Data Case Study

Hank Lo

Business Problem

How to establish an approach to recommend a risk score with intentions to help mitigate the risk of fraudulent events while not at expense of harming legitimate members experience

Assumptions

- The model would be used in enrollment
- Financial cost = dispute costs ($10 each dispute) + acquisition costs
- The risk scores were generated by a binary classification model developed by the Data Science team
- Number of disputes serve as a proxy for fraud
- The default cut-off value is 50% which would be used as the baseline to measure improvement on recommended threshold

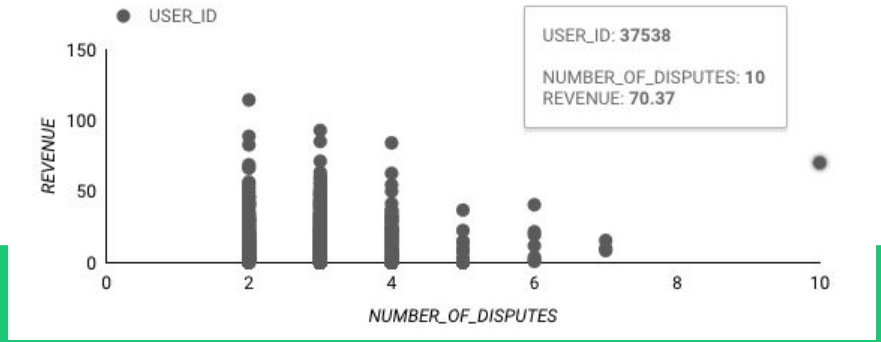
What the data looks like?

1. There are ***72.4K*** distinct USER_IDs in the raw data set. Seven accounts were excluded from the data set due to invalid rick score (outside the range from 0 to 1)
2. No missing values or duplicates were identified in the data set
3. No obvious outlier was found except for USER_ID: 37538 who had ten disputes on the account
4. Feature Engineering
   a. **Total Cost** = dispute costs ($10 each dispute) + acquisition costs
   b. **Net Profit** = Revenue - Total Cost

Metrics to watch

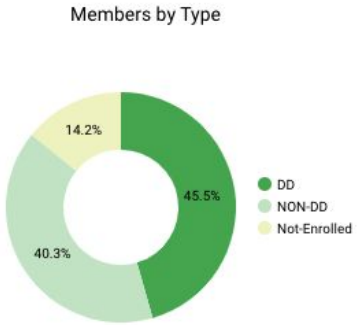
MEMBER_COUNT	REVENUE	NUMBER_OF_DISPUTES	FINANCIAL_COST
72.4K	1.9M	25.3K	311.5K

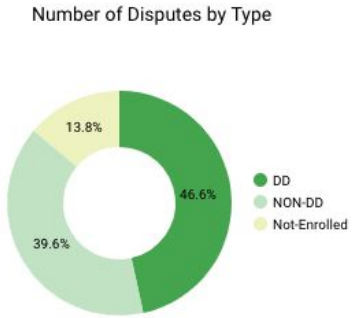
```
-- Create a new table that involves total financial cost and net profit margin
-- financial_total_cost = # of disputes * 10 + acquisition costs (varies depending on channel)
-- Net profit revenue - financial_total_cost

drop table if exists `gentle-complex-312720.chime.member_disputes_cost`;
create or replace table `gentle-complex-312720.chime.member_disputes_cleaned` AS
with temp_cost as (
select USER_ID,
       CHANNEL,
       ENROLLED,
       REVENUE,
       NUMBER_OF_DISPUTES,
       RISK_SCORE,
       ifnull(TYPE, 'Not-Enrolled') as TYPE,
       case when CHANNEL ='SEARCH' then 1.00 + NUMBER_OF_DISPUTES * 10
            when CHANNEL ='SOCIAL' then 1.25 + NUMBER_OF_DISPUTES * 10
            when CHANNEL ='ORGANIC' then 0.5 + NUMBER_OF_DISPUTES * 10
            when CHANNEL ='PARTNER' then 1.5 + NUMBER_OF_DISPUTES * 10
            else NUMBER_OF_DISPUTES * 10 end as TOTAL_COST
from `gentle-complex-312720.chime.member_disputes`
where RISK_SCORE between 0 and 1 -- remove users with invalid risk scores
)
select *,
       REVENUE - TOTAL_COST as NET_PROFIT
from temp_cost
;
```



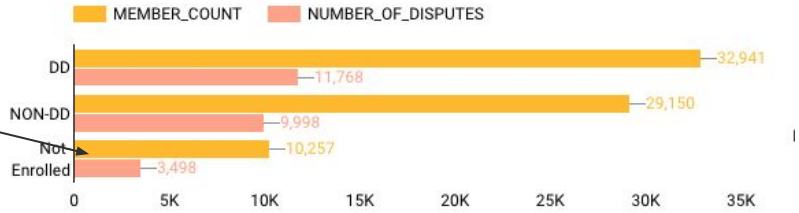
Members and Disputes breakdown

DD type accounts for ***45.5%*** of the total users and ***46.6%*** of the total number of disputes.

