Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data

Proposal Report for the BDM Capstone Project

Submitted by

Name: Bhavya Aditya

Roll number: 21f3001442

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1 Executive Summary

Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data

The project focuses on a Noida based MSME named NCR Eduservices Pvt. Ltd. (hereafter referred to as "NCR" or "the Company"). The company operates in the education sector and caters to both B2C and B2B segments.

The business problem being dealt with in this project is to create a B2C sales funnel for the Company. The purpose of the sales funnel is to understand the customer acquisition process of NCR. This funnel will help NCR understand which services offered by it are the key sales drivers. Further, it will help in finding the B2C conversion rate at each step and how it can be improved.

To approach this problem, the project has been divided into 4 stages. The **first stage** involves collecting the data, and removing nonessential or sensitive information.

The **second stage** involves further cleaning of data on which analysis will be done in the **third stage**.

Finally, the results of the analysis will be compiled into a presentable format in the **fourth stage**.

Various analysis methods and visualizations e.g., pareto analysis, layered proportional charts, Sankey charts, etc. will be used to create a sales funnel and find the best-selling services of NCR.

2 Organization Background

The firm being analyzed in this project is NCR Eduservices Pvt. Ltd. Started in 2012, NCR is based in Noida (Uttar Pradesh) and operates in the education sector. It deals in both B2C as well as B2B segments. In the B2C segment, the company acts as a tutor aggregator across a variety of school and college subjects. In the B2B segment, the Company acts as an outsourcing partner to domestic as well as foreign clients, primarily from the Middle East, USA, UK, Australia, and Japan. Some of the B2B services which it provides are school-teacher recruitment, college-teacher recruitment, curriculum preparation, and content creation. NCR Eduservices was founded by its CEO - Mr. Amit Gupta.

3 Problem Statement/ Objectives

3.1 **Sales Funnel:** Create a sales funnel for NCR to help understand the customer conversion rate.

3.2 **Customer Loss:** Find at which point in the sales process (from generation of leads and customer enquiries to the actual sale/ payment) the maximum customers are lost and investigate the reason for the same.

3.3 **Subject-Class Combinations:** Find which subjects and which classes attract the most customer enquiries/leads, andwhich ones convert more sales.

4 Background of the Problem

4.1 Problem 1 – Sales Funnel

The central problem of this project is gaining a deeper insight into the B2C sales of NCR Eduservices. In the B2C segment, NCR acts as a tutor aggregator – it takes customers' enquiries for tuition in specific subjects and classes, and patches them with the most appropriate tutor from its database. A trial class is offered with the tutor and if the customer is satisfied, the tutor is assigned to that customer else another tutor is assigned.

The sales process can be broken down into 3 broad steps:

1. **Customer enquiry –** The customer posts an enquiry via NCR's website/email or by calling a sales executive.
2. **Tutor matching and trial class –** The most appropriate tutor is matched with the customer for an offline trial class.
3. **Sale and payment –** If the customer decides to proceed with the tutor, it is recognized as a sale. Subsequently the customer makes the required payment.

The above 3 steps constitute the **levels/steps of the sales funnel.**

4.2 Problem 2 – Customer Loss

The number of customers reduces at every step in the above defined process due to multiple reasons – the sales team is not persuasive enough, the customer finds a better deal elsewhere, the customer does not want tuition anymore, the tutor/s provided by NCR are not satisfactory, etc.

The **sales funnel will be used** to find the customer conversion rate for each step.

4.3 Problem 3 – Subject-Class Combination

Since NCR offers/aggregates tutors for multiple subjects and classes, it is natural that the best performing subjects, classes, and their combination be found out. However, at each step of the sales funnel, there might be different best performing subject-class combinations. For instance, most of the enquiries (step-1 in the funnel) may be coming from math tuition for class 5 but most of the realized sales (step-3) may be from class 10

science tuition.

5 Problem Solving Approach

The 3-step sales funnel described above serves as the base for this project's analysis. Each step (customer enquiry, tutor matching and trial class, and sale cum payment) will be analyzed independently first, and the findings of each analysis will be collated and compared to solve the 3 problems/objectives established above and arrive at a comprehensive outcome.

The exact method and algorithm of tutor matching is proprietary to NCR and is a business secret, the details of which cannot be shared. As a result, analyzing the tutor finding method is beyond the scope of this project. However, data pertaining to how many customers proceed from enquiry to trial and then to actual sale is available and will be analyzed.

5.1 Methods To Be Used with Justification

To construct the sales funnel, which serves as the base for analysis, a **layered proportional chart** will be used to identify the relative size of each step of the funnel. A **funnel chart and Sankey/alluvial chart** will be used in addition, for more varied and interpretable visualization.

The Sankey/alluvial chart will help in seeing the flow of customers from one step of the funnel to another and give a visual idea about the successive customer loss.

After constructing the funnel, **Pareto analysis** will be used at each step to find the respective subjects, classes, and cities which have the best customer retention rates.

Finally, to summarize the insights, an **action priority matrix** may be created to indicate which subjects/classes require the maximum effort by the sales team.

5.2 Intended Data Collection with Justification

The most important data which needs to be collected is related to the **customer's tuition requirements**. It shows what subject/class customers demand the most. It is collected when the customer makes an enquiry and includes **variables like** subject, class, contact details of the customer (of which only the customer's city can be used and no personal details are available for this project due to NCR's privacy policy), proposed fees per hour, and if the customer proceeds to the next step which is trial class.

The **data mentioned above** is enough **to construct the first two steps of the sales**

funnel (customer enquiry and trial class). **To complete the third step** of the funnel i.e. realized sales, data pertaining to the sales will be collected. This data will contain the subject/class and the **final rate** at which a customer books a tutor.

5.3 Analysis Tools and Justification

Collecting customer enquiries is an unorganized process at NCR as there are multiple channels via which customers reach out. Moreover, there is no proper database with a defined schema; enquiries as well as sales data is manually entered by the sales team in an Excel sheet and requires extensive cleaning before analysis.

To that end, **MS Excel, Pandas, and OpenRefine** will be the primary tools for cleaning the data. Due to the lack of a proper format, there are some unorganized textual variables (such as subjects, city, remarks, etc.) where cleaning will require the features of OpenRefine.

The primary tools for analysis and visualization would be MS Excel, Pandas, **Plotly, and Flourish**.

To present the derived insights, MS PowerPoint or Canva will be used.

