



# PROPOSITION FOR LEAD SCORING SYSTEM AND BUSINESS GROWTH

- CASE STUDY FOR A FICTITIOUS ONLINES COURSE COMPANY

BY HUAN DENG

# TABLE OF CONTENTS

- Problem Statement
  - 1.1 Project Objective
  - 1.2 Project Overview – Lead Scoring Design
- Current Situation: Where We Are
  - 2.1 Overview
  - 2.2 Customer Profile
  - 2.3 Customer Behavior
- Classification model building
  - 3.1 Model Selection Strategies
  - 3.2 Model Implications
- Recommendations
  - 4.1 Marketing/Sales Strategies
  - 4.1 Next Step: Follow-up with Marketing/Sakes Team

# 1.1 PROJECT OBJECTIVE

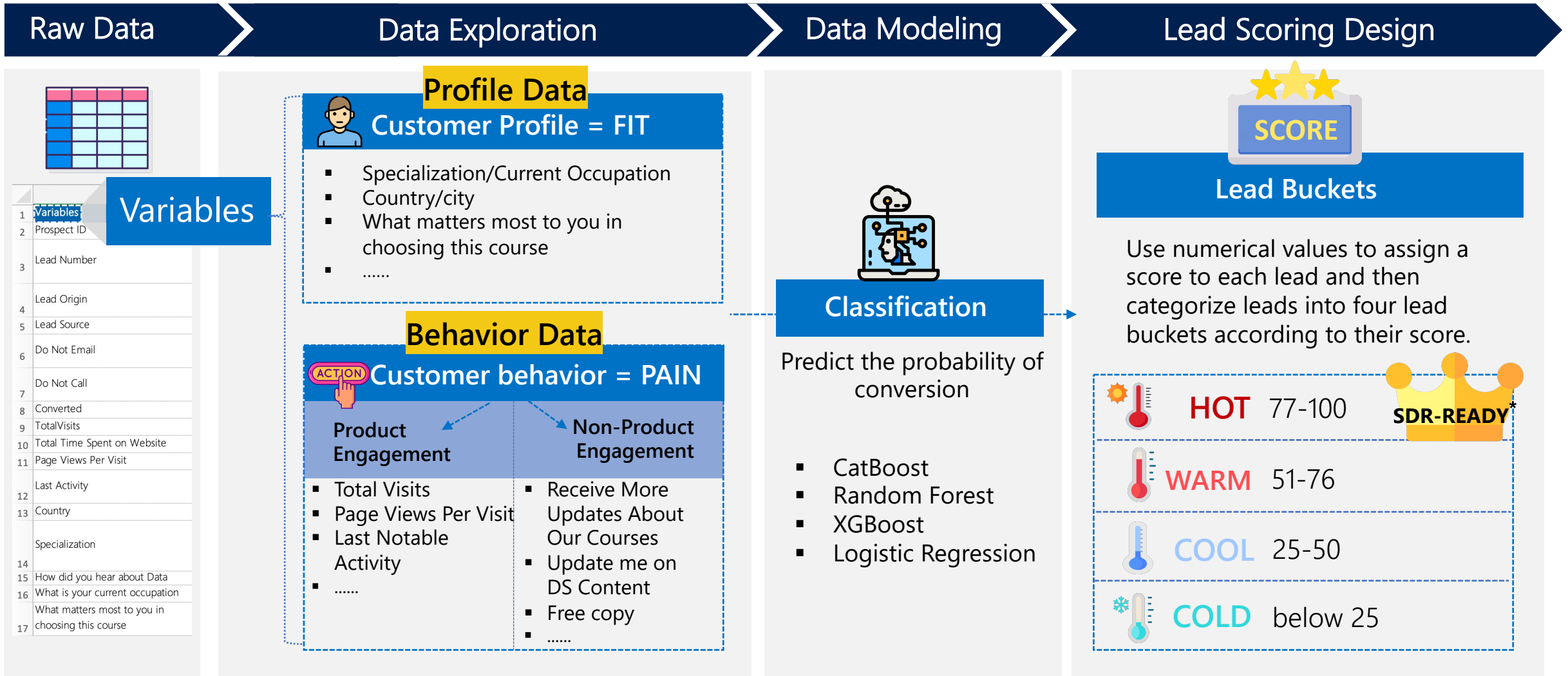
This fictitious company aims to

- 1 Prioritize sales resources
  - 2 Improve conversion rates
  - 3 Shorten sales cycle
- by building a lead scoring system.



