout the project's duration, keeping them informed and engaged.

WHEN:

- Weekly updates with the Client Project Owner to discuss project details, tasks, and immediate concerns.
- Monthly strategic meetings with Senior Managers and Client Managers for exploring new opportunities, attended by Lead Senior Consultants as needed.
- Upon completion of each milestone or as initially established in the Project Plan, which outlines the frequency, format, and content of project progress updates.

HOW:

- Conduct regular meetings to provide comprehensive updates on the project's status.
- Perform frequent reviews to ensure the project's approach and interim findings align with the desired state, adjusting as necessary.
- Customize communication methods based on stakeholders' preferred channels (status report, email, dashboard)
- Adjust communication content and frequency according to the RACI chart, ensuring relevant information reaches the appropriate parties.

WHAT:

- Emphasize significant achievements or milestones reached, celebrating successes, and recognizing team contributions.
- Incorporate flowcharts, presentations, or other visual aids to effectively convey the project's status and progress.
- Discuss delays, emerging issues, identified risks, and their potential impacts on project timelines, along with proposed mitigating actions.
- Share updates on any changes to the budget or resource allocation, explaining their implications for the project timeline and overall objectives.
- Maintain records of meetings and Minutes of the Meeting for documentation purposes, enabling easy access and retrieval for future reference.

7. SUCCESS MEASUREMENTS AND PROJECT IMPACT

Our models and recommendations aim to identify factors that affect customer churn and satisfaction to aid Coursera in increasing revenue, decreasing churn, attract, and retain new customers. Thus, measuring the success and impact of this project will involve evaluating the implementation of our recommendations and assessing their effectiveness on Coursera. Relevant metrics for this include:

- Pre- and post-project metrics like the number of successful course launches, user retention rates, and overall user satisfaction.
- Cost-effectiveness of implementing the project's recommendations by comparison of costs incurred (e.g., implementing recommendations, training instructors) to the benefits gained (e.g., increased user retention, higher course ratings)
- Pre- and post-project metrics for comparison of percentage of correct decisions and the time taken to make them by Coursera’s management and course developers.
- Return on investment in terms of increased revenue, user growth, and improved course ratings.

Because the Topic Modeling results show that user experience is heavily influenced by the quality of course instructors, course materials and course quality, we must evaluate the impact of our project on instructors and partner institutions, ensuring that they are well-supported in their efforts to improve the quality of their courses. It also helps to gauge the project's effect on their reputation and relationship with Coursera. The impact can be measured by:

- Improvements in student satisfaction measured through improved course ratings.
- Course completion rates
- Sentiment towards the courses offered by the partner institutions through ratings and reviews.

- Enrollment rates in courses offered by the instructor/partner institution.

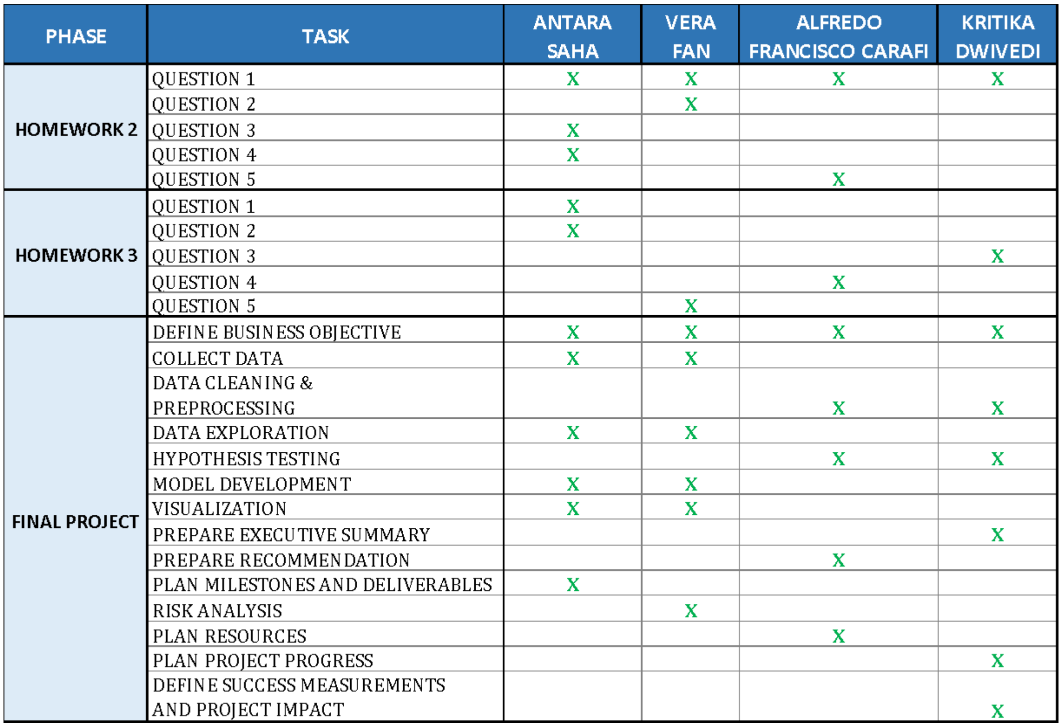
It is also important to ensure that the project has successfully addressed the concerns and pain points of the end-users, leading to improved user satisfaction and retention. By identifying any issues or areas for further improvement for Learners, we aid Coursera in improving user satisfaction and retaining their current customers. The impact can be measured through the following metrics:

- Improvements in course ratings and reviews
- Retention rates
- Course completion rates
- Employability tracking

Criteria to determine when the project is completed:

- Model implementation and performance: All recommendations have been implemented, and their impact on course ratings, user retention, and satisfaction has been assessed and found to be significantly better than previous solutions. The model is accurate, timely and shows significant values for R-squared, precision etc.
- Stakeholder satisfaction: Stakeholders, including Coursera management, course instructors, and partner institutions, have provided feedback on the project outcomes, and their concerns have been addressed. Can be measured by Customer Satisfaction Score (CSAT) expected to be above average (8 – on a scale of 1 to 10).
- Future growth potential: The project has identified and initiated follow-on opportunities for further improvement and growth, measured through assessment of the number of follow-on projects, collaborations, or referrals resulting from the project.

8. DIVISION OF LABOR



PHASE	TASK	ANTARA SAHA	VERA FAN	ALFREDO FRANCISCO CARAFI	KRITIKA DWIVEDI
HOMEWORK 2	QUESTION 1	X	X	X	X
	QUESTION 2		X		
	QUESTION 3	X			
	QUESTION 4	X			
	QUESTION 5			X	
HOMEWORK 3	QUESTION 1	X			
	QUESTION 2	X			
	QUESTION 3				X
	QUESTION 4			X	
	QUESTION 5		X		
FINAL PROJECT	DEFINE BUSINESS OBJECTIVE	X	X	X	X
	COLLECT DATA	X	X		
	DATA CLEANING & PREPROCESSING			X	X
	DATA EXPLORATION	X	X		
	HYPOTHESIS TESTING			X	X
	MODEL DEVELOPMENT	X	X		
	VISUALIZATION	X	X		
	PREPARE EXECUTIVE SUMMARY				X
	PREPARE RECOMMENDATION			X	
	PLAN MILESTONES AND DELIVERABLES	X			
	RISK ANALYSIS		X		
	PLAN RESOURCES			X	
	PLAN PROJECT PROGRESS				X
	DEFINE SUCCESS MEASUREMENTS AND PROJECT IMPACT				X

Figure 9: Task Distribution among team members