EdTech Data Analysis

Created by  Anh Phan

"Understanding Lead Behavior in EdTech: Insights from Course Demos and Sales Interactions"

In the fast-paced world of EdTech, understanding what drives potential learners to engage—or disengage—with online courses is critical. This analysis dives into a rich dataset from an EdTech company, covering the journey of each lead from initial contact to conversion or dropout. The data provides a unique view of leads who attended course demos, the sales managers assigned to them, and each interaction along the way. Crucially, it also includes detailed information on where in the process leads lost interest and why.

By analyzing this information, we aim to answer questions such as:

- Which sources generate the most lead? Which generate the least?
- What are the most common reasons for not being interested in the demo?
- At which points in the sales funnel do leads most commonly drop off, and for what reasons?

This analysis will uncover insights into the lead conversion process, offering actionable strategies to help sales managers engage leads more effectively and improve conversion rates. For EdTech providers, understanding these dynamics is a valuable step toward creating a more responsive, data-driven approach to sales and lead engagement.

Link to dataset on Kaggle: Data Analyst

Data Cleaning

To start off with the dataset, I start by creating database, tables and importing data from CSV file into the tables. Then I did some data cleaning for every table, such as check for data duplicates, outliers, data errors, unique values. Some of them are:

1. lead_details table

- I use quartiles to check for outliers in lead_details table.
- There are 2 records in the table that has the age value of 116 and 221, which could be a typo when entering leads data. To resolve the issue, I replace these values with the median value.

2. not_interested_reason table

- I check for unique values in the table using DISTINCT function to have a quick glance at the columns.
- In the reason_not_interested_in_demo column, there are 2 separate values that have the same meaning are “Can't afford” and “Cannot afford”. To resolve this issue, I use UPDATE function to fix the wrong value.

Data Analysis

1. The first thing that I was curious about when looking at the leads data is which source generates most leads, and which one generates the least.

```
SELECT
    generated_source,
    COUNT(generated_source) AS count_source,
    COUNT(generated_source)*100/(SELECT COUNT(*) FROM lead_deta
FROM lead_details
GROUP BY generated_source
ORDER BY count_source DESC;
```

	generated_source character varying	count_source bigint	source_perc bigint
1	social_media	87	24
2	SEO	75	20
3	email_marketing	73	20
4	user_referrals	66	18
5	website	59	16

- Based on the output, social media generates the most leads, followed by search engine optimization and email marketing. The company's website generates the least number of leads.

2. From the demo_watched_details table, we have information regards what language does the leads watched the demo in and how much percentage of the demo did they watch.

```
SELECT
    percentage,
    COUNT(percentage),
    COUNT(percentage)*100/(SELECT COUNT(*) FROM demo_watched_de
FROM
(SELECT
    lead_id,
    CASE
        WHEN watched_percentage >= 0 AND watched_percentage <= 2
        WHEN watched_percentage > 20 AND watched_percentage <= 4
        WHEN watched_percentage > 40 AND watched_percentage <= 6
        WHEN watched_percentage > 60 AND watched_percentage <= 8
        ELSE '81-100'
    END AS percentage
FROM demo_watched_details)
GROUP BY percentage
ORDER BY percentage ASC;
```

	percentage text	count bigint	percentage_watched bigint
1	0-20	31	15
2	21-40	31	15
3	41-60	44	22
4	61-80	62	31
5	81-100	26	13

- Based on the output of the query above, a larger portion (31%) of leads watched a substantial but incomplete portion (61-80%) of the demo and only a small percentage of leads (13%) finished the entire demo.
- These numbers showing that many leads start watching the demo and watching a significant portion, but only a few are staying to the end. This could indicate that the demo captured initial interest but doesn't maintain it effectively until the conclusion.
- The drop off could suggest that the demo may be too long, overwhelming, and less engaging toward the later stage. Leads may feel they've seen enough without needing to complete the entire demo, or they may lose interest before reaching critical points that emphasize the product's full value.

