Analyzing the Factors Affecting Course Ratings in Coursera

Presented By:

Antara Saha,Vera Fan,Alfredo Francisco Carafi,Kritika Dwivedi

MEET THE FULL TEAM

Antara Saha

Master’s of Science in Information Technology (Business Intelligence Data Analytics)

Vera Fan

Master’s of Science in Information Technology (Business Intelligence Data Analytics)

Alfredo Francisco Carafi

Master’s of Science in Public Policy and Management (Digital Transformation)

Kritika Dwivedi

Master of Information Systems Management - 12 Month

AGENDA

- **VERSION 1:**
  - Executive Summary
  - Problem Framing
  - Hypothesis
  - Gathering and exploring data
  - Analytic Problem Type and Technique
  - Visualization
  - Result and Recommendation
- **VERSION 2:**
  - Milestones and Deliverables
  - Resources
  - Risks and Mitigation
  - Project Progress

VERSION 1

EXECUTIVE SUMMARY

Analysis of approx. 1 million rows dataset to help Coursera maximize its course ratings, thus increasing its enrollments, revenue, and brand value.

Analytic Problem Type: Regression, Multiclass Classification

Analytic Techniques used: Topic Modeling, Aspect Based Sentiment Analysis, Multivariate Regression

Major Takeaways:

- Course ratings are dependent on User sentiment, topics affecting user experience should be addressed (difficulty of course, instructor capability, clear instructions, affordability)
- Difference in the average rating between courses offered by universities compared to courses offered by companies is not statistically significant
- Top 3 highest average rated course categories - Personal Development, Arts and Humanities, and Language Learning.
- 2 highest offered course categories - Data Science and Business, but with average ratings below median rating within all the categories.

PROBLEM FRAMING

- Online learning platforms is an excellent way for students and professionals to expand their knowledge and skill sets at a convenient pace and much less cost than the universities.
- Coursera is one such platform. It offers thousands of courses from 200+ world-class universities and companies.
- User ratings and reviews are key factors in measuring the success of a course, as courses with negative reviews and low ratings tend to lead to dropouts and lower enrollments in the future.
- Analyzing the relationship between the course reviews, the institute offering the courses, and the course category with the course ratings.
- It will enable Coursera to identify ways to improve its existing courses and make data-driven decisions for future new courses.

PROBLEM FRAMING(cont.)

- **Decision To be Improved**
  - Help 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.
- **Key Decision Makers**
  - Instructors and the Curriculum development team at Coursera can use this analysis to further improve the user experience on Coursera.
- **Value of 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

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: Number of positive & negative sentiments of the reviews, the entity offering the course (University or Company), and the topic or category of the course.

DATA SOURCES

Ingested from **Kaggle website** and developed by team.

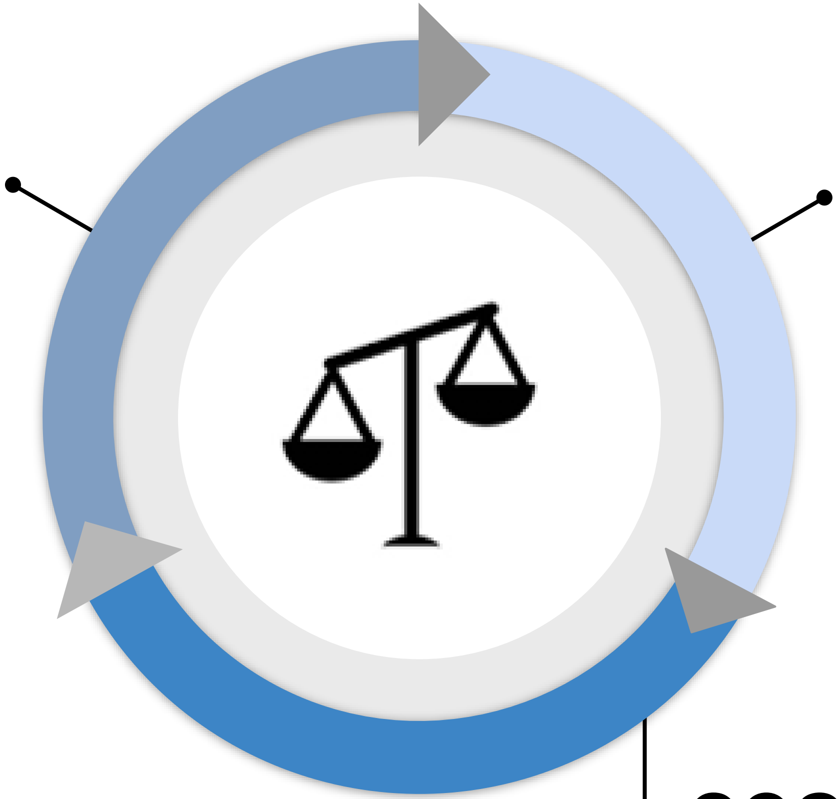
1) **Coursera Courses:** contains course details, like course name, course id, course URL, and institution name offering the course, for 623 courses (https://www.kaggle.com/datasets/imuhammad/course-reviews-on- coursera)

2) **Coursera Reviews**: contains course review details, like course id, review text, reviewer, review date, and rating (https://www.kaggle.com/datasets/imuhammad/course-reviews-on- coursera)

3) **World University Rankings**: contains the world ranking of universities (https://www.kaggle.com/datasets/whenamancodes/world-university-ranking-2022-2023)

4) **Coursera Courses Category**: we manually created this file, where we classified the 623 courses into 11 subjects based on their course description. This file also contain course id as master key.