6 Expected Timeline

6.1 Gantt Chart

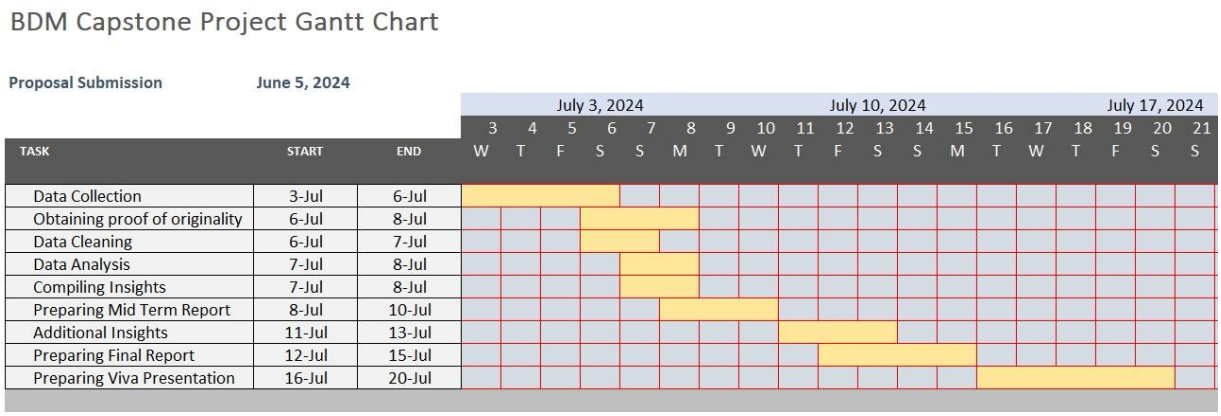


Figure 1 BDM Capstone Project Gantt Chart

6.2 Work Breakdown Structure:

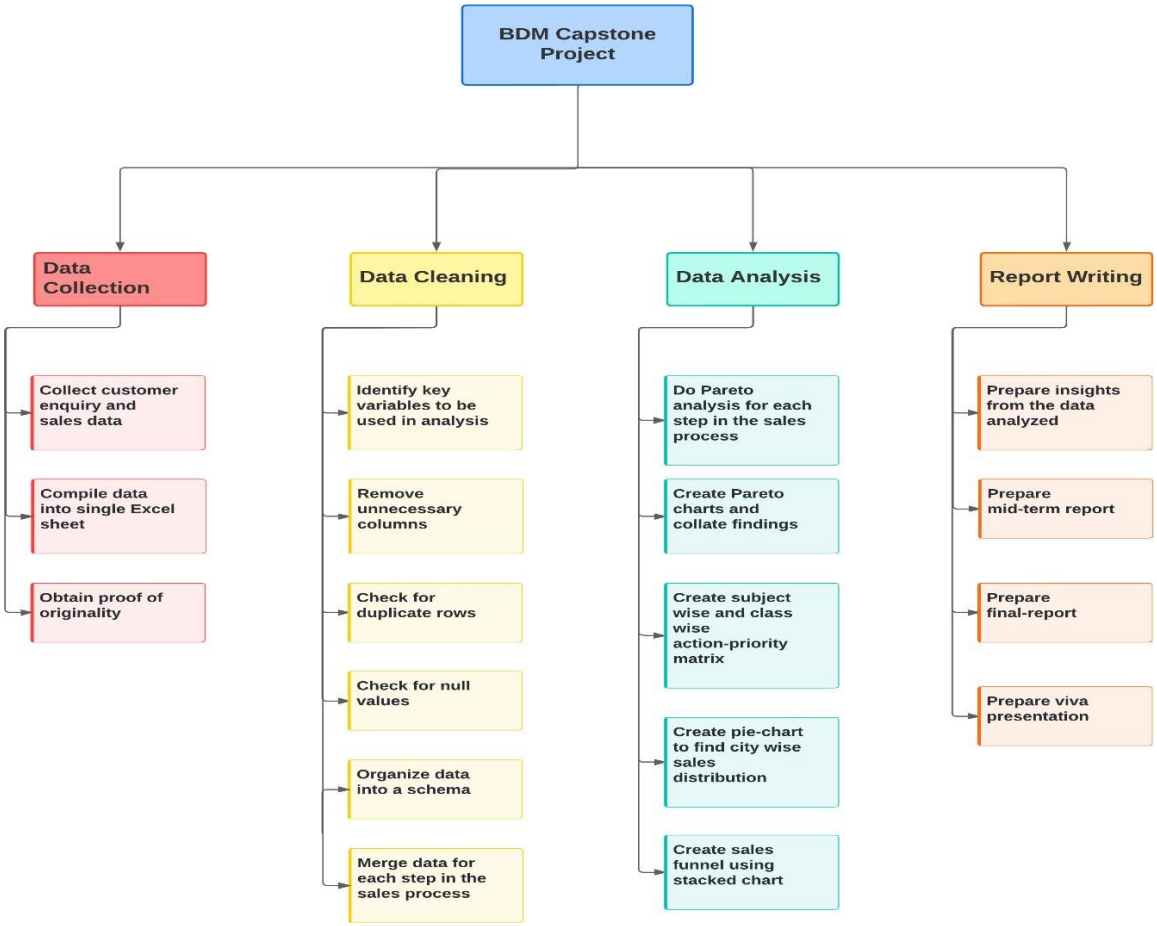


Figure 2 BDM Capstone Project Work Breakdown Structure

7 Expected Outcomes

1. Understanding at which point in the sales process NCR loses most of its customers and why.
2. Which subjects and classes have higher customer conversion rate and hence require more sales effort.
3. Find which cities provide the most sales.
4. How these findings can be used for more targeted marketing in the future.

----------End of the report----------

Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data

Mid Term Report for the BDM Capstone Project

Submitted by

Name: Bhavya Aditya

Roll number: 21f3001442

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Declaration Statement

I am working on a Project titled "**Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data**". I extend my appreciation to **NCR Eduservices Pvt. Ltd.**, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Signature of Candidate: (**Digital Signature**)

Name: Bhavya Aditya

Date: 10 November 2024

Contents

1. Executive Summary

The project aims to analyze the leads data for the tuition business of NCR Eduservices ("NCR"/ "Company"), an educational services company which caters to the B2B as well as B2C segments.

This data, primary in nature, contains information such as the location of the lead (primarily the Oceanian market), educational level, subject requirements, time commitments, mode of tutoring, status of the lead – active, inactive, or converted, etc. but in unstructured form. The data is textual rather than numeric, and has over 2000 data points.

To derive insights from the data, intensive cleaning and structuring has been done using Pandas, Geopy-Nominatim, and Llama-2 LLM. The cleaning process involved duplicate removal, feature engineering such as deriving state/territory form location, subject requirements from paragraph like string values, etc.

Using the clean data, analysis was done primarily using Pandas and Plotly. The distribution of leads based on geography and status was done with Pandas and the sales funnel was plotted using a Plotly funnel chart. The analysis gives an overview of the sales funnel of NCR and also gives insight into the geographical distribution of the leads.

The intermediate results from the cleaning process gave insight into how the sales funnel of the Company looks like. It was discovered that there is maximum customer loss at the first step which is the initial sales call with the customer, while there is minimum loss at the last step which is converting a trial class into sale, indicating a satisfactory level of service provided by the Company.

2. Proof of Originality of Data

2.1. Link to the Primary Dataset Used in the Project

https://docs.google.com/spreadsheets/d/1TeW6_NwV-jtm7LIocAUXc7hfnB-H24Fg/edit?usp=sharing&ouid=103665924288844259183&rtpof=true&sd=true

2.2. Images Related to Organization Along with Images with the Founder (Mr. Amit Gupta) in the Company Office in Sector 62 Noida

https://drive.google.com/drive/folders/1N4oaYw8zZuorsF0eNrTfM0spxtroQL_r?usp=sharing

2.3. Recorded Video with the Founder (Mr. Amit Gupta) in the Company Office in Sector 62 Noida

https://drive.google.com/file/d/1z_uc7Y0Ieouli9O2EQEpk2Z9QarQdWgj/view?usp=sharing

3. Metadata

The data is primary data which contains the customer leads from the international market and the status of these leads for the time period August 2023 to August 2024. The data is stored in an Excel spreadsheet with 2096 data points (i.e. rows excluding the header row), and 9 columns listed below:

1. **Lead Id –** An integer uniquely identifying the lead.
2. **Subject Category –** This is an artificial categorization done by the sales team of NCR Eduservices for their internal use.
3. **Name –** The name of the lead/potential client. The name found here is generally the name of a parent of the student for whom tuition classes are being sought. These names have been hidden as per the Company's instructions.
4. **Location –** Part of the address of the lead.
5. **Date Created (UTC) –** Date the lead was obtained. It follows the UTC (Coordinated Universal Time) format.
6. **Lead Qualifying Questions –** This columns contains some questions asked by the lead vendor to the potential clients about their child's educational background and requirements. This column is one of the most important columns but is unstructured, increasing the manual work of the Company's sales team. Typical questions include:
    - What is the student's education level?
    - How often would you like tutoring?
    - Which day(s) are you available for tutoring?
    - What time(s) are you available for tutoring?
    - Would you consider online or remote tutoring?
    - Which subject(s) are you looking for tutoring in?
7. **Demo Status –** This column too, is one of the most important columns as it indicates the progress of the lead along the sales funnel. Once the lead is obtained, NCR Eduservices tries to convert it into a trial/demo class and subsequently into a sale. Demo status can have one of the following values:
    - **Demo Schedule** (Lead is active) **–** The lead has scheduled a demo class.
    - **Demo Done** (Lead is active) **–** The lead has taken a demo class.
    - **Enrolled** (Lead is converted) **–** Indicates that the lead has been converted into a sale.
    - **Demo Cancelled** (Lead is inactive/dead) **–** The lead has cancelled the demo class and and is no longer interested. This incidcates a dead lead.
    - **Demo Reschedule** (Lead may or may not be active) **–** The lead has

rescheduled the demo class. It may indicate a dead lead in some cases.

(The difference between demo "cancelled" and "rescheduled" is that a cancelled lead need no longer be pursued because the customer is not interested in availing the services of the Company. In case of a rescheduled lead, the customer may either be interested but has changed the time slot of the demo class or the customer may simply reschedule or find another tutoring service and choose not to show up for the rescheduled demo.

8. **Sub status –** The sub status of a lead is linked to its demo status and provides more context about the status of the lead. The sales team of NCR has categorized some common scenarios linked to each demo status and these common scenarios are the sub statuses of the leads.
9. **Owner –** The name of the sales associate pursuing the lead. These names have been hidden as per the Company's instructions.

4. Descriptive Statistics

Some key information regarding the data has been provided below:

- Total no. of leads
  - 2096
- No. of unique leads
  - 2092
- No. of duplicate leads
  - 4 (= 2096 – 2092)
- No. of rows where the column "Lead Qualifying Questions" contains the answer to the question "Which subject(s) are you looking for tutoring in?" i.e. the lead has explicitly stated the subjects for which they are seeking tuition
  - 1637

The **geographical distribution** of the leads is given below:

Table 4-1Geographical distribution of leads

Countrv	Count
Australia	1939
New Zealand	78
Unknown	69
United States of America	6
Italv	2

Country	Count
United Kingdom	1
Algeria	1

Below is the **status based distribution:**

Table 4.2 Frequency Distribution of Values of Column "Demo Status"

Demo Cancelled	1372
Demo Reschedule	538
Enrolled	86
Demo Done	66
Demo Schedule	29

5. Detailed Explanation of Analysis Process/Methods

5.1. Data Cleaning and Feature Engineering

The data is textual, and lacks structure and consistency. Since the subjects for which a client wants tuition are not recorded separately, but provided by the lead vendor in a paragraph form which contains other information glued to it (refer column "Lead Qualifying Questions" in the data file), obtaining the subjects, current education level of the student, etc requires feature engineering. In many cases some of this crucial information is ambiguous or entirely missing as well. Moreover, there were 4 lead IDs each of which had 2 duplicate rows in the data, and there were 5 leads where the ID was missing.