# 1.2 PROJECT OVERVIEW - LEAD SCORING DESIGN

A PROCESS OF ASSIGNING SCORES TO PROSPECTS BASED ON PREDICTION OF CONVERSION PROBABILITY DATA IN ORDER TO PRIORITIZE LEADS.



\*SDR: Sales Development Representative

# 2.1 CURRENT SITUATION - OVERVIEW

Total Leads: 9,240				Website Analytics			
5679 (61.5%) non-converted		3561 (38.5%) converted		Average Total Visits 3.2	Average Total Time 479 =8 min	Average Pages Per Visit 2.3	
Lead Origin/Source/Quality				Tags			
Lead Origin		No	Yes	Tags		No	Yes
Landing Page Submit		3,118	1,768	Will revert after reading email		65	2,007
API		2,465	1,115	Ringing		1,169	34
Lead Add Form		54	664	Current Student →		566	13
Lead Source		No	Yes	Insterested		506	20
Google		1,726	1,147	No Response		347	112
Direct Traffic		1,725	818	Closed by Horizzon		2	356
Olark Chat		1,307	448	Lost		11	171
Organic Search		718	436	invalid number or not provided		156	1
Lead Quality		No	Yes	Not doing further education →		144	1
Others		3,743	1,024	Interested in full time MBA →		114	3
Might be		381	1,179	Diploma holder (Not Eligible)→		62	1
Not Sure		826	266	Have Question		11	3
High in Relevance		34	603				
Worst		589	12				

## Findings:

1. Generally, the company has decent conversion rates, and the website has high levels of user engagement.
2. Most leads filled up the form through Landing Page submission and API.
3. Google Ads was the strongest lead source, but more leads came from Direct Traffic, Olark Chat, and Organic search (combined).
4. Overall, the leads reacted positively to selling pitches. Almost all leads contacted by email were converted – indicating email may be an effective channel.
5. Especially, a lot of leads tagged with diploma information (see red arrow) are not converted. Based on educated guess, the reason of tagging that information can be this **company not only sells data science courses (low cost) but also offer diploma-related program (high cost)**, just like Coursera does.  
=> important assumption for further analysis

\*Yes: number of converted leads. No: number of unconverted leads.  
Tables are descending by the total number of leads.

## 2.2 CURRENT SITUATION - CUSTOMER PROFILE

Occupation		No	Yes
	Unemployed	3,159	2,441
	Working Professional	59	647
	Student	132	78
	Housewife		10
	Businessman	3	5
Specialization		No	Yes
	Finance Management	540	436
	HR Management	460	388
	Marketing Management	430	408
	Operations Management	265	238
	Business Administration	224	179
	IT Projects Management	226	140
	Supply Chain Managem..	198	151
	Banking, Investment An..	171	167
	Media and Advertising	118	85
	Travel and Tourism	131	72
	International Business	114	64

### Findings:

1. More than half of the leads were unemployed. **It is possible that leads' motivation to study is to land a job.**
2. Almost all working professionals were converted. **Since working professionals tend to have higher purchasing powers, we may need to involve more working professional customers.**
3. Most specializations were related to business/management. **It is possible that leads want to study DS courses associated with business.**

Location		
<b>TOP 1</b>		
Country:	India	
Count of City:	4,374	
<b>TOP 2</b>		
Country:	United States	
Count of City:	45	
<b>TOP 3</b>		
Country:	United Arab Emirates	
Count of City:	25	

City	
Mumbai	3,222
Thane & Outskirts	752
Cities of Maharashtra	457
Metro Cities	380
Tier II Cities	74

### Findings:

1. Nearly half of the leads located in India while the rest were scattered around the world.
2. In India, 74% leads were from Mumbai. To better market the service in India, it is necessary to examine what are the factors that contributed to Mumbai's dominant market share.