3. I want to see which primary language is viewed the most in the demo of different cities. Knowing this, the company will be able to tailor its content and campaigns to each city.

```
SELECT
    ld.current_city,
    dw.demo_language,
    COUNT(dw.demo_language),
    ROUND(COUNT(dw.demo_language)*100/SUM(COUNT(dw.demo_language
FROM demo_watched_details dw
JOIN lead_details ld
ON dw.lead_id = ld.lead_id
GROUP BY ld.current_city, dw.demo_language
ORDER BY ld.current_city, COUNT(dw.demo_language) DESC;
```

	current_city character varying	demo_language character varying (20)	count bigint	percentage_within_city numeric
1	Bengaluru	English	22	68.75
2	Bengaluru	Telugu	7	21.88
3	Bengaluru	Hindi	3	9.38
4	Chennai	English	17	56.67
5	Chennai	Telugu	7	23.33
6	Chennai	Hindi	6	20.00
7	Hyderabad	English	21	48.84
8	Hyderabad	Telugu	18	41.86
9	Hyderabad	Hindi	4	9.30
10	Kochi	English	19	61.29
11	Kochi	Telugu	8	25.81
12	Kochi	Hindi	4	12.90
13	Mumbai	English	11	57.89
14	Mumbai	Telugu	7	36.84
15	Mumbai	Hindi	1	5.26
16	Visakhapatnam	English	22	56.41
17	Visakhapatnam	Telugu	15	38.46
18	Visakhapatnam	Hindi	2	5.13

- In the output we will get a list of cities, languages, number of demos viewed in a particular language and the percentage of that language viewed per city.
- The most viewed language in the demo of all cities is English, followed by Telugu and Hindi.
- In some cities like Bengaluru, the demos viewed in English account for more than half of the total. The demos viewed in Hindi in Visakhapatnam are very few, almost negligible, but in Chennai, the demos viewed in Hindi, though few, account for 20%.

4. What is the successful conversion rate? Can we draw some insights from the converted leads? What are the most common characteristics of these leads?

```
SELECT
    COUNT(call_reason)*100/(SELECT COUNT(*) FROM lead_details) A
FROM interaction_details
WHERE call_reason = 'successful_conversion';
-- Successful conversion rate: 17%

WITH converted AS
(SELECT *
FROM lead_details
WHERE lead_id IN
(
SELECT DISTINCT lead_id
```

```
FROM interaction_details
WHERE call_reason = 'successful_conversion'
))

SELECT
    generated_source,        -- change to different column to lo
    COUNT(generated_source),
    COUNT(generated_source)*100/(SELECT COUNT(*) FROM converted)
FROM converted
GROUP BY 1
ORDER BY 3 DESC;
```

	generated_source character varying	count bigint	percentage bigint
1	email_marketing	19	29
2	social_media	17	26
3	SEO	14	21
4	user_referrals	8	12
5	website	6	9

- I first created a CTE to filter only leads that have converted. Then from the CTE, I look into different columns to find the most common characteristics of these leads.
- 29% of successful conversion are people come from source of email marketing, 26% from social media.
- 40% of successful conversion are people with current education of Bachelor of Technology, 29% of people who are currently looking for a job.
- 67% of successful conversion are female, only 32% are male.
- 32% of successful conversion are people who has parent working in Business field, 28% are Government Employee.

5. Throughout the communication process between the junior sales manager and the leads assigned to them, starting from the first call of getting to know the lead, to inviting them to schedule a demo session, to follow up after the demo,

to follow up for consideration, and if possible, follow up for conversion. Out of all the leads that decline to continue the process at any stage, I'm curious which stage has the highest drop rate.

```
SELECT
    COUNT(reason_not_interested_in_demo)*100/(SELECT COUNT(*) FR
    COUNT(reason_not_interested_to_consider)*100/(SELECT COUNT(
    COUNT(reason_not_interested_to_convert)*100/(SELECT COUNT(*
FROM not_interested_reason;
```