A major **pain point for NCR Eduservices** is the **lack of automated data management**. Since the leads are obtained from different sources (primarily lead vendors but also social media and brand website) for different markets, the sales team manually compiles these into different spreadsheets. There is no central repository of the leads or no CRM being used.

As a result, there is very little analytics being performed. Added to this problem is the unstructured and textual nature of the data. Thus, **NCR wants to automate the data management process using AI without investing in a CRM.**

Hence, to clean the unstructured data and derive insights, Python-Jupyter environment has been used. The cleaning is done process is explained in detail below:

- **Pandas** has been used to find duplicates, missing values, frequency distributions of individual columns.
- The **Location column** has been used to obtain the state/territory of each lead's origin using **Geopy** and **Nominatim.** However, manual cleaning is still required as there are a few cases where the geocoding has not delivered the desired results.
- A custom **LLM script** has been written for NCR Eduservices, which extracts

information from the **"Lead Qualifying Questions"** column, structures it, and derives new features out of it.
    - ***Note: LLM has only been used for feature engineering and has NOT been used in any way that violates the requirements of the BDM project. This script has been written not only for this project but also for use by the NCR Eduservices' engineering team to automate the data cleaning process and reduce the manual workload of the sales team. This was a requirement set by the Company to provide primary data for this project.***
- **Handling missing values:** The column "Lead Qualifying Questions" is the column where the problem of missing values is prevalent. 455 leads have not answered the question "Which subject(s) are you looking for tutoring in?". Hence, the subject which the lead wants tutoring in needs to be inferred by checking the asnwer for the question "What is the student's education level?". This inference is done using the LLM script mentioned above.
    - In addition, there are 5 leads where the ID is missing. For the purpose of analysis, these values have been imputed with numbers ranging from 1-5.
- **Dealing with duplicates:** The data has 4 lead IDs which have been duplicated during data entry. Each of these IDs has 2 rows corresponding to it in the data and both the rows contain the same data. Hence, for each of these lead IDs, the duplicate row has been deleted.

Some of the cleaned data has been used for deriving insights in the forthcoming sections. However, there is some cleaning and feature engineering still required to find out the subject wise distribution of leads.

5.2. Calculation of the Sales Funnel

1. Data Used in the Sales Funnel

All the data used to create the sales funnel is obtained from the column **"Demo Status"**. The frequency distribution of values of this column is given in Table 4.2 and the calculation of values is explained in the next section.

2. Stages of the Sales Funnel

The sales funnel was initially planned to have 3 ordered stages – **1. Lead 2. Trial 3. Sale.** However, after a lead is generated and the sales team initiates contact, it is not necessary that the lead is converted into a trail class or the lead becomes dead.

There is an intermediate stage wherein the lead may agree to take a trail class but is yet to take it, the lead may reshcedule an already scheduled trial class and even then may or may not take the trial class. If a lead reschedules a trial class but does not show up for it, the lead may then be considered dead but until then, its status is ambiguous.

Thus **Trial** stage has been broken down into 2 stages –

1. **Trial (Planned) -** At this stage, the lead has agreed to/scheduled a trial class, or rescheduled a trial class. This stage also includes leads which have taken the trial class, and thus is a superset of the **Trial (Taken)** stage to ensure logical consistency.
2. **Trial (Taken) -** At this stage the potential customer has taken the trial class and may or may not enroll into NCR's classes.

Finally, all the stages of the sales funnel along with their formulae (also refer table 5.1) are given below:

Table 5.2-1. Stages of the sales funnel and their calculation

1. Lead	No. of rows in the data – Duplicate leads	**2092**	(= 2096 – 4)
2. Trial (Planned)	Demo Scheduled + Demo Done + Demo Rescheduled + Enrolled	**719**	(= 29 + 66 + 86 + 538)
3. Trial (Taken)	Demo Done + Enrolled	**152**	(= 66 + 86)
4. Sale	Enrolled	**66**	

6. Results and Findings

6.1. Insights From the Sales Funnel

- The no. of leads even planning to take a trial is just 34% of the total leads (refer figure 6.1). This indicates that the initial sales call may not be effective enough to convince leads to schedule a trial. **This stage is where the maximum customer loss is occurring.**
- From the leads who schedule a trial class (719), only 21% (152) of them actually take a trial class. There are deterrents which are responsible for the low conversion to this stage and need to be investigated.
- These deterrents can include unfavourable time slot, customer not being convinced enough to actully take the trial, better options by competing tuition firms, etc.
- From the leads who actually take a trial class (152), 43% (66) are being converted into sale i.e. they are enrolling in NCR Eduservices' classes. The conversion rate form the previous stage here is better than that in the other stages. This indicates that customers who take a trial class, must be generally satisfied with the Company's services and hence they avail them.
- The sales conversion rate at 3% of the total leads is relatively low. The column "Sub Status" needs to be analysed to further understand reasons behind and suggest corrective actions.

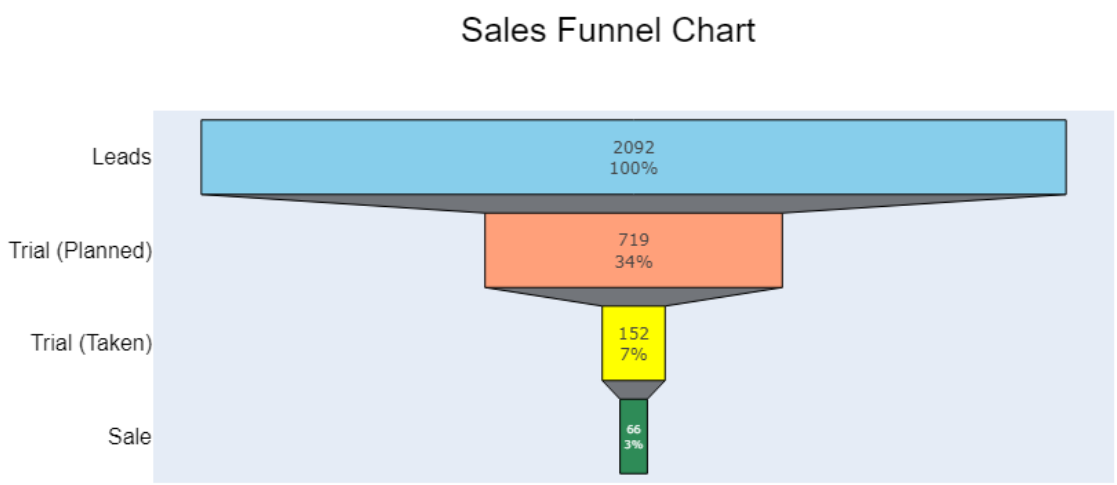


Figure 6.1. Sales Funnel

6.2. Insights from Country and Demostatus

Table 6.2.1 Country Wise Lead Status

	Lead Status ("Demostatus" column)							
	Demo Cancelled	**Demo Done**	**Demo Reschedule**	**Demo Schedule**	**Enrolled**	**Total**	**Demos Cancelled Per Enrollment**	**% Enrollments Per Trial (Taken)**
Algeria	1	0	0	0	0	**1**	0	-
Australia	1310	59	457	26	78	**1930**	17	57%
Italy	0	0	1	1	0	**2**	0	-
New Zealand	13	1	61	2	1	**78**	13	50%
United Kingdom	0	0	1	0	0	**1**	0	-
United States of America	4	1	1	0	0	**6**	0	0%
Unkown	42	4	17	0	6	**69**	7	60%
Total	**1370**	**65**	**538**	**29**	**85**			

- Majority of leads (92%) come from Australia, followed by New Zealand (3.7%). The Oceanian region is the origin for almost 96% of the leads. The country of origin cannot be determined for 69 leads and these come under the category "Unkown". Hence, Australia and New Zealand demonstrate a higher interest in NCR's services as compared to other countries, considering that NCR's marketing efforts are equal for all foreign countries.
- Between Australia and New Zealand, the no. of demos cancelled per enrollment is higher for Australia at 17, while for New Zealand it is 13 (refer table 6.2.1).
- However, the percentage of leads which convert into a sale after taking a demo (indicated by % Enrollments Per Trial (Taken)) is higher in Australia (57%) than in New Zealand (50%).
- This indicates that it may difficult to convince a lead to take a demo in Australia than

in New Zealand, but once a lead has taken a demo class, the sales conversion is likely to be better in Australia.

6.3. Planned Analysis for Final Report

- The current analysis provides insights into the how the overall sales funnel for NCR looks like and which countries show higher interest for NCR's services.
- Further analysis needs to be done to find the subjects which generate the most leads and sales.
- For Australia and New Zealand, further analysis is required to understand which states have better conversion.
- Finally, the analysis needs to be done to find which subjects have higher demand in which states of Australia and New Zealand.

----------End of the report----------

Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data

Final Report for the BDM Capstone Project

Submitted by

Name: Bhavya Aditya

Roll number: 21f3001442

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Declaration Statement

I am working on a Project titled "**Creating a Sales Funnel for an Education MSME by Analyzing Customer Queries and Sales Data**". I extend my appreciation to **NCR Eduservices Pvt. Ltd.**, for providing the necessary resources that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered from primary sources and carefully analyzed to assure its reliability.

Additionally, I affirm that all procedures employed for the purpose of data collection and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to the principles of academic honesty and integrity, and I am receptive to any additional examination or validation of the data contained in this project report.

