NLP CoE

Kristin Chen, Aishwarya Bhangale, Raj Desai, Meet Paradia, Wenxi Li

Purpose

- Find dataset and research problems aligns with business needs and industrial trend

Problem 1: Drug Review

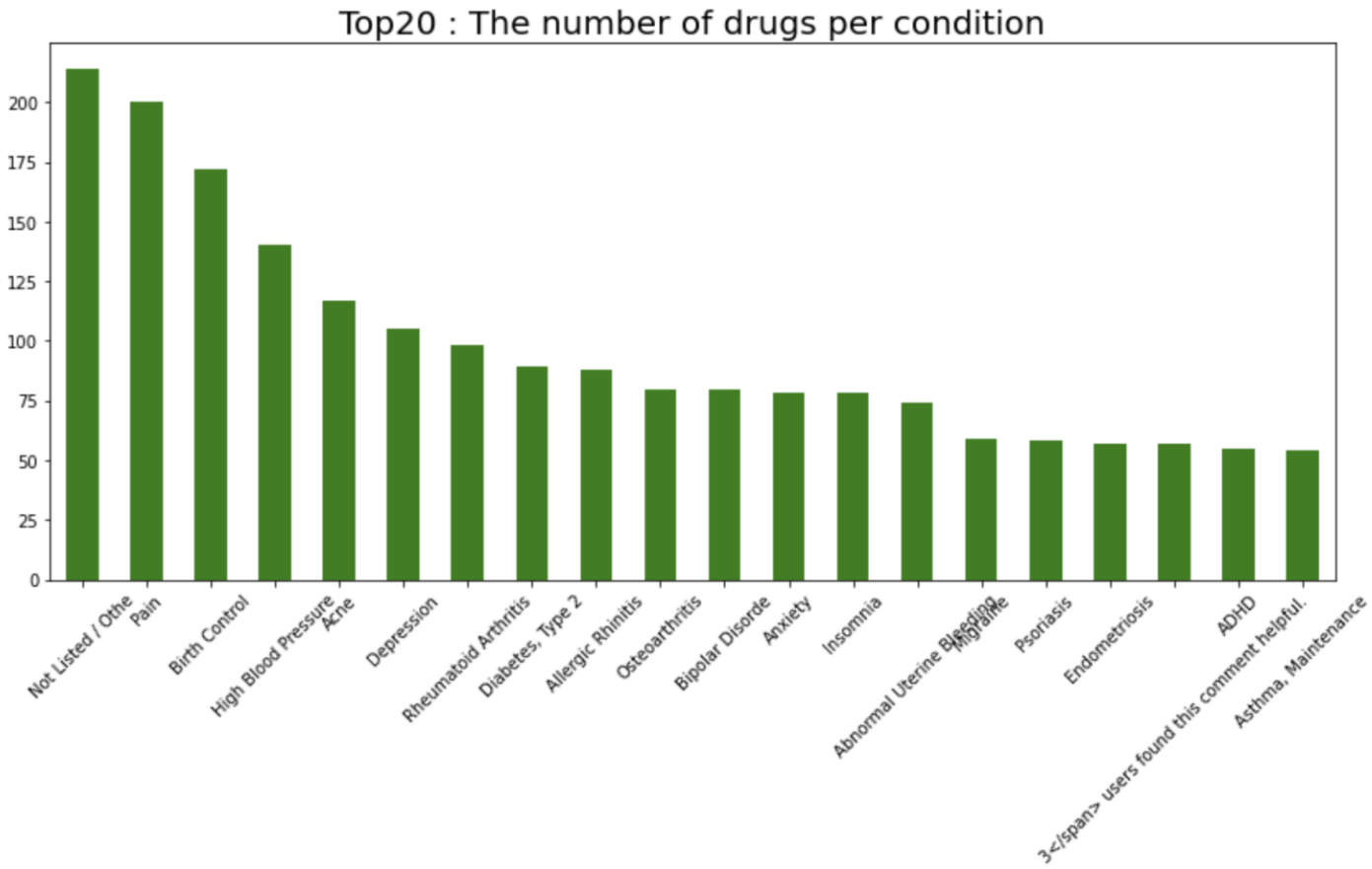
- Data: 161,297 unique reviews for train dataset about drugs related to 884 different health conditions (rating from 1 to 10), and the number of users who found review useful (ranging from 0 to 1291)
- Problem:
  - Text classification: classify health conditions and rating based on the review
  - Regression: predict the rating of the drug based on the review
  - Sentiment analysis
- Potential business application: drug reviews for healthcare/pharma industries