TRADE-OFF



DATA QUALITY

➔ Public

- without the reviewer’s demographic and geographic details
- no privacy concerns or proprietary issues.

➔ Language

- reviews are only in the English language.
- might lose reviewers in other languages.

➔ Date

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

GRANULARITY

➔ Users’ demand

- users' demand for each course and course category.

➔ Profitability

- profitability of each course and course category.

COST

➔ Manual labeling

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

DATA INTEGRATION & PREPROCESSING

STEPS	DESCRIPTIONS	OUTPUT
STEP 1: RAW DATA COLLECTION	▪ Coursera Course Details ▪ Coursera Reviews ▪ University Ranking ▪ Course Category	▪ Coursera_courses.csv ▪ Coursera_reviews.csv ▪ WORLD UNIVERSITY RANKINGS.csv ▪ Coursera_courses_category.csv
STEP 2: MERGE & CLEAN DATA	▪ Coursera_courses.csv is the base dataset. ▪ Join the data based on primary key(course_id). ▪ Delete Duplicate records. ▪ Removing records with blank reviews.	▪ Coursera_Temp_Merge_clean_Data.csv
STEP 3: DATA PREPROCESSING	▪ Lowercasing character. ▪ Removing special character, non-alphabets, stop words and extra whitespaces. ▪ Perform lemmatization.	▪ Coursera_Final_Merge_clean_Data.csv

ANALYTIC PROBLEM TYPE AND TECHNIQUE

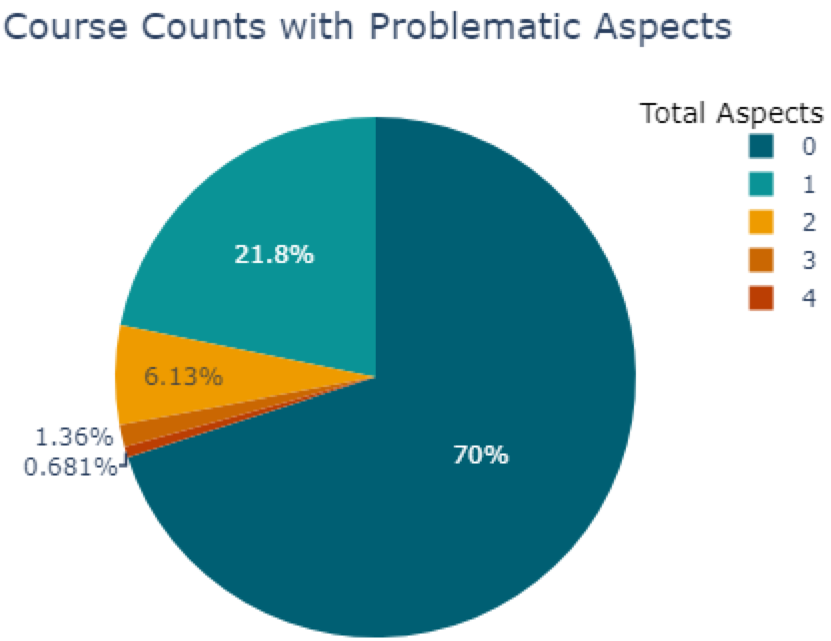
- **Problem Type:**
  - Regression
  - Multiclass Classification