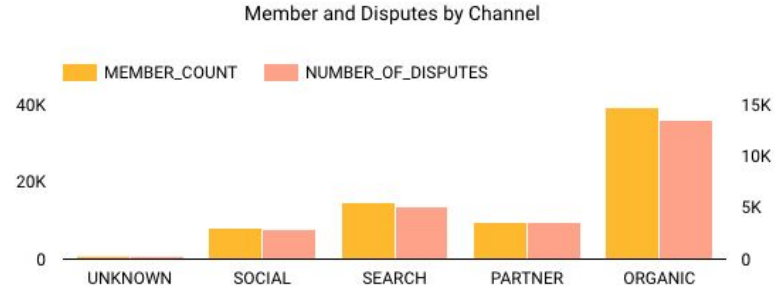




28% of 10,257 non-enrolled users (~2,890 users) have more than one dispute



The majority of members were acquired through ORGANIC channel followed by SEARCH and SOCIAL

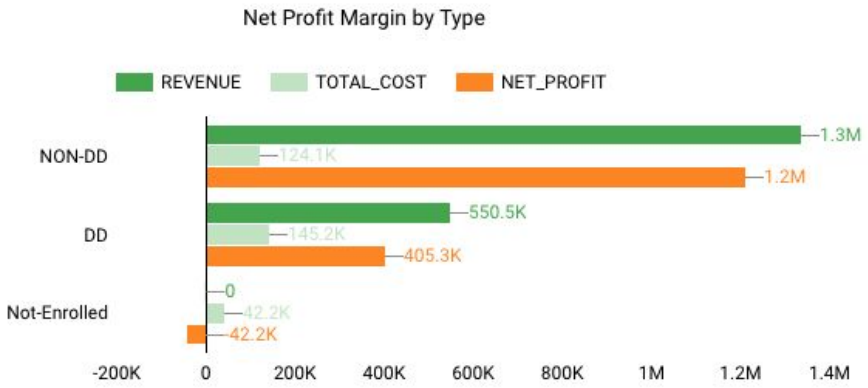


Revenue, Cost and Net Profit Breakdown

~71% of revenue generated by Non-DD members whereas DD members account for ***~30%*** of total revenue. .

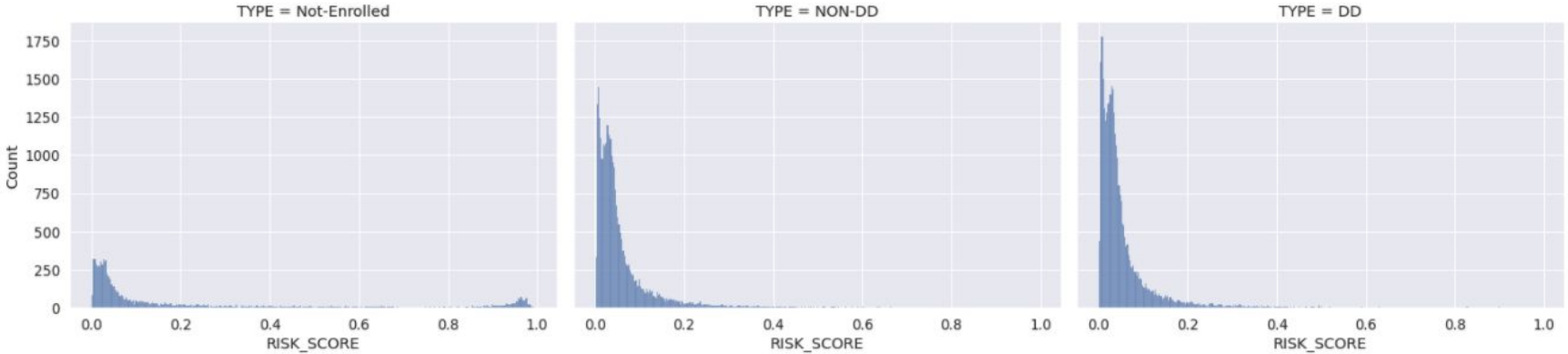
However, the majority of cost comes from DD members which account for ***~47%*** of total cost

	TYPE	REVENUE ▼	PERCENT	TOTAL_COST	PERCENT	AVG_NET_PROFIT
1.	NON-DD	1,337,578.14	70.84%	124,106.5	39.85%	41.63
2.	DD	550,507.07	29.16%	145,186.5	46.62%	12.3
3.	Not-Enrolled	0	0%	42,162.25	13.54%	-4.11



Non-DD Type has highest Average Net Profit (***$41.63*** per user) compared to DD members (***$12.3*** per user). Not-enrolled type has a negative net profit margin given no revenue generated by users who hasn't enrolled.