I understand that the execution of this project is intended for individual completion and is not to be undertaken collectively. I thus affirm that I am not engaged in any form of collaboration with other individuals, and that all the work undertaken has been solely conducted by me. In the event that plagiarism is detected in the report at any stage of the project's completion, I am fully aware and prepared to accept disciplinary measures imposed by the relevant authority.

I understand that all recommendations made in this project report are within the context of the academic project taken up towards course fulfillment in the BS Degree Program offered by IIT Madras. The institution does not endorse any of the claims or comments.

Signature of Candidate: (**Digital Signature**)

Name: Bhavya Aditya

Date: 12 March 2025

Contents

Executive Summary

This report analyzes data of leads of NCR Eduservices, an education firm providing online tutoring. The assignment entailed thorough data cleaning, feature engineering, and analysis of 2092 leads in order to know subject demand, education level wise demand, desired days, and convertion of leads.

The most important findings are that maths and english are the most demanded subjects both in general and among converted leads, with science and exam preparation also having high demand. Primary and secondary school levels are the segments which constitute the majority of demand (above 83%). Weekday classes are more in demnd over the weekends, and the days most preferred by leads are Monday, Tuesday, and Wednesday. One of the key challenges that have been identified is the low overall lead conversion rate (around 4%). Lead status analysis shows that the primary reasons for low lead conversion rate are leads being not interested, wanting offline classes, or having already enrolled elsewhere.

The recommendations of the report concentrate on a number of key areas. NCR Eduservices must give importance to maths and english in curriculum planning and promotion. Bundling options, especially for maths and science, have been suggested based on demand analysis done. The high demand for offline classes indicates a large market opportunity which NCR can look into. In addition, lead sourcing and targeting need to be improved to deal with the high no. of “not interested" leads. The data capture also needs to be enhanced. Lastly, the report recommends exploring and possibly specializing in exam preparation market (LANTITE, NAPLAN, and ATAR in particular), as this is a significant niche market. Optimizing scheduling to fit weekday demand and taking advantage of the flexibility of "any day" preferences have also been also recommended.

Detailed Explanation of Analysis Process/Method

The data is textual, and lacks structure and consistency. Hence, data cleaning and feature engineering has been performed to make the data usable for analysis. Multiple new columns have been created from the existing columns, spelling errors and variations have been corrected, similar values grouped, etc.

The raw data underwent intensive data wrangling to make it consumable for analytics. Python and its package Pandas were the primary tool used for this. In addition, MS Excel and OpenRefine were used as well. Charts have been created using the Plotly and Seaborn libraries of Python, and MS Excel.

Data Cleaning

Cleaning the data primarily involved the following 3 tasks:

- removing duplicates,
- handling spelling variations and errors within columns,

- identifying unique values and grouping similar values

There were 4 duplicate rows in the data, which were removed.

A major problem, which persisted even after feature extraction/engineering, was handling spelling errors in the columns. For instance, in many columns where there was a value "unknown", there were multiple occurrences of the erroneous spelling "unkown". Similarly, "maths" had variations/errors such as "math" and "mathss".

Such spelling variations and errors increased the redundancy within the data. In many columns, the no. of unique values was more than the no. of logically unique items. These were handled manually in Pandas.

In some of the newly created columns, many similar values were grouped for the purpose of analysis. For instance, in the "subjects" column – values "physics", "chemistry", and "biology" were grouped under the name "science". There were many subjects which occurred very few times (or even just once) in the data. Many such subjects were grouped under the category "others".

There were 34 unique subjects identified in the data but since most of them occurred just once, the no. of unique subjects was grouped reduced to 6 subjects.

When analyzing the education level or the grades of the leads, multiple grades were clubbed into categories like "high school" or "primary school". The details of this grouping and the rationale behind it are discussed later on, in the analysis section.

Feature Extraction and Engineering

The column "Lead Qualifying Questions" contains information about which subjects are demanded by a lead, the grade/education level for which they want tuition, and their preferred days of the week when they want classes. This information is crucial for the analysis done in this report.

However, all this information is in form of subjective questions answered by a lead. There are no standard answers or choices to choose from, making the information in this column highly varied and unstructured. Hence, the majority of the data wrangling and feature engineering/extraction done is on this column alone.

A custom LLM script has been written for NCR Eduservices, which extracts information from the column "Lead Qualifying Questions", structures it, and derives new features out of it. The LLM used is Google's Gemini Flash
Note: LLM has only been used for feature engineering and has NOT been used in any way that violates the requirements of the BDM project. This script has been written not only for this project but also for use by the NCR Eduservices' engineering team to automate the data cleaning process and reduce the manual workload of the sales team. This was a requirement set

by the Company to provide primary data for this project.

Another problem in the column "Lead Qualifying Questions" is the column where the problem of missing values is prevalent. 455 leads have not answered the question "Which subject(s) are you looking for tutoring in?". Hence, the subject which the lead wants tutoring in needs to be inferred by checking the answer for the question "What is the student's education level?". This inference is done using the LLM script mentioned above.

In addition, there are 5 leads where the ID is missing. Since lead IDs have no inherent meaning and are only used for uniquely identifying rows, all the lead IDs have been changed to numbers ranging from 0 to 2091.

Below are the 3 most important columns were extracted using the above mention LLM script:

1. **Subjects**: This column contains the occurences of the unique subjects demanded by a lead.
2. **Student Education Level**: This column contains the grade information of a lead.
3. **Available Days**: This column contains the preferred days when the lead wants classes.

The above 3 columns form the basis for the analysis done in this report. Hence the preprocessing and analysis done on each column is discussed in detail in the subsequent section.

Analysis Process and Methodology

Subjects Column

This column was created by extracting the subjects demanded by a lead using Gemini. For each row, a list of subjects was obtained since a lead may demand tuition for multiple subjects. The extracted subjects were cleaned and grouped into similar subjects. Initially, there were 34 unique subjects obtained which were grouped into 6 unique subject categories on which the analysis was done.

Before moving further, it is important to understand the following two terms:

i. **Unique subject**: A unique subject is a specific requirement/subject for which the lead is looking for tuition. For e.g. maths and english are two unique subjects as they are different from each other. However, a unique subject may not necessarily be an academic "subject". For instance, "exam preparation" can be a unique subject as it represents a unique but frequently occuring requirement in the dataset. Identifying unique subjects is important because:
   a. A single lead may demand tuition for more than one subject
   b. Multiple leads may demand the same subject but using refer to it differently

ii. **Subject occurrence:** As mentioned above, a subject might occur in the dataset

multiple times but with spelling variations since there is no standardized list of subjects and leads mention their subject requirements in their own words. Hence, to really understand the demand for subjects, it crucial to first identify how many subjects are present in the dataset and second, how many times is that subject demanded. Each time a lead demands a unique subject, that subjects "occurrence" increases by one.

The 6 subject categories which have been used for analysis are maths, english, science, exam preparation, others, and unkown.

Once these 6 categories were identified, the subjects column was one hot encoded to capture the occurrence of each subject across all the leads.

One hot encoding was the preferred technique here as a lead may demand multiple subjects and hence the subjects column had values in form of lists of subjects. This gave 6 new columns, each corresponding to a unique subject and the value of a column would be 1 if the lead demanded that subject else it would be 0.

Student Education Level Column

This column was created by extracting the grade/education level from the column "Lead Qualifying Questions".

This was one of the most unstructured columns in the data since a lead answers their grade subjectively, not by choosing from a pre-defined set of options. As a result in many cases it was difficult to infer the exact grade of the lead. Examples of such values are: "['Year 4', 'Year 5']", "Year 11 or 12 (age 16-18)", "Grade 5-6". As evident here, the lead can be attributed to multiple grades.

To tackle this situation, the following 5 broad classifications were made:

1. Primary school (Grades 1-5)
2. Secondary school (Grades 6-9)
3. High school (Grades 10-12)
4. Pre-primary school (Kindergarten or below)
5. Unkown

These classifications are in line with NCR's requirement as well because, NCR does not typically consider different grades as different segments, rather it follows the classification given above. This is because all grades within the primary school category (grades 1-5) have similar level of curricula and hence each grade does not require different sets of teachers.
A teacher who is qualified to teach grade 1 can also handle grade 5, but might not be as well-suited for pre-primary or high school grades due to the differing levels of students and curricula. Pre-primary education often demands teachers who possess a friendly demeanor and the ability to teach the fundamental concepts of all subjects to young children with shorter attention spans. In contrast, high school education requires teachers who are

specialized in specific subjects and capable of instructing a more mature age group.

After making the above classification, each lead was assigned a single education level thus making analysis easier. This column was used to create pie charts to understand the distribution of education level in leads. It was also used with the column “Demostatus” to understand the distribution of education level in converted leads.

Available Days Column

The available days column contains the list of preference of days of tuition classes for a lead. Here also, spelling errors and variations were corrected. The cleaned column was then one hot encoded to obtain 7 new columns for each day of the week.

There were mutiple cases were leads had mentioned “any” or “anyday” as there preferred days. In this case, all the days were considered as preferred and the value for all the days’ columns were set to 1 for that lead.

Like the subjects column, here too, a lead can give multiple preferences. Hence, the concept of occurrences is used in the analysis of this column as well. Using the one hot encoded columns, the distribution of the occurrences of the days of the week was found and then correlation analysis was done in an attempt to find possible combinations in day preferences.