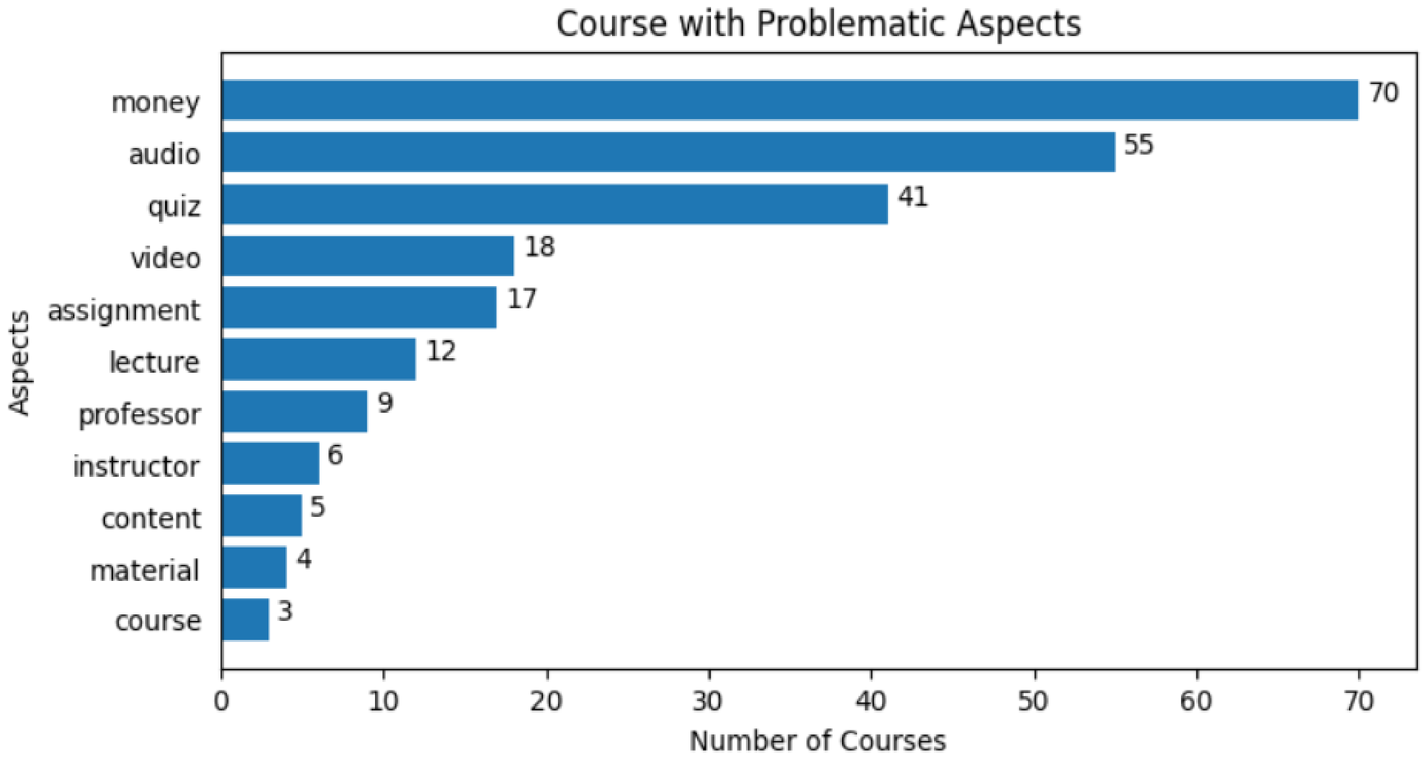
- **Analytic Techniques Used:**
  - Multivariate Regression
  - Topic Modeling – Using LDA,TFIDF Vectorizer
  - Aspect Based Sentiment Analysis – Using Vader (Valence Aware Dictionary for Sentiment Reasoning)