Reference: https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

Problem 1: Drug Review

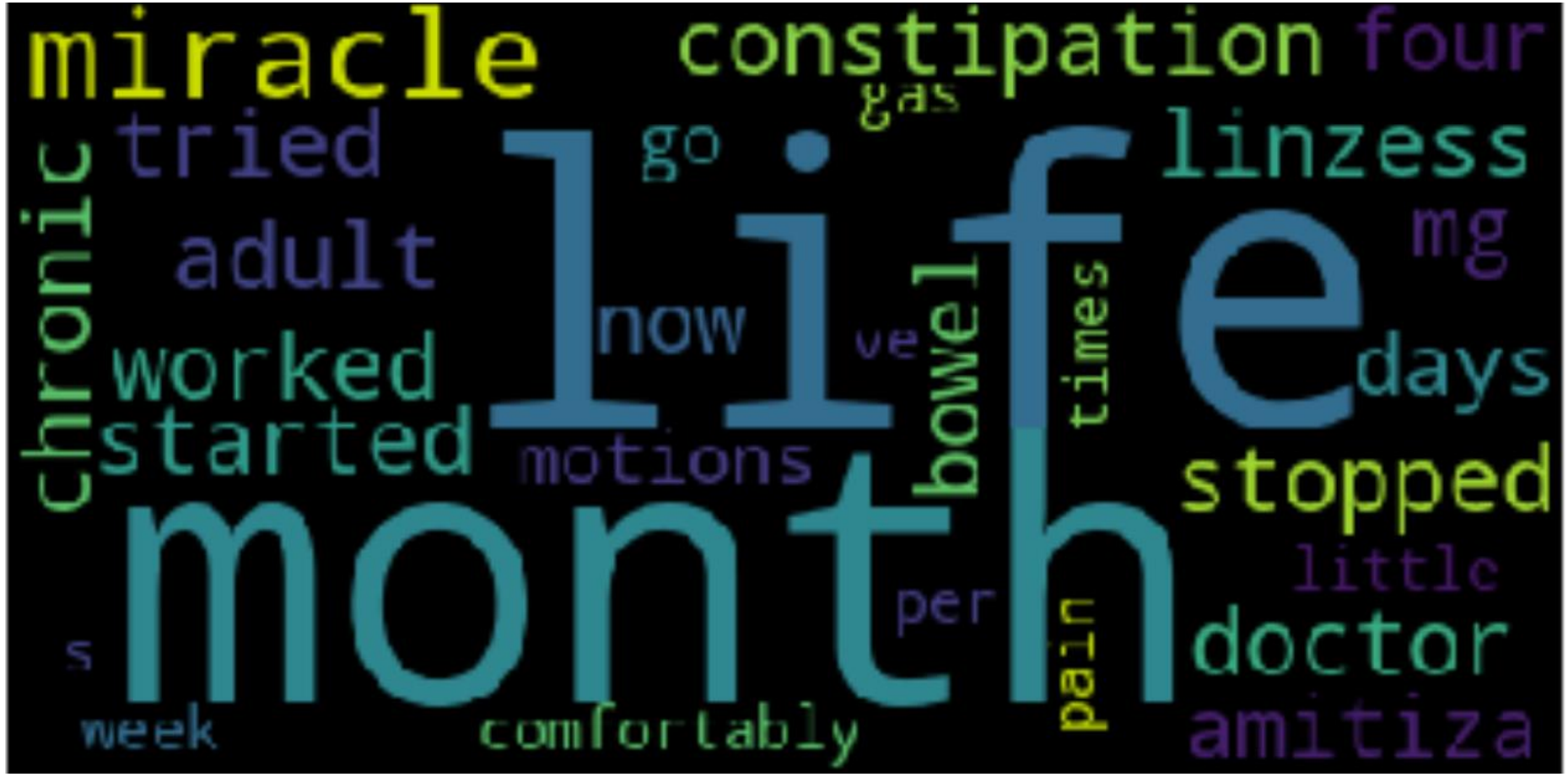
- There are 885 unique conditions , 3436 unique drugs and 389 unique usefulCounts.
- The reviews are given in the timeframe between 1-Apr-08 to 9-Sep-17.