Demostatus and Substatus Columns

The “Demostatus” column contains the current status of a lead and the column “Substatus” contains the sub-status linked to that status (a sub-status may be linked to multiple statuses). The sub-status is more specific and gives a better idea about the position of a lead. It is an important column as it sheds light on the reasons nehin the conversion or non-conversion of leads.

The demostatus column is standardized with the following possible values:

- **Demo Schedule (Lead is active) –** The lead has scheduled a demo class.
- **Demo Done (Lead is active) –** The lead has taken a demo class.
- **Enrolled (Lead is converted) –** Indicates that the lead has been converted into a sale.
- **Demo Cancelled (Lead is inactive/dead) –** The lead has cancelled the demo class and and is no longer interested. This incidcates a dead lead.
- **Demo Reschedule (Lead may or may not be active) –** The lead has rescheduled the demo class. It may indicate a dead lead in some cases.

The substatus column however, is not as standardized and there were multiple spelling errors/variations which were corrected. This resulted in 25 unique sub-statuses. The demostatus and substatus columns were analyzed using a heatmap visualization.

Results and Findings

Analysis of the Status and Sub-status of Leads

The heatmap below (Fig. 1) shows the status as well as the sub-status of each lead, providing a better understanding of the reasons behind the loss or conversion of a lead.

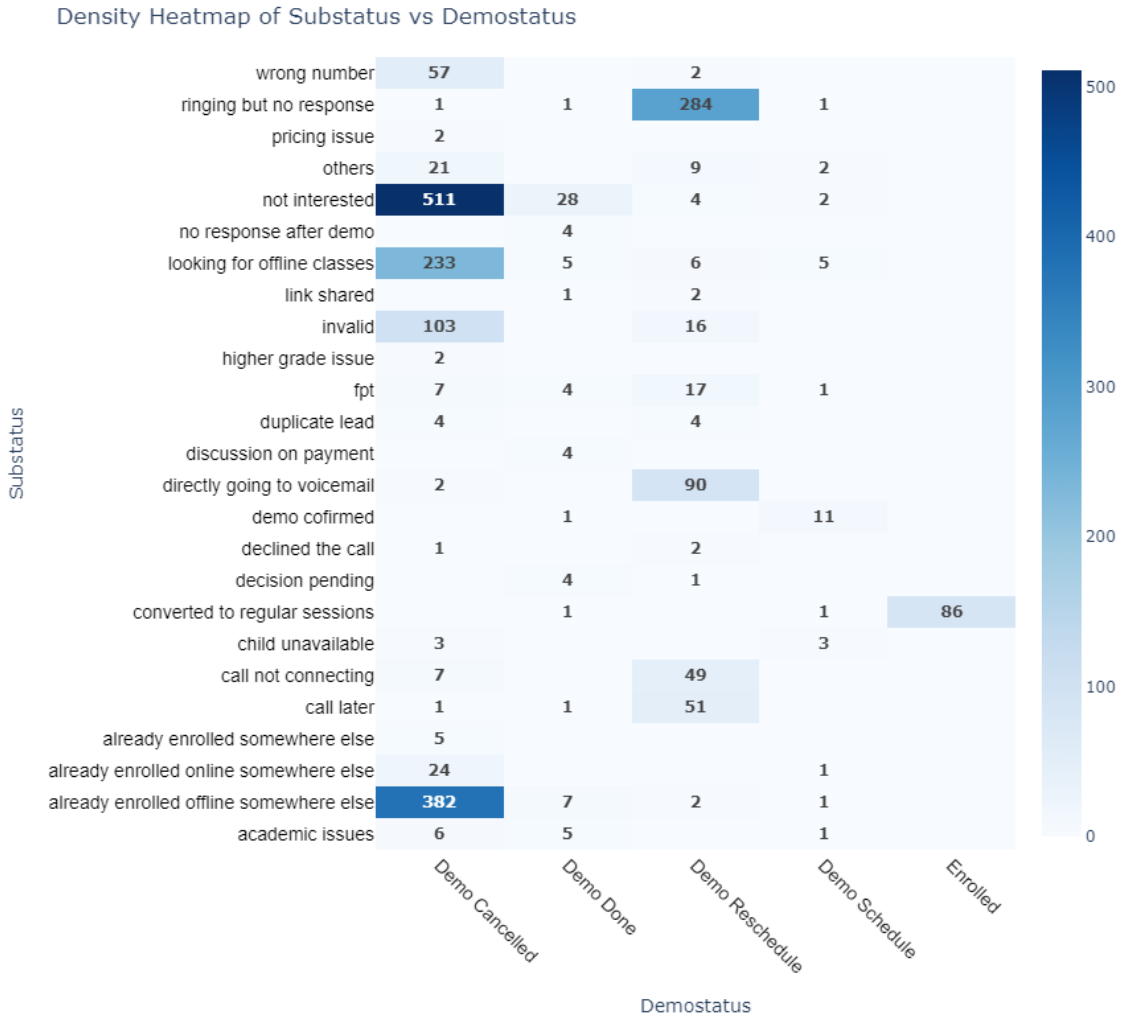


Figure 1 Heatmap of Status and Sub-status of Leads

- **Primary Reasons for Cancellation:** Looking at the heatmap, it is evident that the three most frequent reasons for demo cancellations are:
    - the lead is no longer interested, (“not interested” - 511 leads)
    - the lead is “looking for offline classes” not online (233 leads)
    - the lead has “already enrolled elsewhere” (382+24 i.e. 406 leads)
- In cases where the lead enrolled elsewhere, only few of the enrollments are into online classes (24 leads i.e. 1.7%) by other providers while majority are in offline

classes (382 leads i.e. ~28% of the cancelled demos).

- **Rescheduled Demos:** In leads having the status "Demo Reschedule", majority of them have the sub-statuses related to not being able to contact the lead, such as “ringing but no response” (284 leads i.e. ~53% of demo reschedules) or “directly going to voicemail” (90 leads – 17%) or “call not connecting” (49 leads – 9%) , indicating that reschedules rarely convert to completed demos.
- **Pricing:** Only 2 leads out of 2092 leads have cited “pricing issue”. This is just 0.1% of the leads, approximately. Hence, pricing issue is a trivial reason/sub-status.

Demand of Subjects

An important point to reiterate here is that there are 2092 leads in the dataset, but that does not imply that there will be 2092 occurrences. One subject may be demanded more than once by different leads and one lead may demand multiple subjects at once.

There are a total of 4121 subject occurrences. Given below is a bar chart along with pareto line visualizing the distribution of subjects demanded by the leads.

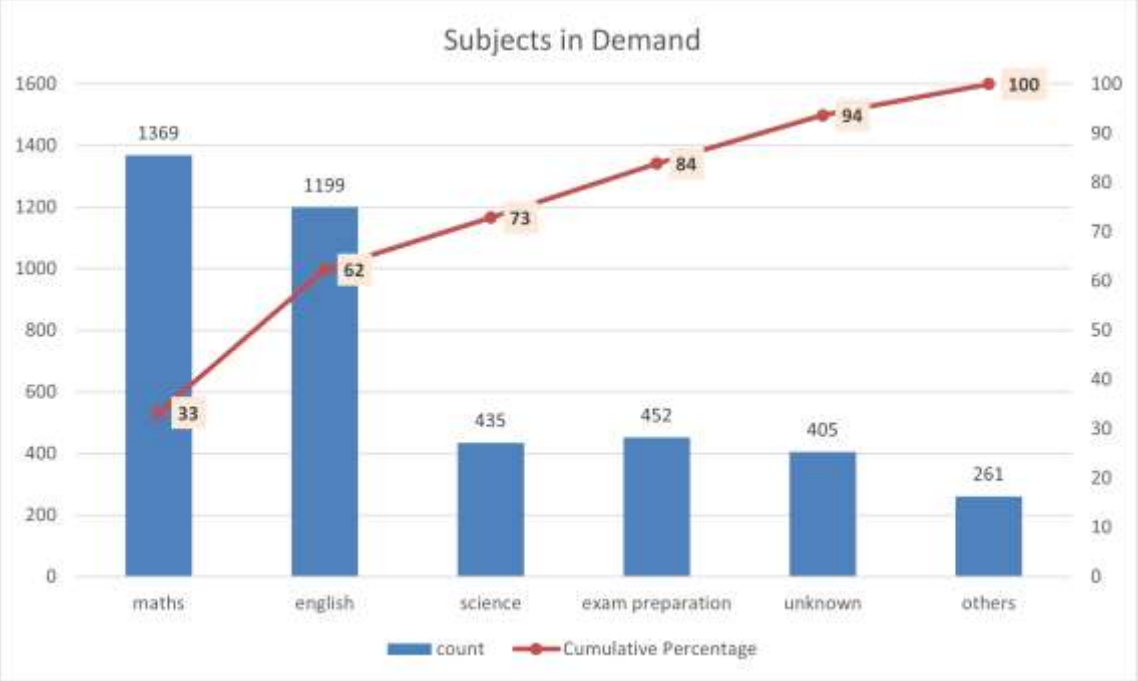


Figure 2 Distribution of Subjects Demanded and Pareto Analysis

- As evident from Fig. 2, maths and english are the most demanded subjects accounting for 33% and 29% of the occurrences respectively.
- These are followed by science and “exam preparation”, which account for approximately 11% of the total occurrences each.
- Maths, having 1369 occurrences, is demanded by 65% of the leads. English, having 1199 occurrences is demanded by 57% of the leads.
- Science and exam preparation have 435 and 452 occurrences respectively and are

demanded by 21% and 22% of the leads respectively.