Common Themes in Negative Feedback and Associated Aspects

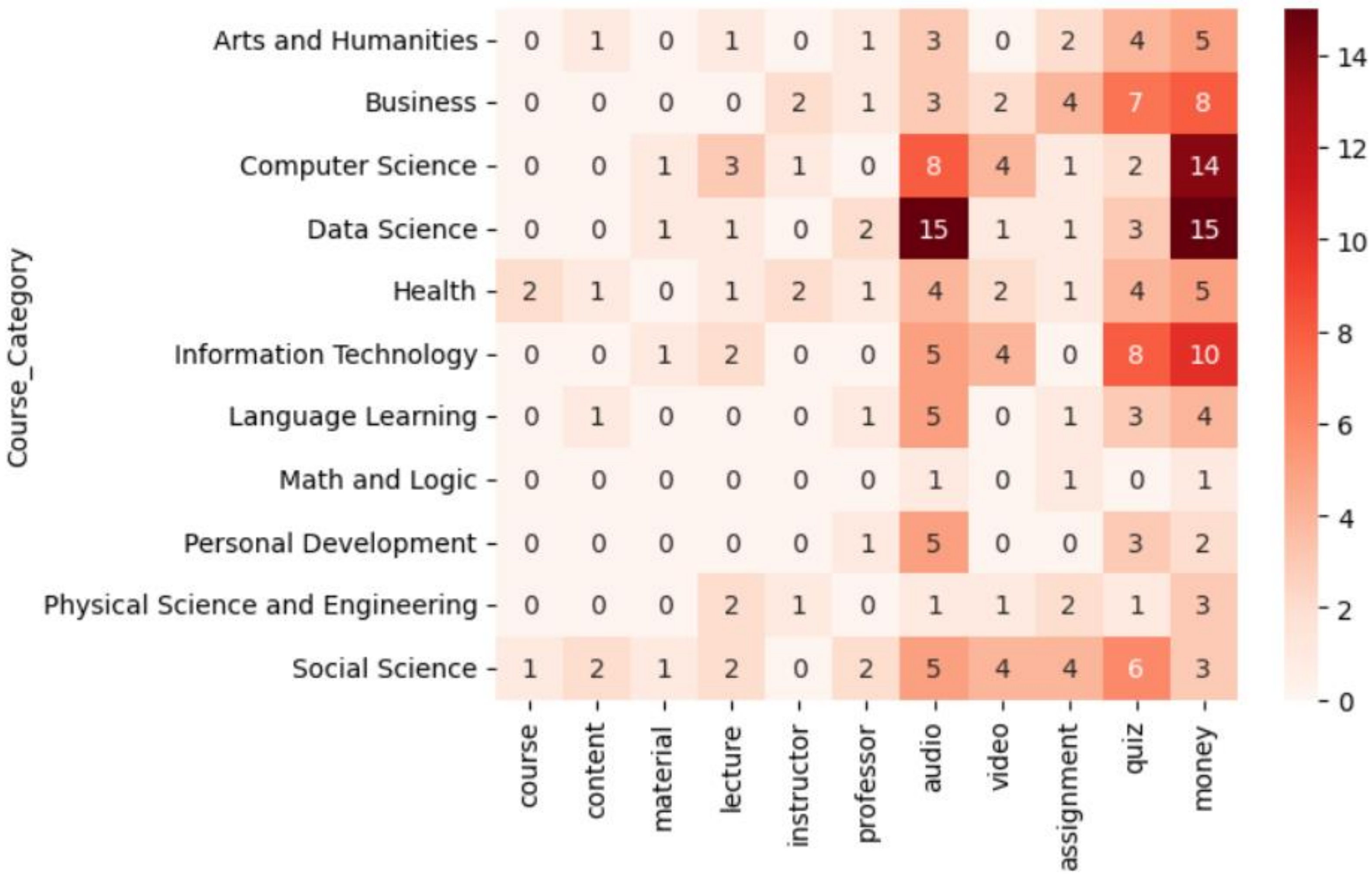
	TOPIC 1	TOPIC 2	TOPIC 3	TOPIC 4	TOPIC 5	TOPIC 6
Top Trigrams seen under each topic.	▪ bad course ever ▪ not learn anything ▪ not like course	▪ quiz question not ▪ little bit tough ▪ little bit complicated	▪ instruction not clear ▪ not clearly explain ▪ assignment bit difficult	▪ video not good ▪ course 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

We find ~30% of the courses have mean negative sentiment for at least one of the aspect, and Money (or cost) is the top concern