## 2.2 CURRENT SITUATION - CUSTOMER PROFILE

Likes & Contact	
What matters?	How did you hear about DS course?
Better <b>Career Prospect</b> 6,528 (99.97% chosen)	<b>TOP 1</b> Online Search - 808 <b>TOP 2</b> Word of Mouth - 348 <b>TOP 3</b> Student of School - 310
Free copy of case study?	Phone/Email Permission
<b>YES</b> - 2,888 (35.7% converted) <b>NO</b> - 6,352 (39.8% converted)	<b>YES Phone</b> - 8,506 (35.7% converted) <b>YES Email</b> - 9,238 (29% converted)
Update content/Payment Method	
<ul style="list-style-type: none"><li>▪ No update on DS/DM content/course</li><li>▪ Do not agree to pay through check or credit card.</li></ul>	

### Findings:

1. Almost all leads attributed better career prospect as the reason for choosing the course. **It is highly possible that career development is the main motivation for prospects to convert.**
2. Most of the leads heard this company from online search while others from classmates and word-of-mouth. **But we don't know the exact type of online search (organic, paid, etc.). Considering much fewer people heard the DM course from social media and advertisement, it is rather safe to say that this company needs to improve their communication strategies.**
3. About 90% leads opted in phone and email communication and eventually 30% of them were converted. **We need to further utilize phone and email to communicate with leads.**

## 2.3 WHERE WE ARE – CUSTOMER BEHAVIOR

Last Activity		No	Yes
	Email Opened	2,184	1,253
	SMS Sent	1,018	1,727
	Olark Chat Conversation	889	84
	Page Visited on Website	489	151
	Converted to Lead	374	54
	Email Bounced	300	26
	Email Link Clicked	194	73
	Form Submitted on Websi..	88	28
# of leads saw ad	DS Forums	1.00	
	Digital Advertisement	4.00	
	Magazine	0.00	
	Newspaper	1.00	
	Newspaper Article	2.00	
	Search	14.00	

Findings:

- As for last activity, most of the leads opened/clicked the email, sent SMS, initiated Olark chat conversation, or visited the website. **We need to later examine which last activity(s) is significantly associated with conversion in order to optimize communication strategy.**
- Very few customers claimed they have seen the ad on search (result), magazine, newspaper article, data science online forums, or digital advertisement. **The number of leads who saw the ad from "search" (14 leads) was much smaller than the number of leads who had Google Ads as lead source (>3000 leads). This discrepancy calls for further investigation**

Lead Source	No	Yes
Google	1,726	1,147

Contradicts



# KEY TAKEAWAYS

- The company offers data science courses (low cost) and diploma-related program (high cost).
- Key motivation to study online courses is landing a job/career development. It is possible that working professionals have the strongest purchase power.
- Except for Google Ads, the overall advertising performed poorly.
- Email and phone call are effective communication tools to nurture leads.
- Main business market is in India and Mumbai.
- Last activity seems to be highly correlated with conversion.

# 3.1 MODEL SELECTION STRATEGY

HOW TO SELECT MODELS AND MAKE STRATEGY TO MEET THE CURRENT BUSINESS NEEDS?

## PERFORMANCE VS INTERPRETABILITY

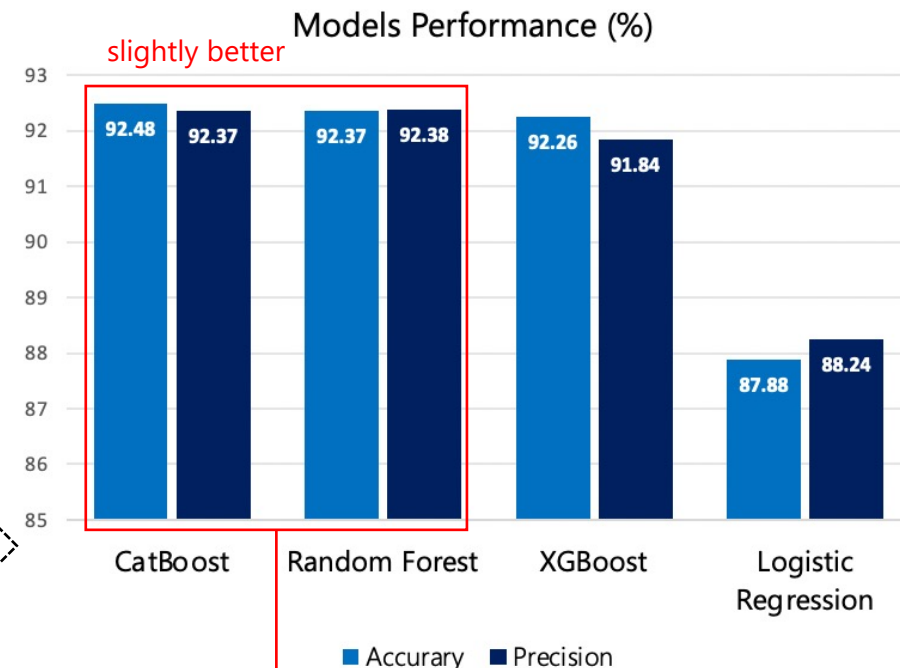
Generally, more complex models like XGBoost tend to have **better performance** but **less interpretability** while the relatively simple models, like logistics regression, are the other way around.