	no_demo_perc bigint	no_consider_perc bigint	no_convert_perc bigint
1	55	26	17

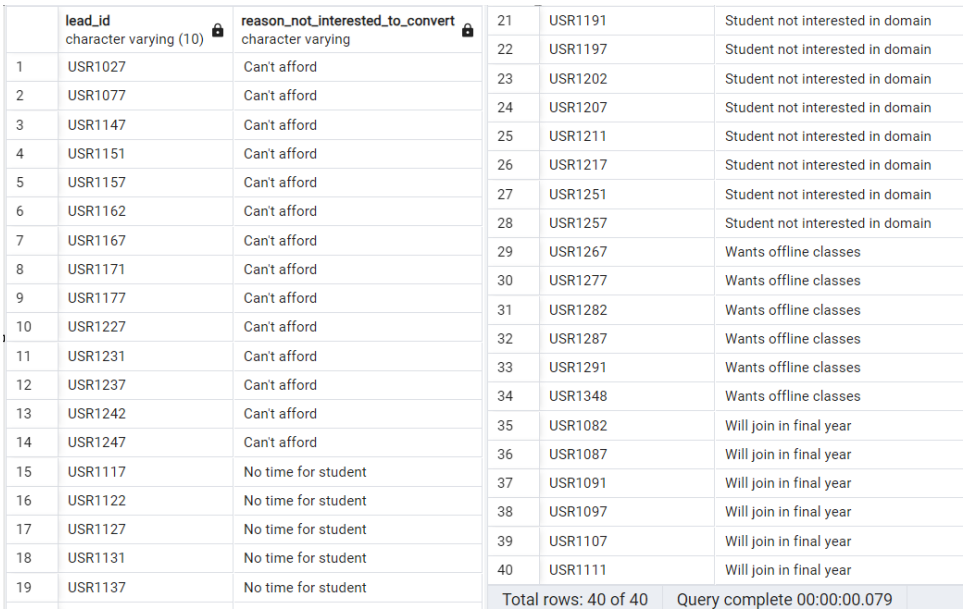
- Out of all the leads that dropped out at any stage during the communication process, 55% dropped due to not interested in demo, where some people watched the demo, some people don't.
- 26% dropped due to not interested in considering the course, and 17% dropped at the last stage of not wanting to convert.

6. The last stage during the communication process is the most important node of the chain because it is where the leads actually becoming the customer. I was curious if there are any leads that reached the last stage in the communication process, but ended up not converting? And if so, what are the reasons that made them change their mind in the end?

```
WITH followup AS
(SELECT *
FROM interaction_details
WHERE call_reason = 'followup_for_conversion'
ORDER BY lead_id ASC)

SELECT DISTINCT
    nir.lead_id,
    nir.reason_not_interested_to_convert
FROM followup fl
```

```
INNER JOIN not_interested_reason nir
ON fl.lead_id = nir.lead_id
ORDER BY nir.reason_not_interested_to_convert, nir.lead_id;
```



	lead_id character varying (10)	reason_not_interested_to_convert character varying
1	USR1027	Can't afford
2	USR1077	Can't afford
3	USR1147	Can't afford
4	USR1151	Can't afford
5	USR1157	Can't afford
6	USR1162	Can't afford
7	USR1167	Can't afford
8	USR1171	Can't afford
9	USR1177	Can't afford
10	USR1227	Can't afford
11	USR1231	Can't afford
12	USR1237	Can't afford
13	USR1242	Can't afford
14	USR1247	Can't afford
15	USR1117	No time for student
16	USR1122	No time for student
17	USR1127	No time for student
18	USR1131	No time for student
19	USR1137	No time for student
21	USR1191	Student not interested in domain
22	USR1197	Student not interested in domain
23	USR1202	Student not interested in domain
24	USR1207	Student not interested in domain
25	USR1211	Student not interested in domain
26	USR1217	Student not interested in domain
27	USR1251	Student not interested in domain
28	USR1257	Student not interested in domain
29	USR1267	Wants offline classes
30	USR1277	Wants offline classes
31	USR1282	Wants offline classes
32	USR1287	Wants offline classes
33	USR1291	Wants offline classes
34	USR1348	Wants offline classes
35	USR1082	Will join in final year
36	USR1087	Will join in final year
37	USR1091	Will join in final year
38	USR1097	Will join in final year
39	USR1107	Will join in final year
40	USR1111	Will join in final year