Problem 1: **Drug Review**



Condition	
Not Listed / Other	214
Pain	200
Birth Control	172
High Blood Pressure	140
Acne	117
Depression	105
Rheumatoid Arthritis	98
Diabetes, Type 2	89
Allergic Rhinitis	88
Osteoarthritis	80
Bipolar Disorder	80
Anxiety	78
Insomnia	78
Abnormal Uterine Bleeding	74
Migraine	59
Psoriasis	58
Endometriosis	57
3 users found this comment helpful.	57
ADHD	55
Asthma, Maintenance	54

WordCloud


miracle
constipation
four
gas
tried
go
linzess
life
chronic
adult
mg
now
ve
bowel
times
worked
days
started
motions
stopped
month
little
per
s
pain
doctor
week
comfortably
amitiza

Problem 1: **Drug Review**

- the number of users who found review useful distribution

count	161297
mean	28.00476
std	36.40374
min	0
25%	6
50%	16
75%	36
max	1291

- Rating distribution

10	31.61%
9	17.07%
1	13.40%
8	11.71%
7	5.86%
5	4.97%
2	4.30%
3	4.04%
6	3.93%
4	3.11%

- Top 10 health conditions

Birth Control	17.97%
Depression	5.69%
Pain	3.86%
Anxiety	3.65%
Acne	3.48%
Bipolar Disorder	2.62%
Insomnia	2.29%
Weight Loss	2.27%
Obesity	2.22%
ADHD	2.11%

Problem 2: Twitter Conversational Bot

Unique Tweet ID

Flag to identify incoming message to rep

Tweet ID of the response to which the current record text is(Following Response)

	tweet_id	author_id	inbound	created_at	text	response_tweet_id	in_response_to_tweet_id
0	1	sprintcare	False	Tue Oct 31 22:10:47 +0000 2017	@115712 I understand. I would like to assist y...	2	3.0
1	2	115712	True	Tue Oct 31 22:11:45 +0000 2017	@sprintcare and how do you propose we do that	NaN	1.0
2	3	115712	True	Tue Oct 31 22:08:27 +0000 2017	@sprintcare I have sent several private messag...	1	4.0
3	4	sprintcare	False	Tue Oct 31 21:54:49 +0000 2017	@115712 Please send us a Private Message so th...	3	5.0
4	5	115712	True	Tue Oct 31 21:49:35 +0000 2017	@sprintcare I did.	4	6.0

Customer Service Agent ID / Customer ID

Tweet Creation Time

Text Response

Tweet ID of the response to the current record text (Preceding Response)

Twitter Agent Description

With 1.5MM total responses sent by Twitter agents and 1.27MM total responses received, following is a table showing agents classified into various businesses.