- Only 2 subjects, maths and english account for 62% of the occurrences. Maths, english, and science together account for 73% of the demand and if exam preparation is taken into account as well, then just 4 subjects constitute 84% percent of the demand.
- There are 405 leads for which the subject requirements could not be determined. If these leads follow the same distribution, then the 4 subjects may account for an even higher proportion of the total demand.

The table (Table 1) and the figure (Fig. 3) given below show the distribution of leads by the no. of subjects demanded. They gives an insight into the composition of the demand.

Table 1 Distribution of Leads by No. of Subjects Demanded

No. of Subjects	No. of Leads	Percentage of Toal Leads
1	964	46.1 %
2	645	30.8 %
3	319	15.2 %
4	98	4.7 %
5	66	3.2 %

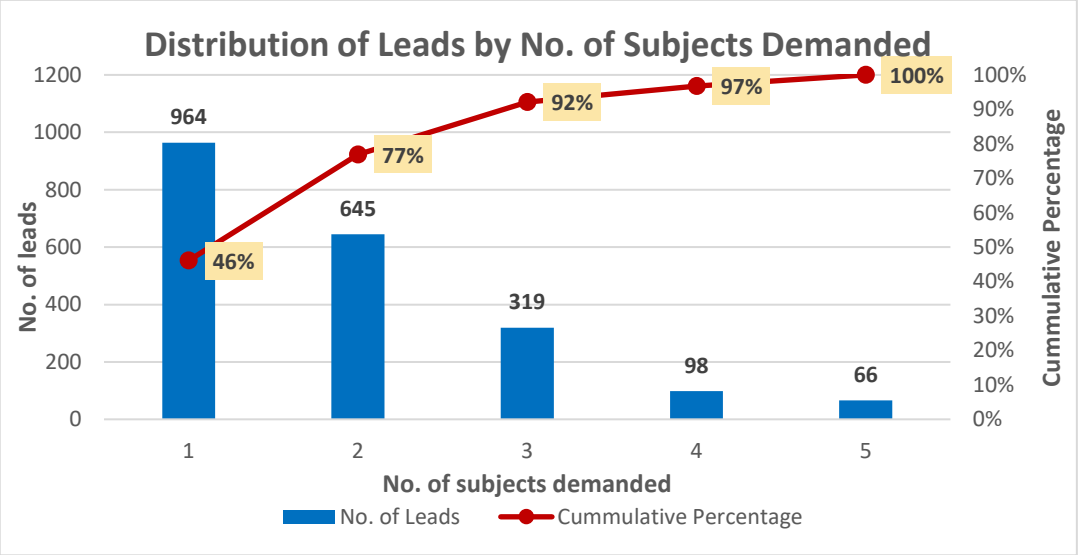


Figure 3 Distribution of Leads by No. of Subjects

- From Table 1, it can be seen that the demand for classes for a single subject are the highest, at 964 leads or 46% of the total leads.
- This is followed by demand for 2 subjects at 31% (645 leads) and 3 subjects at 15% (319 leads).
- Together, the 1-3 subjects segment generates 92% of the demand.

An important result which needs to be highlighted separately is that 80% of the science

leads demand math as a subject as well, however, only 25% of the math leads demand science as well. Similarly 80% of the english leads demand math as a subject as well and 70% of the math lead demand english as well.

Subject-wise Sales/Enrollemnts

The table (Table 2) and the chart (Fig. 3) show the occurrences of subjects in the leads with the status "Enrolled" i.e. leads which have been converted into sales. In total, there were 86 enrollments/converted leads in the data.

Table 2 Distribution of Subjects in Enrolled Subjects i.e. Sales

Subject	Occurence	Percentage
Math	59	69 %
English	48	56 %
Science	23	27 %
Exam preparation	17	20 %
Others	7	8 %

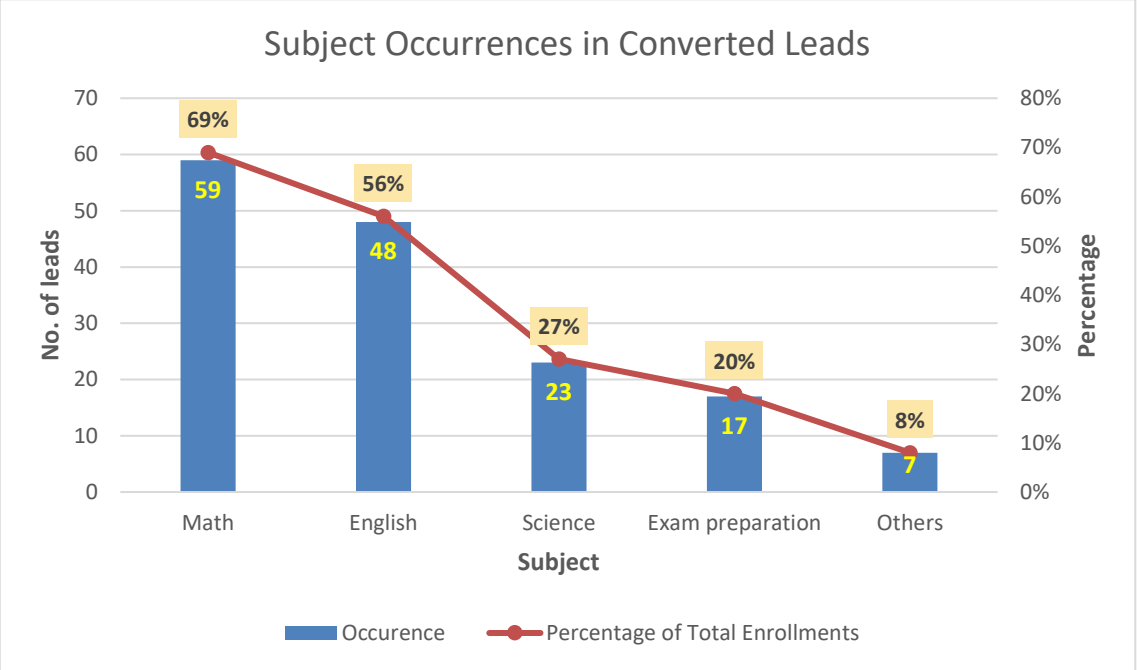


Figure 4 Subject Occurrences in Converted Leads

- Maths and english are the leading subjects in the converted leads as well, with maths being a subject in 59 (69%) enrollments and english being there in 48 (56%) enrollments.
- In enrollments too, maths/english are followed science (27% enrollments) and exam

preparation (20% enrollments).

- However, while the leads were more for exam preparation (452 leads) than science (435 leads), the conversion is better for science (23 enrollments vs 17 enrollments for exam preparation).
- Only 8% of the enrollments have subjects belonging to the "others" category. This confirms the dominant position of maths, english, science, and exam preparation as the leading subjects in not only demand but also conversion.

The chart below (Figure 5) shows the distribution of enrollments by subject combinations.

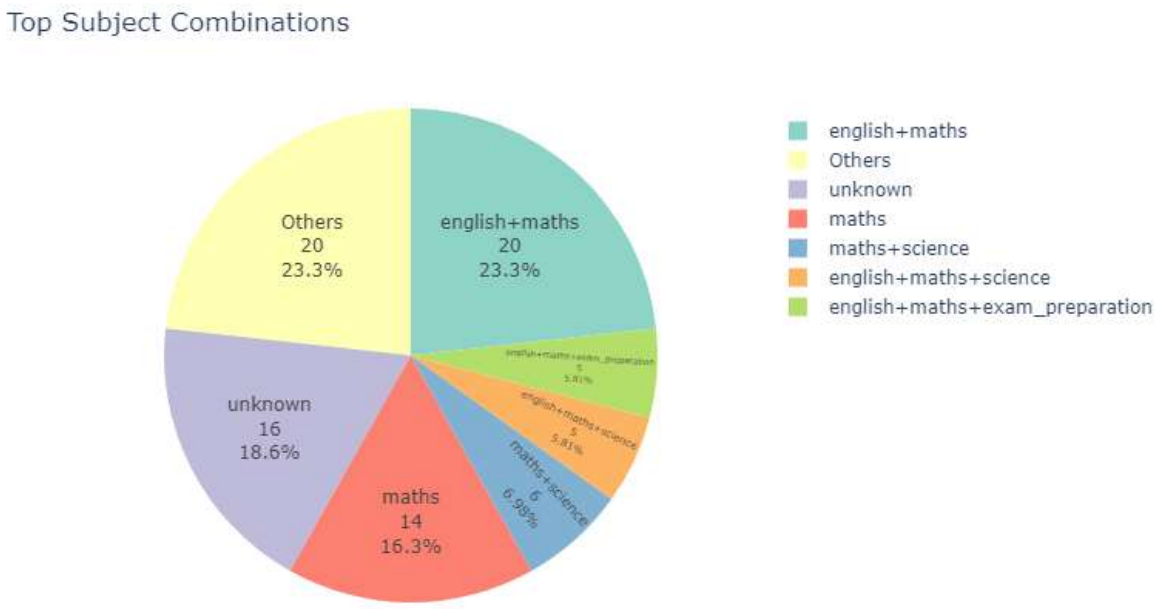


Figure 5 Subject Combinations in Enrollments

- In single-subject enrollments, math has the highest share (16.3%) with 14 enrollments.
- The dominant subject combination is maths plus english with 23% share (20 enrollments). This combination also has the highest share overall.