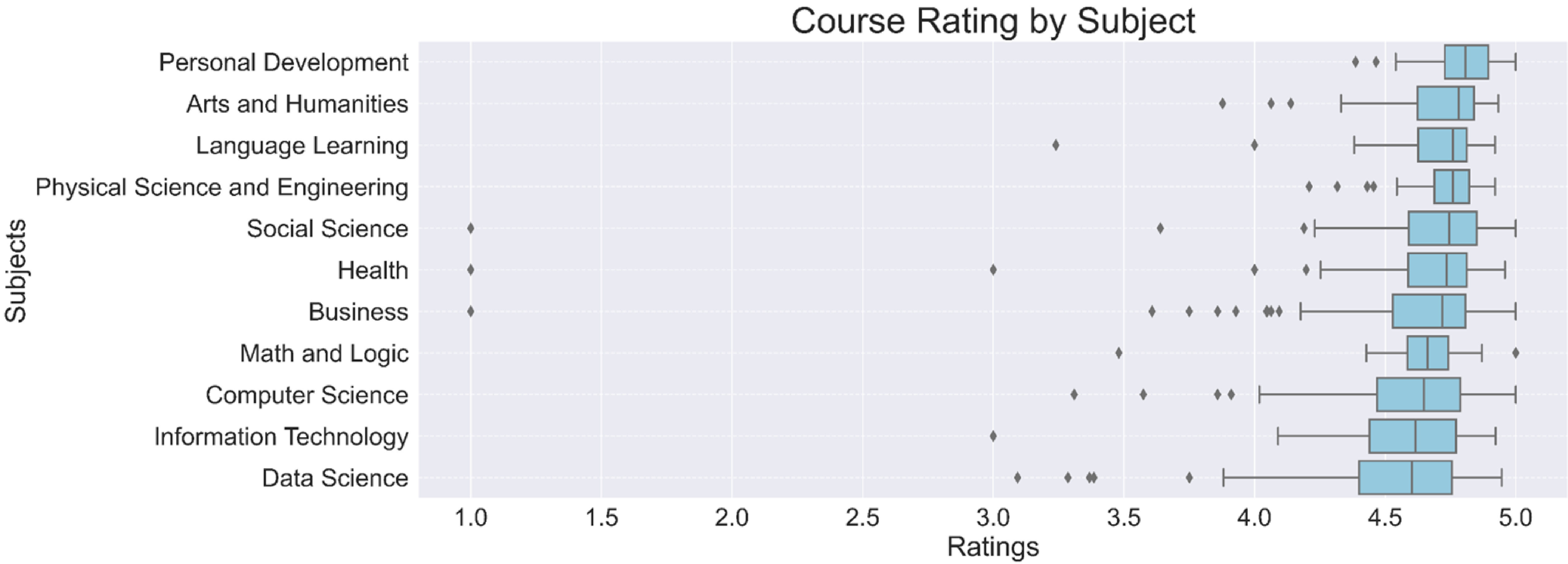




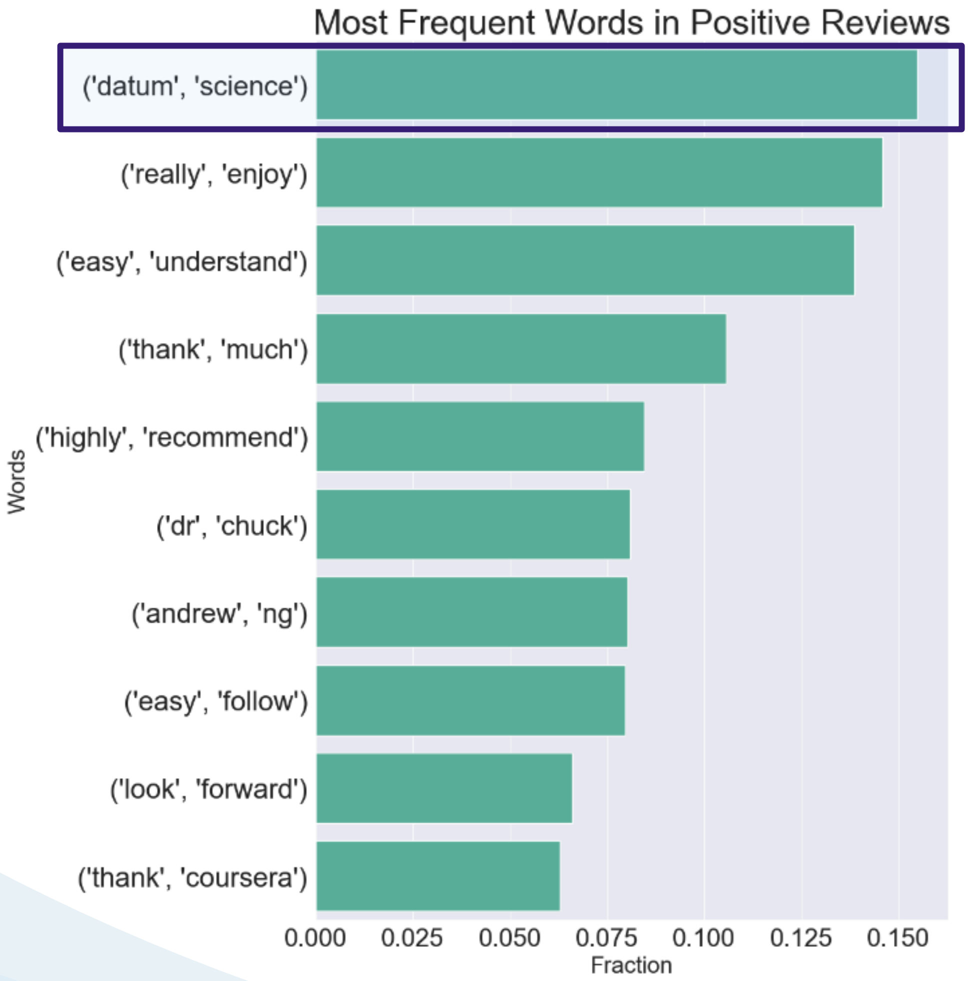
Further, majority of these courses belong Data Science, Computer Science, and Information Technology