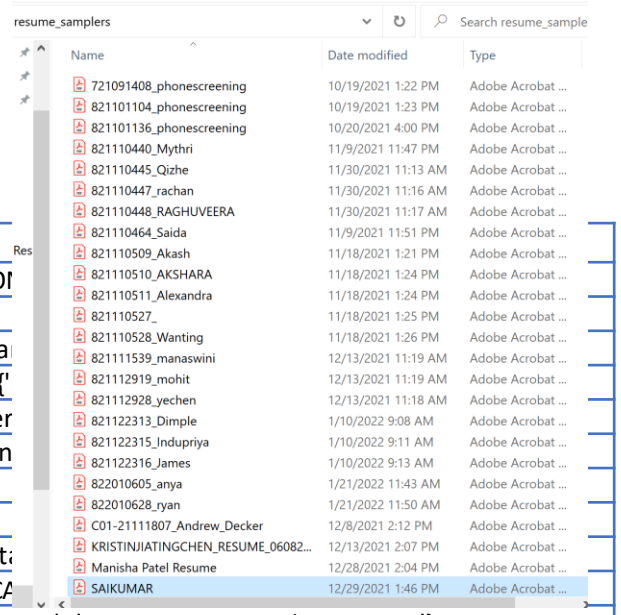
Business Name	Agents
Airline	AirAsia Support, Delta, AmericanAir, SouthwestAir, VirginAtlantic, AlaskaAir, VirginAmerica, JetBlue
Software	AdobeCare, YahooCare, SpotifyCare, AskeBay, DropBoxSupport, AzureSupport, GoDaddyHelp, asksalesforce, GooglePlayMusic, OfficeSupport, TwitterSupport, NortonSupport, SCsupport, AWSSupport, mediatemplehelp, AskTigogh, PandoraSupport, AmazonHelp
Super Market	Morrisons, Tesco, sainsburys, ArgosHelpers, AskTarget, Walmart, AldiUK
Transport	Nationalrailenq, AskLyft, UPSHelp, VirginTrains, GWRHelp, TfL, LondonMidland

**Additional business include Beauty, Education, Clothing, Electronics, Finance, Food, Gaming, Hotel and Rentals*

Most Frequently Used Words By Customers

Agent	Responses Sent	Responses Received	Most Common Words Used By Customer				
			1st	2nd	3rd	4th	5th
Sprintcare	22,381	13,876	phone	store	bill	account	upgrade
ChipotleTweets	18,749	21,593	burrito	order	queso	bowl	chip
XboxSupport	24,557	28,083	update	account	live	buy	console
JetBlue	8,020	9,475	delay	seat	book	bag	cancel
AskPayPal	11,298	10,164	account	money	bank	card	hold

Problem #3: resume parsing



name	phone_num	email	
Karthik	(410)-292-1151	karthikr2194@gmail.com	{'EDUCATIO
Robert		hpundir@umd.edu	
Data Analysis	469-370-9437	tirth2410@gmail.com	{'University', 'Ha
Mythri	(281) 725-7080	mythripartha8@gmail.com	{'
Qizhe		ziyaotingyu@gmail.com	{'EDUCATION Univer
Stamford	(475) 685 0166	rachan_vamsi.bhooshi@uconn.edu	{'Un
Github	240-713-8296	rmadireddy1@student.gsu.edu	
Data Science	443-833-6344	saidam1@umbc.edu	
Akash Patel	+1 4845387112	adp178@scarletmail.rutgers.edu	{'St
Analytics		akshara@ou.edu	{'EDUCA
Alexandra	630.818.6275	manetas.alexandra@gmail.com	{'DePaul University', 'University Research Institutes'}
Pranav Premdas Gulghane	-4695140739	pranavpremdasgulghane@gmail.com	{'University'}
Lu Berkeley	(559) 387-0880	luwinnie12@gmail.com	{'University'}
Khoury College	(617) 818-4953	nagaraj.m@northeastern.edu	{'National University', 'D Institute', 'REVA University'}
Data Management	18572077337	mohitmanjaria55333@gmail.com	{'UNIVERSITY Master', 'UNIVERSITY Masters'}
Yuchen	949-413-2863	yuchen724@ucla.edu	{'EDUCATION University', 'University'}
Dimple		dimple8997@gmail.com	set()
Kompi S		indupriyakompi@gmail.com	{'GPA Northeastern University'}
James	(706) 305-6369	james.domingo@gatech.edu	{'University', 'EDUCATION Georgia Institute'}
Machine Learning	805 453 1532	anyampatel@gmail.com	set()
Ryan Martin Goodwin LinkedIn Github	(910) 547-7027	rgoodwin1997@gmail.com	{'Science University'}
Andrew	(443)-742-3540	eckan01@gettysburg.edu	set()
Git	(202) 212 -9607	jiatingchen0107@gmail.com	{'UNIVERSITY Master'}
Data Scientist	774-994-4106	manishap2690@gmail.com	{'State University'}
Python	12678819188	sgheereddy@gmail.com	{'EDUCATION University', 'Jawaharlal Nehru Technological University'}
Senthil	-8553247568	msnathan55@yahoo.com	{'University', 'Toyohashi University', 'PSG College'}