The chart below (Fig 6) shows the distribution of leads by no. of enrollments.

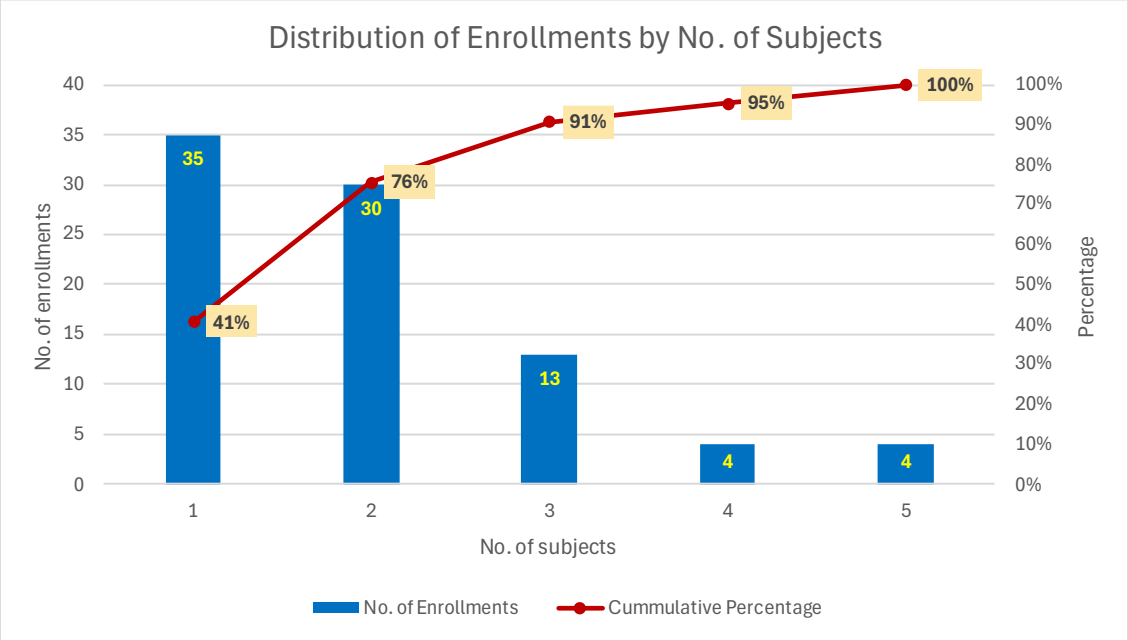


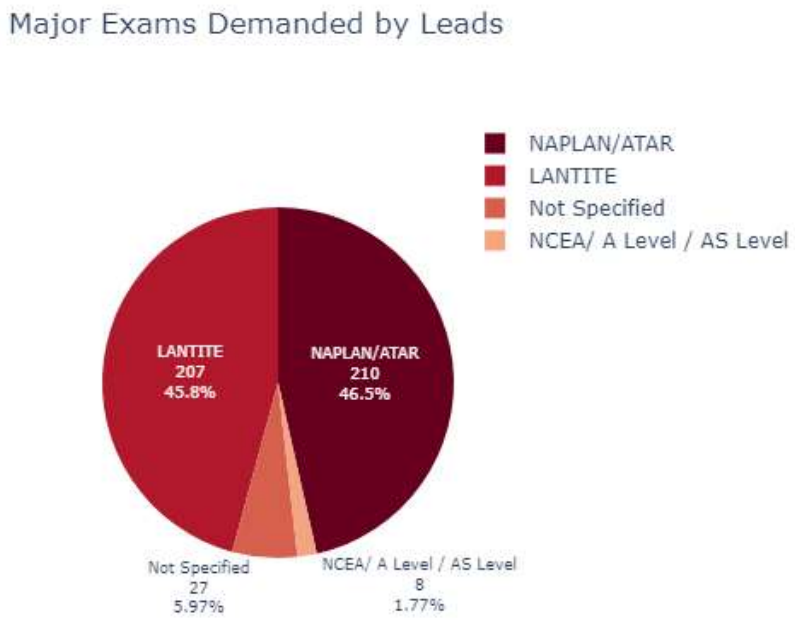
Figure 6 Distribution of Enrollments by No. of Subjects

- The distribution in enrollments follows a similar distribution as the one seen in Fig. 3 which showed distribution of all leads by no. of subjects.
- Here too, 1 subject is the leading category (41%) followed by 2 subjects (35%).
- The 1 subject and 2 subjects categories constitute to roughly three quarters of enrollments (76% here as compared to 77% in Fig. 3), and combined with the 3 subjects categories, they constitute 91% of the enrollments (92% in Fig 3).

In enrollments too, there is a science-math relationship which was observed when analyzing demand: 83% of students who enroll in science also enroll in math. Similarly, 81% of english enrollments also enroll in math. However, the converse does not hold as math enrollments are less likely to take enrollment in other subjects (66% enroll in english, 32% enroll in science)

Demand of Tuition for Exams

There are 452 occurences of exams in the dataset, i.e. 22% of the leads are looking for tuition for some specific exam rather than the regular school curriculum. Exams have been separately analyzed as they can be a potential market for NCR Eduservices to specialize in.



- The most demanded exams are:
    - **LANTITE** (Literacy and Numeracy Test for Initial Teacher Education) is a mandatory exam for prospective school teachers enrolled in the Intital Teachers Education (ITE) program in Australia.
    - **NAPLAN** (National Assessment Program – Literacy and Numeracy) is a standardized school level exam in Australia for students in grades 3, 5, 7, and 9.
    - **ATAR** (Australian Tertiary Admission Rank) is a ranking system used for university admissions in Australia. It is calculated based on a student's performance in their final year of high school i.e. grade 12, and it determines eligibility for university courses.
- While LANTITE is the single most demanded exam (~46% of all the exams), NAPLAN and ATAR together constitute for about 47% of the exam demand.

Analyzing the Educational Level Demanded

The chart below (Fig 7) shows the distribution of leads by education level.

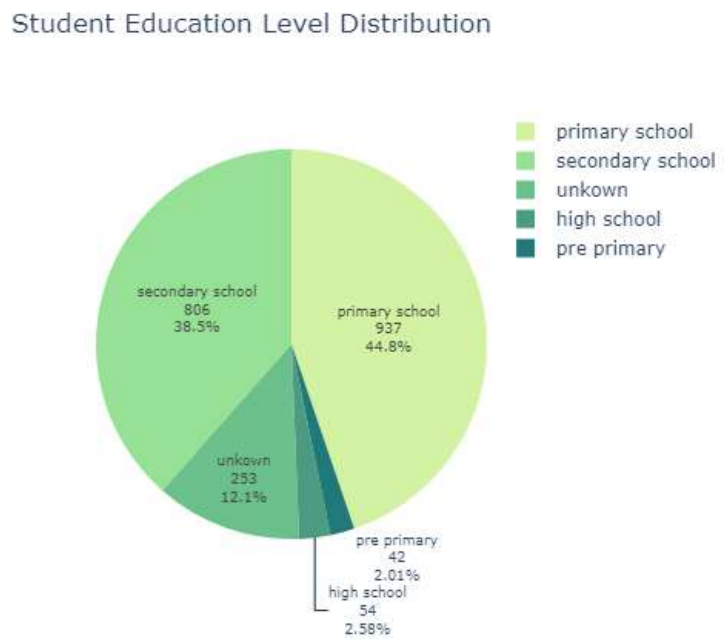


Figure 7 Demand Distribution by Education Level

- Primary school is the leading segment with a share of 45% i.e. 937 leads.
- Secondary school is te second most demand generating segment with 39% share i.e. 806 leads.
- Together, primary and sceondary school constitute over 84% of the demand.
- There are 253 leads for which the education level could not be determined and they have been categorized as “unkown”. Due to this, there might be a possibility that primary and secondary school segments together may have a higher share.

The chart below (Fig 8) shows the education level wise distribution for converted leads i.e. leads with status “Enrolled”.