Risk score follows a right-skewed distribution across types



Non-Enrolled type has the highest median and average risk score across TYPE, which suggests the threshold should be determined based on cohorts with aims to minimize overall costs.

	RISK_SCORE				
	mean	median	std	q1	q3
TYPE					
DD	0.062873	0.0341	0.099793	0.0179	0.0624
NON-DD	0.068778	0.0367	0.106119	0.0192	0.0691
Not-Enrolled	0.239203	0.0659	0.315375	0.0265	0.3486

The best cut-off minimizes cost of losing legitimate members and maximizes average net profit margin

1. ***Average net profit margin*** and ***the number of member loss*** (users with fraud flag = 1) were chosen to be the success metrics to optimize
2. ***0.5*** was used as baseline threshold to evaluated the performance improvement on new threshold
3. Business metrics (net profit, average net profit margin, number of retaining/ loss members) were calculated for each thresholds
4. For DD, ***0.55*** is recommended given it has the highest avg_net_profit_margin and decrease the flagged DD members by 17%
5. For NON-DD, ***0.45*** is recommended as it leads to a highest avg_net_profit_margin with a potential loss of 550 members
6. Given no revenue generated by Not-Enrolled users, ***0.4*** threshold was recommended with intention to minimize the acquisition costs and processing loss from disputes

DD Member

	TYPE	threshold	net_profit	member_loss_cnt	member_cnt	avg_net_profit_margin
8	DD	0.40	396240.15	633	32308	12.2645
9	DD	0.45	398533.61	512	32429	12.2894
10	DD	0.50	400342.71	409	32532	12.3061
11	DD	0.55	401693.25	339	32602	12.3211
12	DD	0.60	402060.31	285	32656	12.312
13	DD	0.65	402649.95	233	32708	12.3104
14	DD	0.70	403004.41	184	32757	12.3028

Non-DD Member

	TYPE	threshold	net_profit	member_loss_cnt	member_cnt	avg_net_profit_margin
28	NON-DD	0.40	1207642.36	692	28458	42.436
29	NON-DD	0.45	1208532.80	550	28600	42.2564
30	NON-DD	0.50	1209647.60	445	28705	42.1407
31	NON-DD	0.55	1210140.88	357	28793	42.029
32	NON-DD	0.60	1211149.06	284	28866	41.9576
33	NON-DD	0.65	1211634.30	218	28932	41.8787
34	NON-DD	0.70	1212141.84	163	28987	41.8167

Non-Enrolled User

	TYPE	threshold	net_profit	member_loss_cnt	member_cnt	avg_net_profit_margin
48	Not-Enrolled	0.40	-23093.25	2356	7901	-2.92283
49	Not-Enrolled	0.45	-24513.25	2172	8085	-3.03194
50	Not-Enrolled	0.50	-25857.25	1998	8259	-3.1308
51	Not-Enrolled	0.55	-27109.75	1846	8411	-3.22313
52	Not-Enrolled	0.60	-28302.75	1700	8557	-3.30756
53	Not-Enrolled	0.65	-29397.25	1546	8711	-3.37473
54	Not-Enrolled	0.70	-30317.25	1431	8826	-3.43499

Conclusion and Next Step

1. Recommended thresholds on the risk score for each type
    a. DD Member: 0.55
    b. Non-DD Member: 0.45
    c. Non-Enrolled User: 0.40
2. Further investigation is needed for Non-Enrolled users who has more than one dispute
3. Integrate Model Evaluation Metrics (Precision, Recall, F1-score) with ground truth label in order to strike a balance model performance with business objectives
4. Incorporate additional features
    a. Demographic Feature (Region, language, Platform (Desktop/ Mobile Web/ App)
    b. Member Attributes (Number of transactions, Account Tenure, IP Address, Number of Accounts..etc)
    c. Timeframe (WoW, MoM, QoQ)
    d. Guardrail Metrics

Thank you