Total rows: 40 of 40 Query complete 00:00:00.079

- I started off by using the CTE to filter out all the call between leads and sales manager with call reason regarding follow up for conversion. Then I used JOIN clause to combine the CTE with the reason_not_interested_to_convert table to look into what are the reasons that made them ended up not converting.
- The CTE alone returns 102 leads were follow up for conversion and a total of 189 follow up attempts both successful and unsuccessful. After running the JOIN clause between the CTE and reason_not_interested_to_convert table, there are 40 leads ended up not converting due to several reasons such as pricing, scheduling, etc.
- Running a COUNT function on this query, it shows that most leads (14) changed their mind in the last stage due to pricing.

7. Since we're looking at the pricing problem, I would like to know what percentage of people drop out at any given point because of the price.

```
SELECT
    'not_intersted_in_demo' AS reason,
```

```
    COUNT(CASE WHEN reason_not_interested_in_demo = 'Can''t affo
    COUNT(reason_not_interested_in_demo) AS total_count,
    COUNT(CASE WHEN reason_not_interested_in_demo = 'Can''t affo
FROM not_interested_reason
WHERE reason_not_interested_in_demo IS NOT null

UNION ALL

SELECT
    'not_intersted_to_consider' AS reason,
    COUNT(CASE WHEN reason_not_interested_to_consider = 'Can''t
    COUNT(reason_not_interested_to_consider) AS total_count,
    COUNT(CASE WHEN reason_not_interested_to_consider = 'Can''t
FROM not_interested_reason
WHERE reason_not_interested_to_consider IS NOT null

UNION ALL

SELECT
    'not_intersted_to_convert' AS reason,
    COUNT(CASE WHEN reason_not_interested_to_convert = 'Can''t a
    COUNT(reason_not_interested_to_convert) AS total_count,
    COUNT(CASE WHEN reason_not_interested_to_convert = 'Can''t a
FROM not_interested_reason
WHERE reason_not_interested_to_convert IS NOT null;
```

	reason text	cant_afford_count bigint	total_count bigint	cant_afford_perc bigint
1	not_intersted_in_demo	48	164	29
2	not_intersted_to_consider	32	79	40
3	not_intersted_to_convert	19	51	37

- Out of 164 leads that dropped due to lack of interest in the demo, 29% cited price as the main reason. Out of 79 leads that dropped due to not interested to consider, 40% cited price as the reason. Out of 51 leads that dropped due to not interested to convert, 37% said that they can’t afford the course.

8. Out of all the people who dropped out midway, I think there are some who are still in school and that's why they can't arrange the time. I'm curious to know who dropped out midway at any stage with the excuse of "Will join in final year" so that the sales manager can follow up with them later.

```
SELECT
    sma.jnr_sm_id,
    nir.lead_id,
    ld.current_education,
    nir.reason_not_interested_in_demo,
    nir.reason_not_interested_to_consider,
    nir.reason_not_interested_to_convert
FROM not_interested_reason nir
INNER JOIN lead_details ld
ON nir.lead_id = ld.lead_id
LEFT JOIN sales_manager_assigned sma
ON ld.lead_id = sma.lead_id
WHERE reason_not_interested_in_demo = 'Will join in final year'
OR reason_not_interested_to_consider = 'Will join in final year
OR reason_not_interested_to_convert = 'Will join in final year'
```

	jnr_sm_id character varying (10)	lead_id character varying (10)	current_education character varying	reason_not_interested_in_demo character varying	reason_not_interested_to_consider character varying	reason_not_interested_to_convert character varying
1	JNR1001MG	USR1008	B.Tech	[null]	Will join in final year	[null]
2	JNR1001MG	USR1010	B.Tech	[null]	Will join in final year	[null]
3	JNR1005MG	USR1082	B.Tech	[null]	[null]	Will join in final year
4	JNR1005MG	USR1087	Looking for Job	[null]	[null]	Will join in final year
5	JNR1005MG	USR1091	B.Tech	[null]	[null]	Will join in final year
6	JNR1005MG	USR1097	Intermediate	[null]	[null]	Will join in final year
7	JNR1006MG	USR1105	Intermediate	[null]	[null]	Will join in final year
8	JNR1006MG	USR1107	Intermediate	[null]	[null]	Will join in final year
9	JNR1006MG	USR1111	B.Tech	[null]	[null]	Will join in final year
10	JNR1001MG	USR1353	Looking for Job	Will join in final year	[null]	[null]
11	JNR1001MG	USR1354	B.Tech	Will join in final year	[null]	[null]
12	JNR1001MG	USR1355	Looking for Job	Will join in final year	[null]	[null]
13	JNR1001MG	USR1359	B.Tech	Will join in final year	[null]	[null]
14	JNR1001MG	USR1360	Intermediate	Will join in final year	[null]	[null]