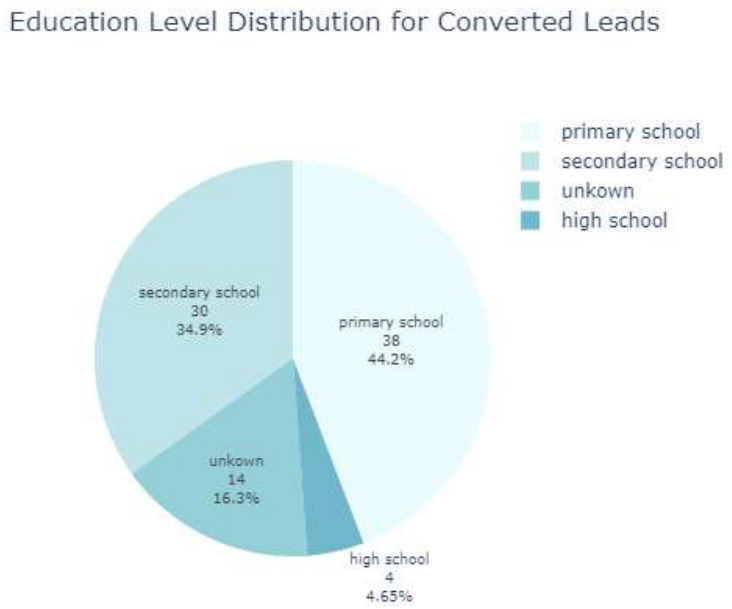


Figure 8 Distribution of Converted Leads by Student Education Level

- The converted leads follow a similar distribution as the one seen in Fig 7.
- Here too, primary school is the leading category (44% share) followed by secondary school (35% share).
- Two noticeable differences here are:
  - Pre-primary has 0% share in enrollments
  - High school, on the other hand, has a higher share in this distribution (4.65%) than in Fig 7 (2.6%)
- Since there are enrollments with unkown education level, pre-primary may have some enrollments which are not listed here. This is one of the improvements required in the data management process at NCR as the data of enrolled leads has not been updated in the sales data
- Ignoring the pre-primary and unkown categories, it can be seen that though high school category has the lowest no. of enrollments, it has the highest conversion rate; significantly higher than the primary and secondary school categories.
- The conversion rates are given below:
  - Primary School: **4.06%** (= 38 / 937 * 100%)
  - Secondary School: **3.72%** (= 30 / 806 * 100%)
  - High School: **7.41%** (= 4 / 54 * 100%)

Day Wise Demand of Leads

- From the chart below which shows the distribution of days, it is evident that there is a greater demand for weekdays than weekends.
- Monday (1375 occurrences), Tuesday (1302), and Wednesday (1333) have the highest demand, while Saturday (1066) and Sunday (968) have the lowest demand.

- 788 leads i.e. 38% of the total leads have expressed availability for all any day of the week and prefer any of the 7 days.

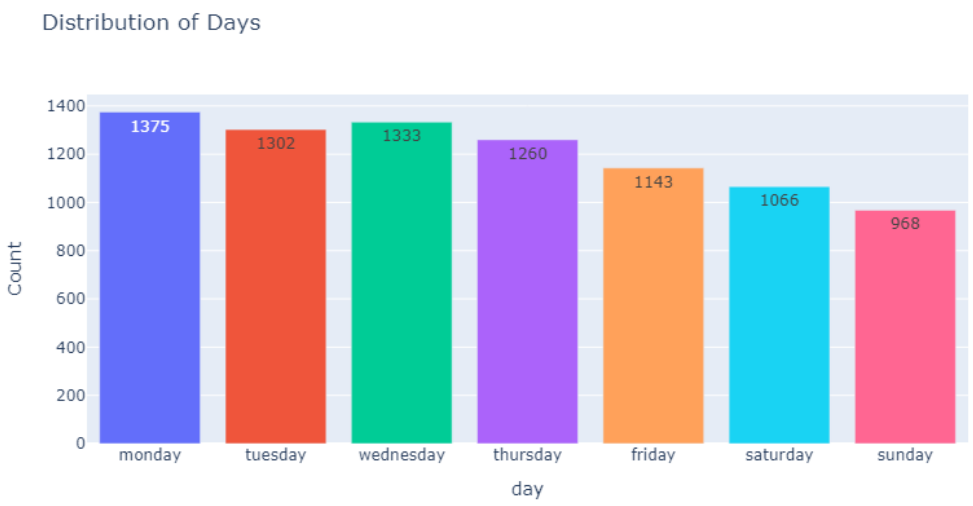


Figure 9 Distribution of Days

The matrix below (Fig. 10) is the correlation matrix among the days of the week.

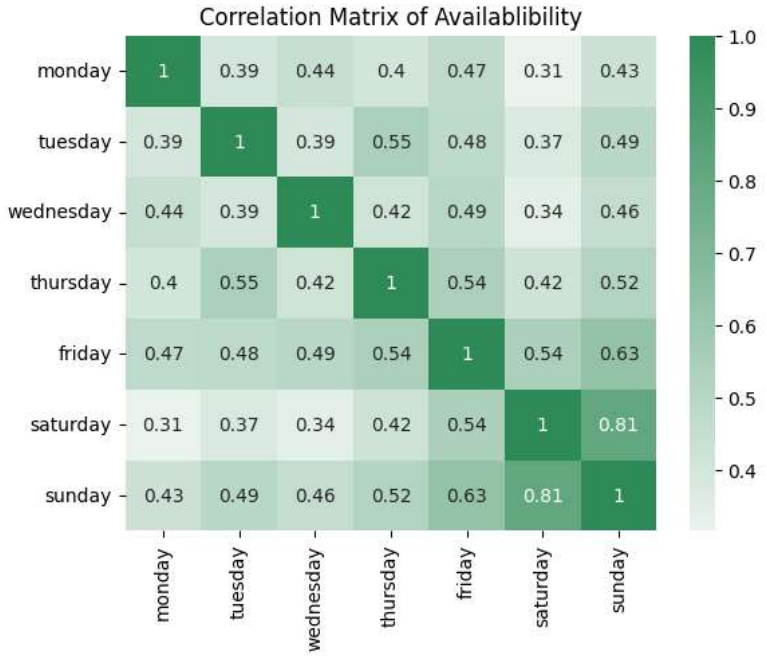


Figure 10 Correlation Matrix of Days

- **No Negative-Relations in Preferences:** From the above figure, it is evident that all the days are positively correlated with each other, implying that there are no reverse preferences as such, i.e. if a lead prefers Monday then it will not reduce the lead's preference for some other day, say Saturday.
- **Weak Overall Correlation:** Correlation among all the days, though positive, is not strong enough to suggest strong relations among days. Hence it is difficult to confidently identfiy groups of days which are preferred together.

- **Pairwise Correlation:** While the overall correlation is weak, the correlation between consecutive days or days which are at max 1 day apart is marginally higher than between non-consecutive days. For instance, Friday and Saturday have relatively higher correlation coefficient (0.54). Similarly, Friday and Sunday have a coefficient of 0.63, and Saturday and Wednesday have a weaker coefficient of 0.34.
- **Weekend Preference:** The only strong preference which can be identified from the above matrix is the "weekend preference". Saturday and Sunday have a correlation of 0.81, suggesting that both are demanded consecutively.

Interpretation of Results and Recommendations

- **Improve lead quality and conversion strategies:** As of now, the conversion rate of the leads is extremely low, at ~4% (86 ÷ 2092). Very few leads (~7%) even take the free trial class ("demo") offered by NCR.
  The reasons for this, as discussed earlier, are that leads are no longer interested in availing NCR's services or have enrolled in offline classes.
  Hence NCR needs to improve the quality of leads it is able to source. Since offline classes is preferred over online in the current set of leads, it is an indication that these leads might not be accurately targeted.
  NCR needs to source more leads which are looking specifically for online classes.
- **Rescheduled leads:** As discussed earlier, a lead which has resheduled the demo is highly unlikely to actually take the rescheduled demo class. There is a very high probability that this lead is inactive and no longer needs to be pursued actively. NCR's sales team should not focus on these leads and must close these leads with a simple reminder to take the demo class.
  Also, since a erscheduled demo is less likely to actually happen, it impacts the tutor availability as well: NCR does not need to actually block a tutor for a rescheduled demo unless the lead has explicitly confirmed that they will attend the demo class.
- **Pricing:** Analyzing the data revealed that NCR's pricing may not be a major blocker in converting leads. This is contrary to the first instinct in a business where poor conversion of leads into sales is attributed to higher pricing.
  There is no sufficient evidence in the data to suggest that price may be costing NCR sales. In this case, preference for offline classes and inaccurate leads are more likely reasons behind poor conversion than pricing.
- **Bundling strategies:** There were clear subject combinations which are in demand:
  - maths+english,
  - maths+science,
  - maths+english+science.

  Moreover, 80% of the leads which demand english or science, also demand maths. A similar percentage is found in converted leads as well, suggesting that maths as a subject is sold along with science and english.
  NCR should utilize this to create special bundles of subjects and offer incentives such as discounts on enrolling in these bundled tuition packages.
- **Market Expansion – Offline Tuition Segment:** Offline classes are the biggest

competing segment that NCR loses its leads to. NCR can look into the offline classes segment to expand into.

- **Market Expansion – New Segment:** The analysis revealed that the current market of NCR is primary and secondary school with focus on maths, english, and science. However, the data also revealed a new market opportunity, that is exam preparation. NCR can research more into the exam preparation market in Australia. Specifically the following 3 exams need to be looked into:
    - LANTITE
    - NAPLAN
    - ATAR

  They must, at the same time, evaluate their existing teachers' preparedness to take classes for these exams.
- **Scheduling batches:** While scheduling batches, NCR should consider the following two sets of batches:
    - Monday, Tuesday, Wednesday, and
    - Saturday-Sunday

  These two sets of batches are when the availability of tutors should be maxmium and downtime should be minimum. Also, for each batch, the same set of tutors should be available throughout the batch days to ensure consistency.
  While marketing their services, NCR should strategically highlight the weekday and weekend class schedules. At the same time, they should also convey flexibility in choosing days.

----------End of the report----------