## CURRENT SALES OBJECTIVE

- **Diploma-related program ✓**
  - Price: high cost
  - How to convert: sales involved
- Data science course – *not current objective\**
  - Price: Low cost
  - How to convert: drip emails

- Find a balance between performance and interpretability.
- To not waste any sales resources, we prefer a classifier that rejects many good leads (relatively low recall) **but keeps only the true hot leads** (high precision). In this sense, we prefer use **accuracy** and **precision** as main metrics to evaluate model performance.



## After Applying K-fold Cross Validation

### CatBoost:

Accuracy: 92.98 %, Standard Deviation: 0.84 %

### Random Forest:

Accuracy: 92.59 %, Standard Deviation: 0.85 %

**CatBoost Wins!**

\* This is an assumption that I made. In real business situation, I will confirm with the marketing/sales team.

# 3.2 BEST MODEL - CATBOOST

## Data Cleaning

- NA value/Select -> mean value/"Others"
- Put new label with few results  
=> avoid shift between train and test data
- Outliers -> 95% quantile value
- Convert variable to factor
- Reduce redundancy by merging similar categories

## Data Processing

- One-Hot Encoding
- Data Splitting
- Feature Scaling

## Data Modeling

- **Grid Search**  
=> Find the best parameters  
No need for CatBoost
- Model Building
- K-Fold Validation

Converted	Prediction Class	Conversion Probability	Lead Score	Lead Buckets
1828	0	0	0.04	4
1829	0	0	0.01	1
1830	0	0	0.01	1
1831	0	0	0.00	0
1832	1	1	0.94	94
1833	0	0	0.18	18
1834	0	0	0.01	1
1835	0	0	0.07	7
1836	0	0	0.01	1
1837	1	1	0.72	72
1838	1	1	0.98	98

multiplied by 100

## CatBoost Advantages

- Great quality without parameter tuning
- Reduce overfitting

## Performance

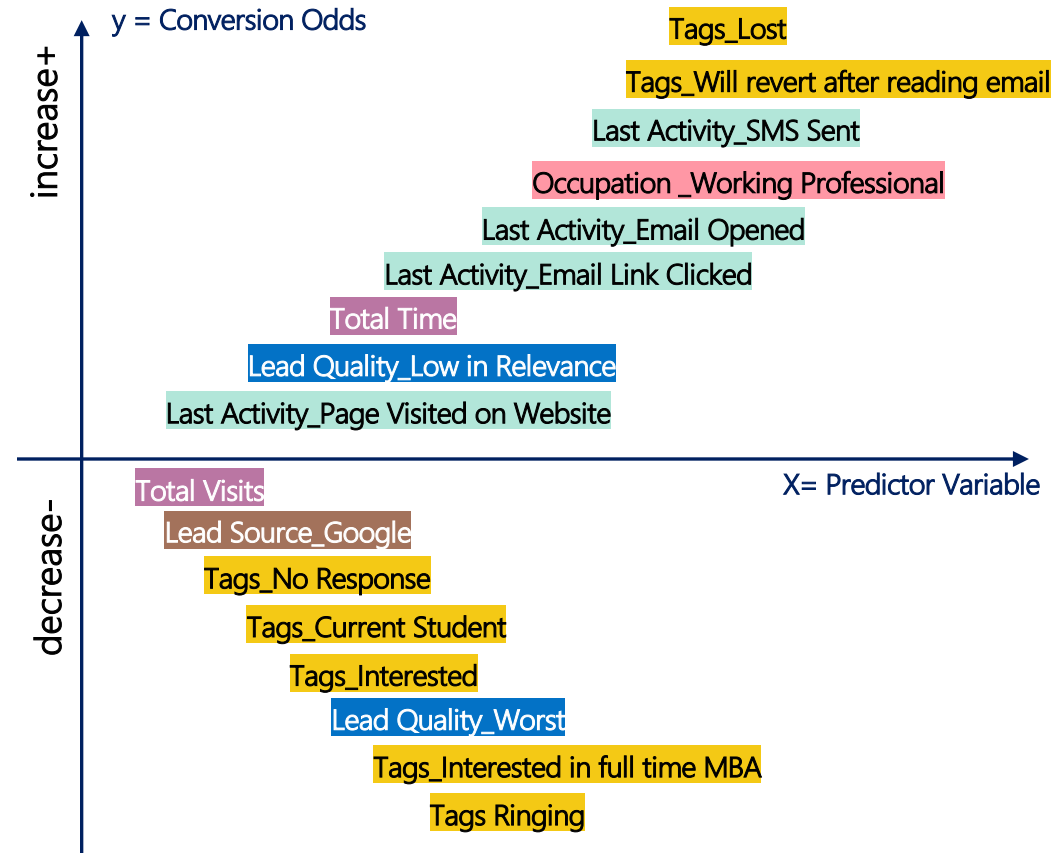
- Accuracy = 92.4784%
- Precision = 92.3685%, Recall = 91.6803%
- F1 = 91.9975%