- Knowing the list of leads who dropped and stated that they “will join in final year”, sales manager will have a plan to follow up with these leads in order to maintain their interest.

9. Wanting to know how efficient the team of sales manager are doing. I look into the converted rate of each junior sales manager.

```
WITH converted AS
(
SELECT
    jnr_sm_id,
    COUNT(lead_id) AS converted_number
FROM interaction_details
WHERE call_reason = 'successful_conversion'
GROUP BY jnr_sm_id
ORDER BY jnr_sm_id ASC),

total AS
(SELECT
    jnr_sm_id,
    COUNT(lead_id) AS total_number
FROM sales_manager_assigned
GROUP BY jnr_sm_id
ORDER BY jnr_sm_id ASC)

SELECT
    total.jnr_sm_id,
    converted.converted_number*100/total.total_number AS convert
FROM total
LEFT JOIN converted
ON total.jnr_sm_id = converted.jnr_sm_id;
```

	jnr_sm_id character varying (10)	converted_rate bigint
1	JNR1001MG	17
2	JNR1002MG	35
3	JNR1003MG	30
4	JNR1004MG	20
5	JNR1005MG	10
6	JNR1006MG	20
7	JNR1007MG	10
8	JNR1008MG	20
9	JNR1009MG	10
10	JNR1010MG	19
11	JNR1011MG	7
12	JNR1012MG	12
13	JNR1013MG	10
14	JNR1014MG	20
15	JNR1015MG	14
16	JNR1016MG	29

- I got the converted rate for each junior sales manager by creating CTE to divide number of successful conversions by the total of leads assigned to each sales manager.
- This could help the senior sales manager to manage and monitor their subordinates more closely, especially those with alarmingly low conversion rates.

10. As far as the communication process between a manager and a prospect goes, the communication only really ends when the lead becomes a customer, or when they decide to drop out mid-process. Therefore, I was curious if there are any situation when sales manager failed to continually the interaction process, for example the lead was still showing interest and there's no sign of rejection, but sales manager failed to reach out for next step.

```
WITH ranked AS
(SELECT
    ids.lead_id,
    ids.call_date,
    ids.call_reason,
    ROW_NUMBER() OVER(PARTITION BY ids.lead_id) AS rn
FROM interaction_details ids
```

```
LEFT JOIN not_interested_reason nir
ON ids.lead_id = nir.lead_id
WHERE call_date =
(
SELECT MAX(call_date)
FROM interaction_details ids2
WHERE ids.lead_id = ids2.lead_id
)
AND nir.lead_id IS NULL)

SELECT
    lead_id,
    call_date,
    call_reason
FROM ranked
WHERE rn =
(
SELECT MAX(rn)
FROM ranked AS r
WHERE r.lead_id = ranked.lead_id
)
AND call_reason != 'successful_conversion';
```

	lead_id character varying (10)	call_date date	call_reason character varying
1	USR1072	2022-01-20	interested_for_conversion
2	USR1093	2022-01-29	followup_for_conversion
3	USR1113	2022-01-20	demo_schedule
4	USR1133	2022-01-20	demo_schedule

- To start off, I create a CTE to find a list of last interactions between leads and sales manager, in which I use LEFT JOIN clause to combine with not_interested_reason table to only select records that leads do not show up in not_interested_reason table. In another word, I filter the data to only show records of leads that did not drop out of the interaction process, and I use ROW_NUMBER function to number each row within a specific partition

(lead_id). The output I get from the CTE alone are list of last interactions with different leads who did not choose to drop out of the process themselves.