Problem #3: resume parsing (variants in phone numbers)

name	phone_num	email	education
Robert		hpundir@umd.edu	{'University'}
Dimple		dimple8997@gmail.com	set()
Senthil	-8553247568	msnathan55@yahoo.com	{'University', 'Toyohashi University', 'PSG College'}

HARSH PUNDIR
3425 Tulane Dr. Hyattsville, MD 20783 240.423.5453
hpundir@umd.edu | www.linkedin.com/in/hpundir | https://github.com/HARSHPUNDIR

EDUCATION

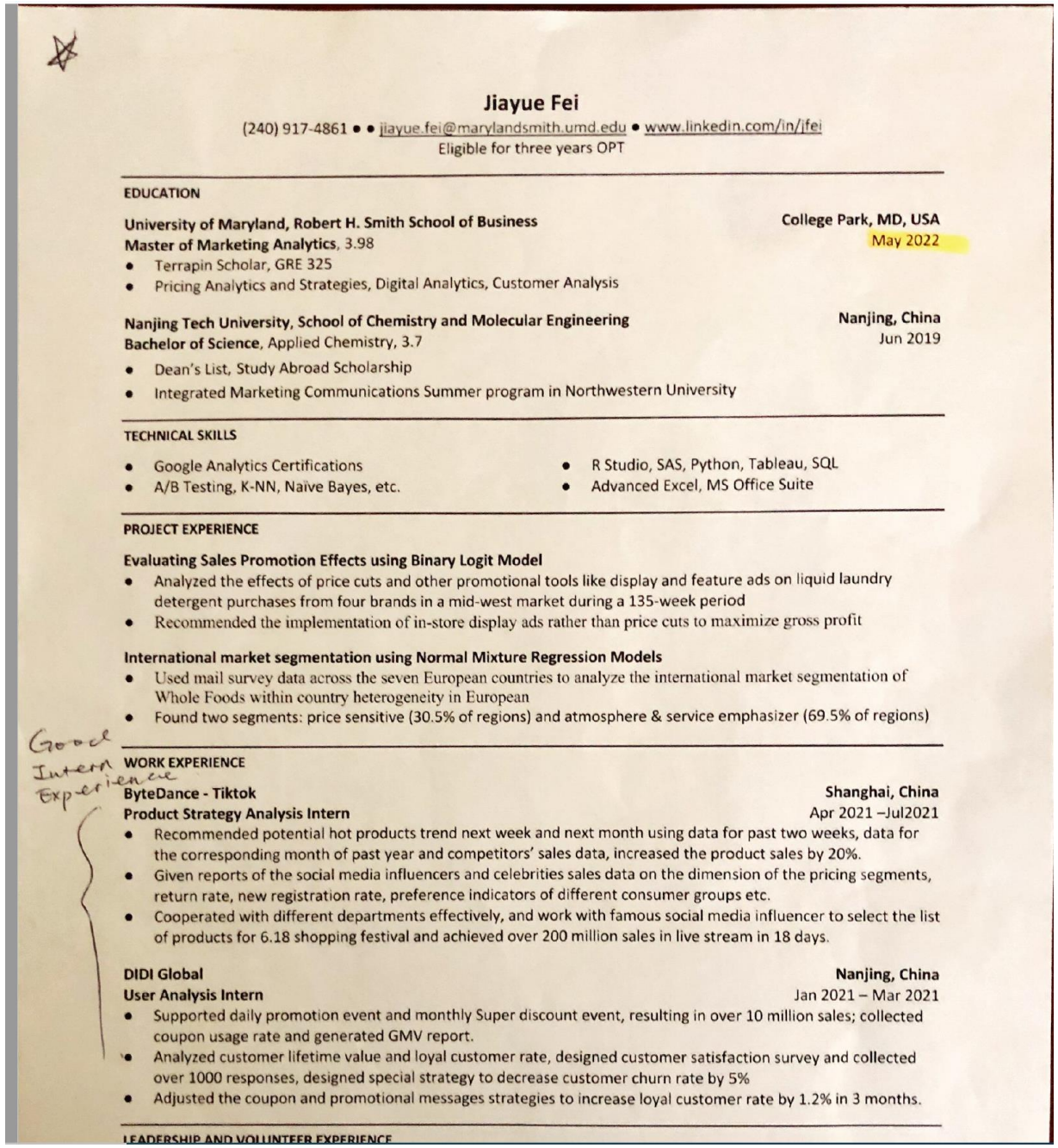
Robert H. Smith School of Business, University of Maryland, College Park, MD August 2020 - Present
Master of Science in Business Analytics, Focus Area: Data Science GPA: 4.0
- Relevant Coursework – DBMS, Python, Data Models & Decisions, Data Mining, Big Data & AI, Data Visualization.