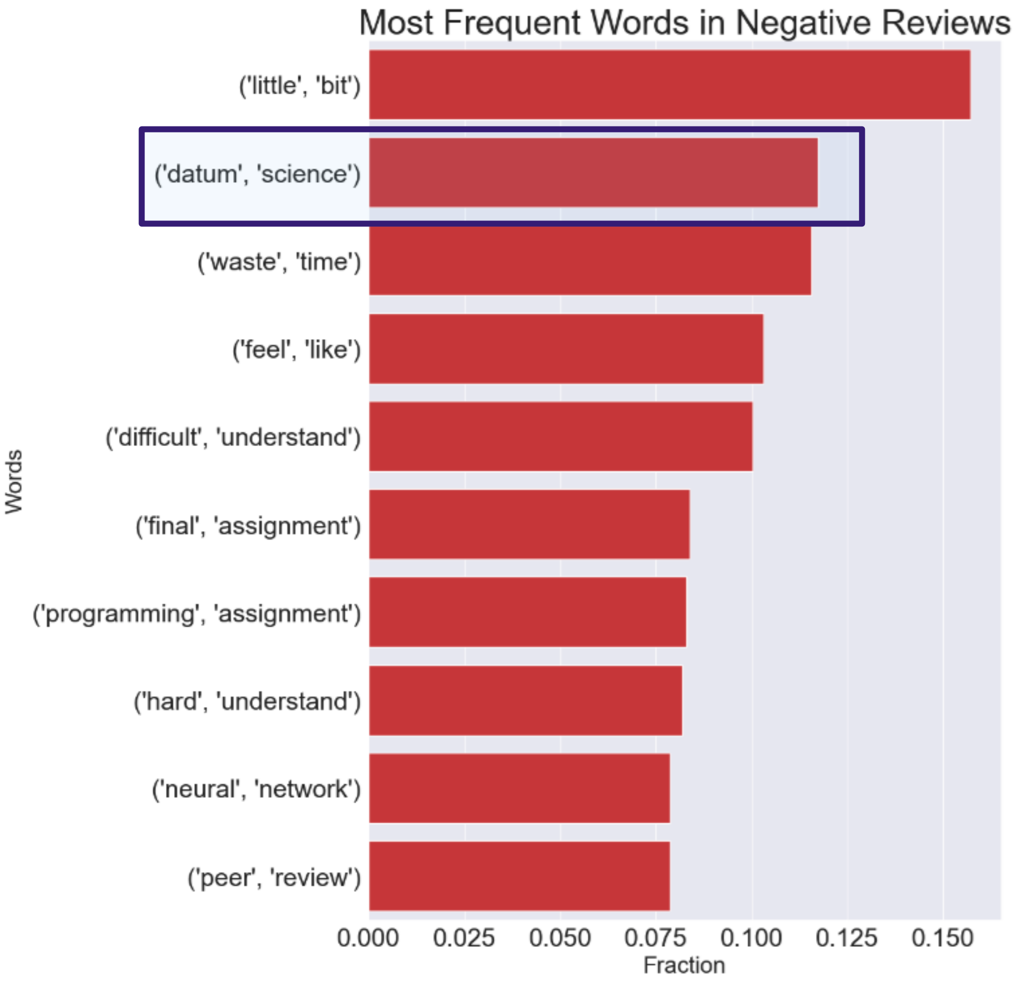


VISUALIZATION

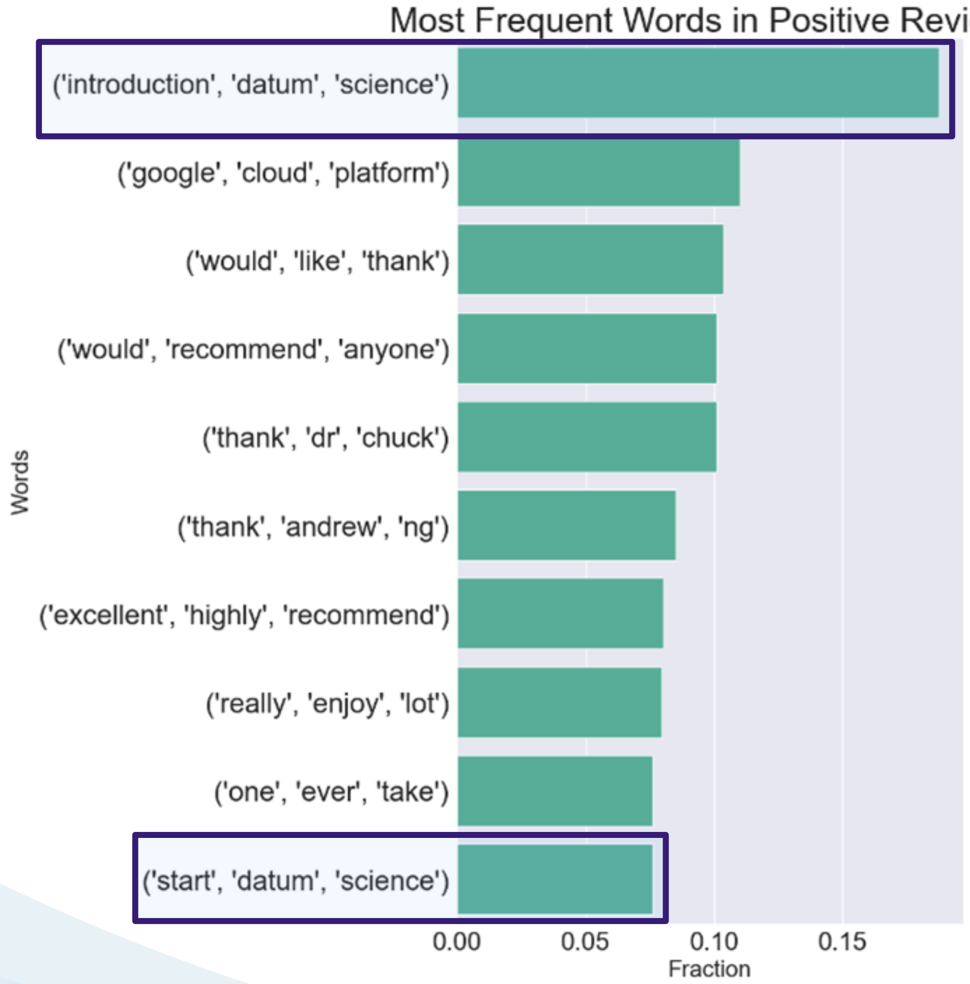


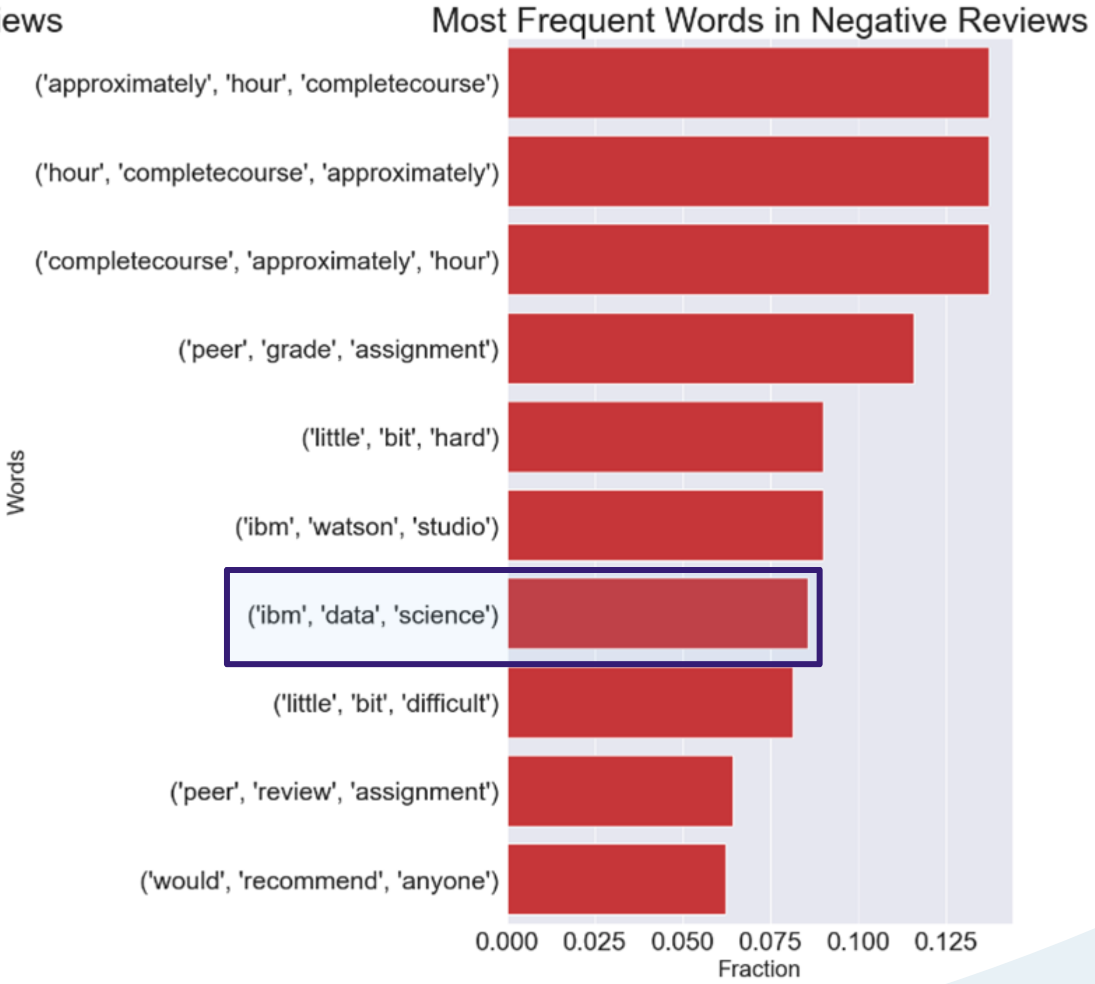
VISUALIZATION





VISUALIZATION


Most Frequent Words in Positive Reviews


Most Frequent Words in Negative Reviews

RESULTS AND RECOMMENDATIONS

Results

- The results sustain the alternative hypothesis confirming that the **course ratings are dependent on the user sentiment** mapped from the reviews.
- Ordinary Least Squares Regression results show that rating is **explained by the sentiment score (excluding neutral sentients) only by about 28.5% for categories, and 29.9% for institutions**. If we only consider university reviews, R^2 lowers to 0.281. The **p-value is always below 0.05** which indicates that the sentiment score relation with the rating is significant.
- There is a difference of **+0.08 points in the average rating between courses offered by universities** compared to courses offered by companies, and also that the number of courses for **university exceeds five times company courses**.
- The top five most offered categories are Business, Data Science, Social Science, Computer Science and Health.
- The **top five highest average ratings categories** are Personal Development, Arts and Humanities, Language learning, Physical Science and Engineering and Social Science.
- Users associate **bad experiences** with courses being too “difficult”, with much “assignments”, “time consuming”, “hard to follow” or involving “peer reviews”
- **Positive sentiments** over courses are associated with “learning a lot”, being “easy to understand”, “easy to follow”, having “good introductions”, being “enjoyable”, and suitable for “beginners”.
- 6 topics were mapped for courses that users considered: 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.
- The problematic aspects where: money, video and audio, quizzes, lectures and professors, and content & material

RESULTS AND RECOMMENDATIONS

Recommendations

1. Continue focusing the courses in **partnership with universities**.
2. There is an opportunity to offer more courses related to Personal Development, Arts and Humanities, Maths and logistics, Physical Science and Engineering, and Language Learning.