## Confusion Matrix

True Label	0	1
	<b>1081</b>	84
1	55	<b>628</b>
Predicted Label		

## 3.2 MODEL IMPLICATION – LOGISTIC REGRESSION

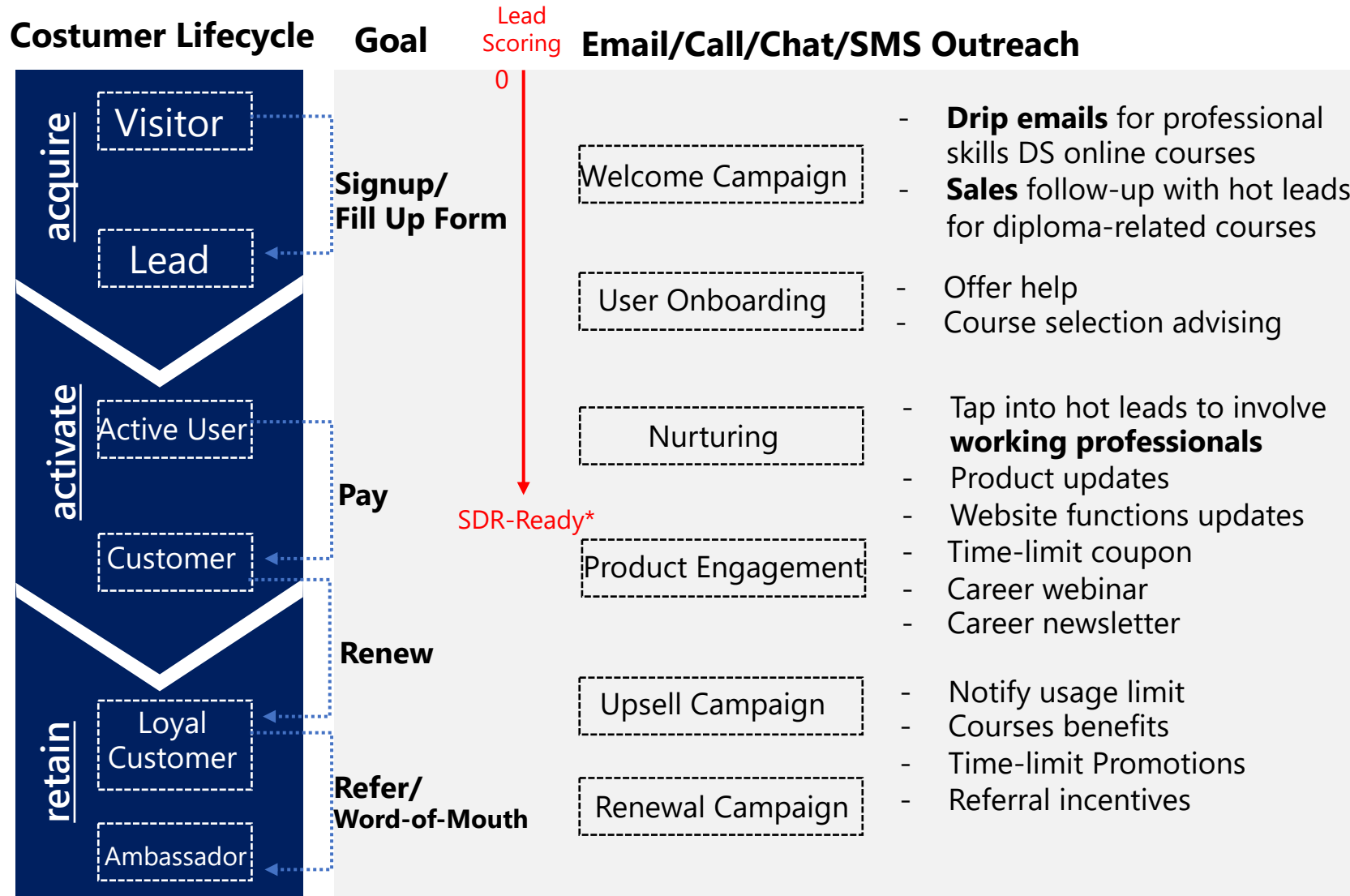
```
logit(p) = +3.76*Tags_Lost
          +3.53*Tags_Will revert after reading the email
          +2.41*Last Activity_SMS Sent
          +1.96*CurrentOccupation_Working Professional
          +1.21*Last Activity_Email Opened
          +0.97*Last Activity_Email Link Clicked
          +0.92*TotalTime
          +0.83*Lead Quality_Low in Relevance
          +0.64*Last Activity_Page Visited on Website
          -0.31*TotalVisits
          -0.34*Lead Source_Google
          -0.78*Tags_No Response
          -2.13*Tags_Current Student
          -2.51*Tags_Insterested
          -2.91*Lead Quality_Worst
          -3.23*Tags_Interested in full time MBA
          -3.50*Tags_Ringing
          -19.64
```