Dimple Mehra in
San Jose, CA 95035

Email : dimple8997@gmail.com
Mobile : +1-(312-358-1359)

Senthil Nathan M
Bangalore, Karnataka, India • +91-8553247659 • msnathan55@yahoo.com • https://www.linkedin.com/in/senthil-nathan-murugappan-585a4967/

Problem #3: resume parsing



Jiayue Fei

(240) 917-4861 • • jiayue.fei@marylandsmith.umd.edu • www.linkedin.com/in/jfei

Eligible for three years OPT

EDUCATION

University of Maryland, Robert H. Smith School of Business — **College Park, MD, USA**

Master of Marketing Analytics, 3.98 — May 2022

- Terrapin Scholar, GRE 325
- Pricing Analytics and Strategies, Digital Analytics, Customer Analysis

Nanjing Tech University, School of Chemistry and Molecular Engineering — **Nanjing, China**

Bachelor of Science, Applied Chemistry, 3.7 — Jun 2019

- Dean's List, Study Abroad Scholarship
- Integrated Marketing Communications Summer program in Northwestern University

TECHNICAL SKILLS

- Google Analytics Certifications
- A/B Testing, K-NN, Naïve Bayes, etc.
- R Studio, SAS, Python, Tableau, SQL
- Advanced Excel, MS Office Suite

PROJECT EXPERIENCE

Evaluating Sales Promotion Effects using Binary Logit Model

- Analyzed the effects of price cuts and other promotional tools like display and feature ads on liquid laundry detergent purchases from four brands in a mid-west market during a 135-week period
- Recommended the implementation of in-store display ads rather than price cuts to maximize gross profit

International market segmentation using Normal Mixture Regression Models

- Used mail survey data across the seven European countries to analyze the international market segmentation of Whole Foods within country heterogeneity in European
- Found two segments: price sensitive (30.5% of regions) and atmosphere & service emphasizer (69.5% of regions)

Good Intern Experience

WORK EXPERIENCE

ByteDance - Tiktok — **Shanghai, China**

Product Strategy Analysis Intern — Apr 2021 –Jul2021

- Recommended potential hot products trend next week and next month using data for past two weeks, data for the corresponding month of past year and competitors' sales data, increased the product sales by 20%.
- Given reports of the social media influencers and celebrities sales data on the dimension of the pricing segments, return rate, new registration rate, preference indicators of different consumer groups etc.
- Cooperated with different departments effectively, and work with famous social media influencer to select the list of products for 6.18 shopping festival and achieved over 200 million sales in live stream in 18 days.