- Next, I select only records that are very last interaction between sales manager and a specific lead, and that last interaction wasn't a successful conversion.
- In the output, I get 4 records shows the lead ID, the last call date, and the reason for the call. These records show 4 leads and the last time the sales manager reached out to them. These 4 leads have not explicitly expressed disinterest in continuing with learning about the course or converting, but sales manager failed to continue to follow up with them.

Actionable Insights

1. **Bring products closer to leads using different channels:**

- Social media is currently the most significant driver of leads. The company should allocate additional resources to create and optimize ad campaigns across diverse social media platforms. The analytics team should conduct a more detailed analysis to identify high-performing platforms and tailor strategies accordingly.
- Invest more in search engine advertisements, such as Google Ads, to capture the attention of users actively searching for related services. Focus on targeted keywords and ad placements to maximize return on investment.
- Increase the frequency and personalization of marketing emails to leads. These emails should include comprehensive course details, exclusive offers, and special promotions to engage potential customers and encourage conversions.
- Since the company's website generates the fewest leads, enhancements are necessary to boost its effectiveness. Consider upgrading the design, improving the user experience, and implementing engaging features to capture visitor interest. A/B testing and feedback from users can guide these changes.

2. **Demo:** To improve engagement and completion rates, the demo should be shortened while ensuring it delivers all essential information. The demo

content should be refined to make it both informative and interesting to keep leads entertained until the end of the demo. Getting a lead's interest from a demo is important because it will keep them engaged longer in the whole process and will increase conversion rate.

3. **Localized Strategies:** In Bengaluru, where English accounts for more than half of the total demos, focusing on English content will maximize reach. In Visakhapatnam, Hindi demos have very low viewership, highlighting a need to deprioritize Hindi content and focus on Telugu and English for better impact. In Chennai, even though Hindi demos are relatively few, they account for 20% of views, suggesting a niche but significant audience that could benefit from targeted Hindi content. To boost engagement, a localized approach should be adopted—developing more English demos for cities like Bengaluru, emphasizing Telugu in Telugu-speaking regions, and offering selective Hindi content where demand exists, such as in Chennai.
4. **Strategizing Conversions Based on Lead Profiles:** Strengthen email marketing and social media campaigns to leverage their proven effectiveness. Develop content and messaging tailored to Bachelor of Technology students and job seekers, focusing on career-oriented benefits. Design campaigns that resonate with female audiences, potentially through personalized offers or relatable narratives. Consider marketing strategies that subtly align with the values of families in business and government sectors to further improve conversions.
5. **Targeting specific drop-off stages with tailored strategies:** Besides refining the demo as mentioned earlier to retain leads' interest, several things the company can do with leads during the consideration phase such as emphasize the value and unique selling points of the course during communication, such as career benefits, success stories, or special offers. With leads who are at the last stage of the process, the company should provide tailored incentives, such as limited-time discounts or flexible payment plans, to encourage conversion. Urge the decision-making process by using urgency tactics like deadlines for enrollment or limited slots. Based on the drop rate at each stage, the company will be able to allocate resources and finance to focus on each one more sufficiently.

6. **Last-Stage Leads Who Dropped-off:** During the interaction with last-stage leads who wants to drop-off from the interaction process, the company can apply different resolutions to different concerns from the leads. For those who drop out due to pricing reasons, the company can offer some flexible pricing, such as installment plans, discounts, or scholarship to make the course more accessible depends on leads' financial situation. For those who drop out due to time constraints, the company can consider providing self-paced options or recording of study sessions to accommodate busy schedules. For leads who plan to join in later, the company and managers should stay connect by using periodic follow-ups, newsletter, reminders to keep them engaged until they are ready to enroll.
7. **Sales Team Performance:** Conduct a performance review of the sales team to identify gaps in their follow-up practices. Provide training sessions on lead management and the importance of consistent follow-ups in maintaining engagement. Establish a structured follow-up system to ensure no lead is neglected, especially those who have not explicitly expressed disinterest. Implement CRM tools with automated reminders to track follow-up actions and ensure timely communication.