The coefficients show the change in log(odds) in conversion, for a unit change in the predictor variable, holding all other predictor variables constant.

- Roughly speaking, for **occupation**, working professional is positively associated with conversion.
- For **last activity**, leads who sent SMS, opened/clicked Email, or visited website pages are positively related with conversion.
- **Tags** with education info are all negatively related with conversion. It implies that **Diploma-related program didn't sell well**.
- **Lead quality** label correctly reflects conversion.
- Leads tend to convert when they spend more **time** on website while **total visits** are not necessarily the case.

# 4.1 MARKETING/SALES STRATEGIES



## Communication

### #Career-oriented

- Skills obtained from courses
- Course curriculum
- Payment plan
- Free trial courses

### #Working skills for Next Job

- Skills obtained from courses
- Service information
- Mock interview courses

### #How to survive in a changing Job Market

- Educate the importance to learn in a changing job market
- Referral promotion info

\* Make sure that SDR team can effectively follow up with every lead that has reached an SDR-ready score within 48 hours.

## 4.2 NEXT STEP: FOLLOW-UP WITH MARKETING/SALES TEAM

*Hypothesis 1: Except for Google ads, other advertising methods did not work.*

To be done:

- Figure out whether it is because the survey “How many lead had seen the ads” did not reflect the real situation or advertising outlet selection fails.
  - Check Google analytics/campaign report.
  - Check if the target audience of advertising media matches product users.

*Hypothesis 2: Career development is the biggest motivation for users to purchase online courses.*

To be done:

- Conduct A/B test to see how leads react to the courses with career information/or not.
- Conduct survey to collect users’ needs for career related courses.

*Hypothesis 3: Leads prefer low-priced professional skills online courses than diploma-related courses.*

To be done:

- Figure out the reason why leads don’t like diploma-related courses. Is it because of sales strategies, product design, or users’ preference?

*Hypothesis 4: Increase Email/SMS/Call communication frequency can enhance conversion rates.*

To be done:

- Conduct A/B tests with different communication frequency to see if conversion rates differ.



# FINAL THOUGHTS

- To keep the model simple and efficient, **current lead scoring system is static** based on past data on customer profile, action, and other features. **We can make lead scoring dynamic** by constantly scoring leads with new data and improve its accuracy.
- We should always keep in mind that the **main goal of building a lead scoring system is to prioritize sales resources** and therefore sales can allocate their capacity effectively to convert leads.
  - If the course price is not high enough to warrant in-person sales pitches, **email could be a better cost-effective communication channel.**
  - This company's business model decides how much efforts we should put into improving accuracy and timeliness for the lead scoring system and therefore making it cost-efficient.

THANK YOU. MERCI. GRACIAS. VIELEN DANK. 谢谢!

---

“A STORYTELLER WITH INTELLECTUAL CURIOSITY  
ABOUT DATA AND CONSUMER BEHAVIOR.”

-HUAN DENG