DIDI Global — **Nanjing, China**

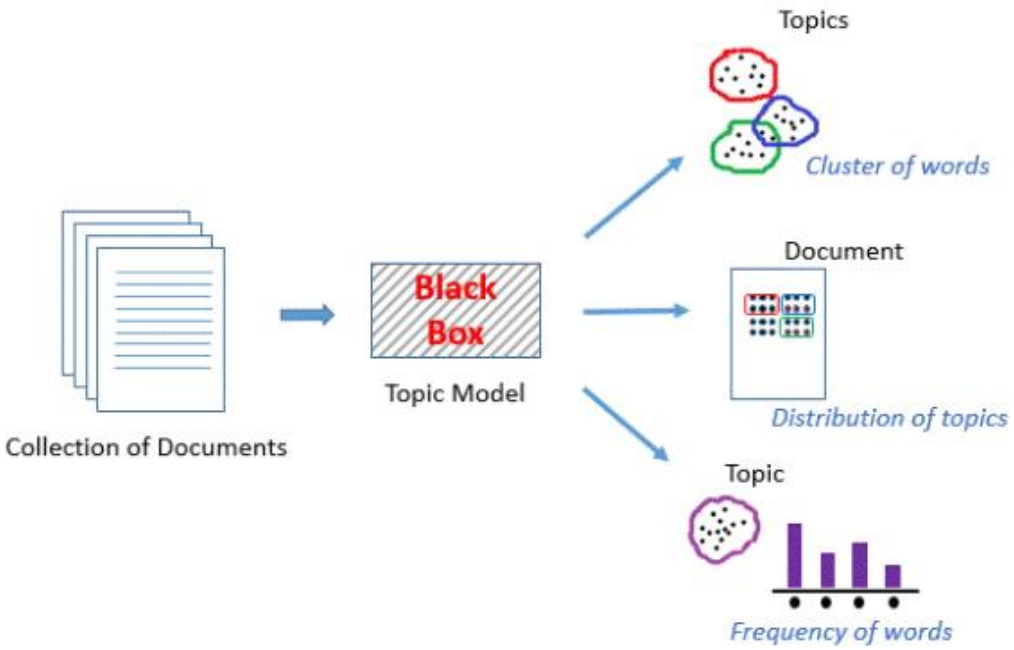
User Analysis Intern — Jan 2021 – Mar 2021

- Supported daily promotion event and monthly Super discount event, resulting in over 10 million sales; collected coupon usage rate and generated GMV report.
- Analyzed customer lifetime value and loyal customer rate, designed customer satisfaction survey and collected over 1000 responses, designed special strategy to decrease customer churn rate by 5%
- Adjusted the coupon and promotional messages strategies to increase loyal customer rate by 1.2% in 3 months.

LEADERSHIP AND VOLUNTEER EXPERIENCE

Topic Modelling

- Unsupervised ML technique to discover a set of topics that best segregate a corpus of documents/ bag of words
- Tries to figure out which topics are present in the documents of the corpus and how strong is that presence
- Eg : Latent Dirichlet Allocation *(Gensim Library in Python)*



Latent Dirichlet Allocation

Two key assumptions:

- Documents are a distribution of topics
- Topics are a distribution of words

Document Term Matrix (DTM)

	W1	W2	W3	W4	W5	W6	W7	W8
D1	0	1	1	0	1	1	0	1
D2	1	1	1	1	0	1	1	0
D3	1	0	0	0	1	0	0	1
D4	1	1	0	1	0	0	1	0
D5	0	1	0	1	0	0	1	0

Shape: 5 * 8

Document Topic Matrix

	K1	K2	K3	K4	K5	K6
D1	1	0	0	0	0	0
D2	0	1	0	0	1	1
D3	1	1	0	0	0	0
D4	1	0	0	1	0	1
D5	0	0	1	1	0	0

Shape: 5 * 6

Topic Word Matrix

	W1	W2	W3	W4	W5	W6	W7	W8
K1	0	1	1	0	1	0	1	0
K2	1	1	1	1	0	1	1	1
K3	1	0	0	0	0	1	0	0
K4	1	1	0	1	1	0	0	1
K5	0	0	1	1	0	1	1	1
K6	1	0	1	1	1	0	0	1

Shape: 6 * 8

How the Algorithm functions?

- For each document in the corpus, a topic word distribution is generated
- Each document will be randomly assigned to topics in the first iteration
- Post the first iteration, LDA provides per document topic distribution and per topic word distribution
- End goal is to optimize these two output matrices by updating the topic for each word in each document keeping the topics assigned to other words in the document constant
- This update is done by calculating the following two probabilities:
  1) ***Proportion (Topic k / Document D)***
  2) ***Proportion (word w / Topic k)***
- Based on the product of these two probabilities, LDA assigns a new topic to the word