3. Improve **audio and video** quality.
4. Make **courses easy to onboard**.
5. Promote courses that are **easy to follow and to understand**.
6. Work on improving the **examination methods and assignments**.
7. **Reduce the workloads** on assignments.
8. Make courses **shorter in time**.
9. Improve **quality of content and material.**
10. Train **Instructors and professors.**
11. Disaggregate extense courses by difficulty level.
12. **Work on the pricing strategy** and **user segmentation** to identify users by needs and required value proposition.

VERSION 2

MILESTONES AND DELIVERABLES

PHASES	OBJECTIVE/ACTIVITIES	DELIVERABLES
Business Issue Understanding	▪ Define Business Objective ▪ Gather Required Information ▪ Form Hypothesis	Project Proposal with Hypothesis.
Data Understanding and preparation	▪ Identify Data Requirements ▪ Collect Initial Data ▪ Cleanse, Format, Merge the data	Master Dataset with preprocessed texts.
Exploratory Analysis	▪ Determine Important Variables ▪ Find Patterns	Report with Initial EDA showing patterns between important variables.
Hypothesis Testing	▪ Perform Hypothesis testing using Multivariate Regression to confirm if the relationship between dependent and independent variable is significant.	Hypothesis Testing Results with R-Squared, Coefficient and p-value associated with important variables.
Modeling	Use Analytic Technique: ▪ Topic Modeling (LDA) ▪ Aspect Based Sentiment Analysis(VADER) ▪ Multivariate Regression	Final Model with Result and Recommendation

RESOURCES

Project Duration: **4 months**

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. 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.	> 5 years	40	12 weeks
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 technical aspects.	3-5 years	40	12 weeks
Sr. Functional Business Analyst (lead Business Consultant)	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

RISKS AND MITIGATION ASSUMPTION

KEY AREA	KEY CONSIDERATIONS
Data	● Data updates to more recent years are not straightforward – a web scraping needs to be conducted first, then followed by a model refresh, but the process needs to be supported. ● The granularity of the company needs to be categorized and studied in a meaningful manner. ● User demand and profitability data could be collected and incorporated in the analysis.
Process	● The topic modeling might be undergoing iterations to improve performance and provide insights. ● The text language model in English has a clearly defined process, while a model update could be needed if foreign language reviews are still applicable.
People	● A personnel shortage could be experienced due to personal matter. This risk is mitigated by developing contingency plans and responsibility coverage.
Technology	● Platform availability: As a cloud-based platform, Google Colab is dependent on internet connectivity and the availability of Google's servers. If there is a disruption in service or downtime, this could impact the ability to run models and cause delays or disruptions. ● Version control: When collaborating on models in Google Colab and Google Slides, there may be challenges with version control and ensuring that all team members are using the most up-to-date version of the code and slide. This could lead to errors or inconsistencies in the model and presentation output.

PROJECT PROGRESS

Communication of project progress to Stakeholders:

WHEN:

- Weekly, with Client Project Owner
- Monthly, Senior Manager with Client Manager for ideation of new opportunities, can be attended by Lead Senior Consultants
- At the end of each milestone, or
- As decided initially in Project plan outlining the frequency, format, and content of project progress updates.

HOW:

- Regular meetings to provide updates about status of project
- Frequent reviews to ensure that approach and interim findings are aligned with desired state
- Communication methods decided according to stakeholder’s preferred method of communication (status report, email, dashboard)
- Communication content and frequency dependent on RACI charts

WHAT:

- Highlight any significant achievements or milestones reached
- Use of flow charts and presentations to convey project’s status
- Communication on delays and emerging issues, identified risks, and how they affect timelines
- Updates on any changes to the budget or resource allocation, and how they may impact the project timeline
- Recorded meetings and Minutes of the Meeting preserved for documentation and future retrieval

THANK YOU