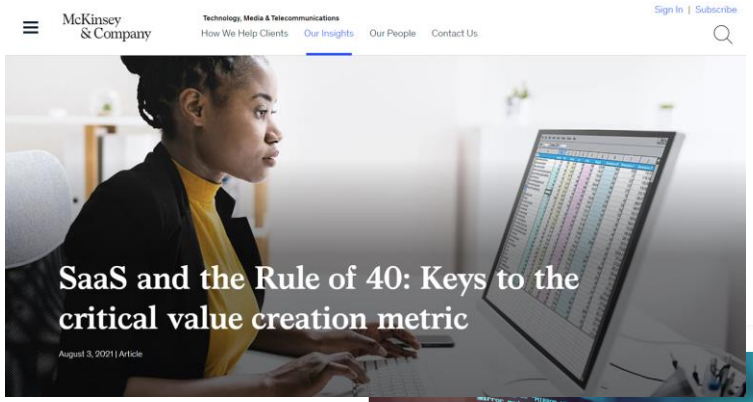
Can We Predict Churn Rate?: The Lifeline of Startups

森島 晴也 Hal Morishima

TABLE OF CONTENTS



Hacking Software's Rule of 40

Software companies need to balance growth and profitability to create lasting value.

By Thierry Depeyrot and Simon Heap

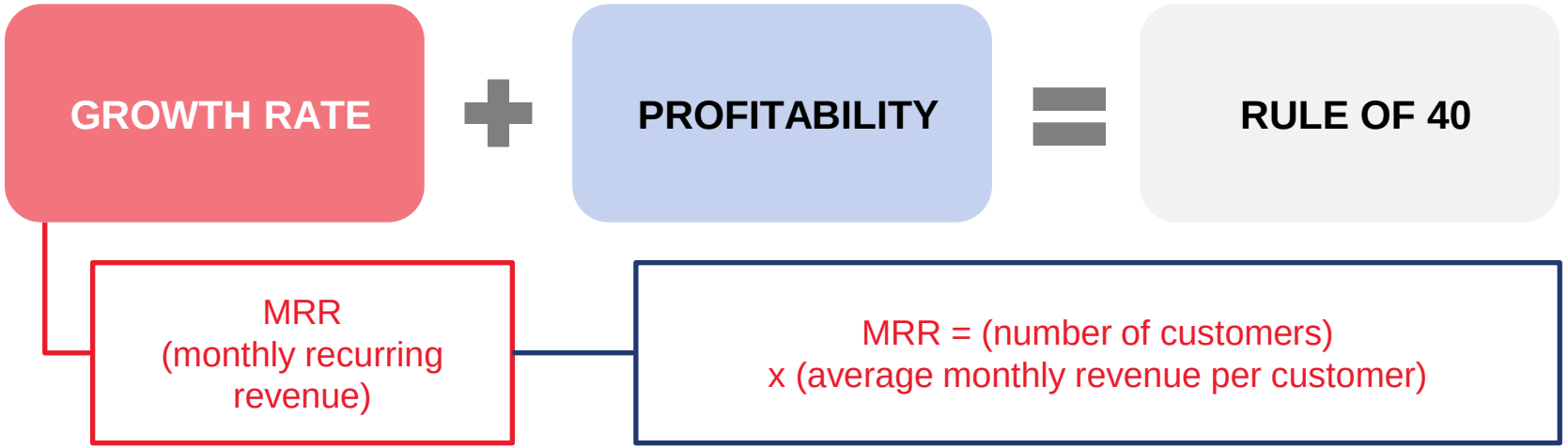
BAIN & COMPANY

What is The Rule of 40 in SaaS and Why it Matters

01.INTRODUCTION

02. Importance of Metric “RULE of 40”

The "Rule of 40" is an important financial metric for startups, calculated by summing the company's revenue growth rate and its profitability margin, and MRR serves as an indicator of the growth rate.



02. Understanding MRR and Churn-rate

MRR is not constituted by a single factor; instead it is observed from various perspectives such as new business, existing customers, and churn.

MRR
(monthly recurring revenue)

- New MRR: total MRR gained from new customers
- Churn MRR: total MRR lost from customers who have canceled their subscription
- Expansion MRR: total MRR gained from existing customers (e.g. customers who upgrade their plan)
- Contraction MRR: total MRR lost from existing customers who have downgraded their plan
- Reactivations MRR: total MRR gained from old customers who have reactivated their subscription.

[(#) Total customers churned this time period /

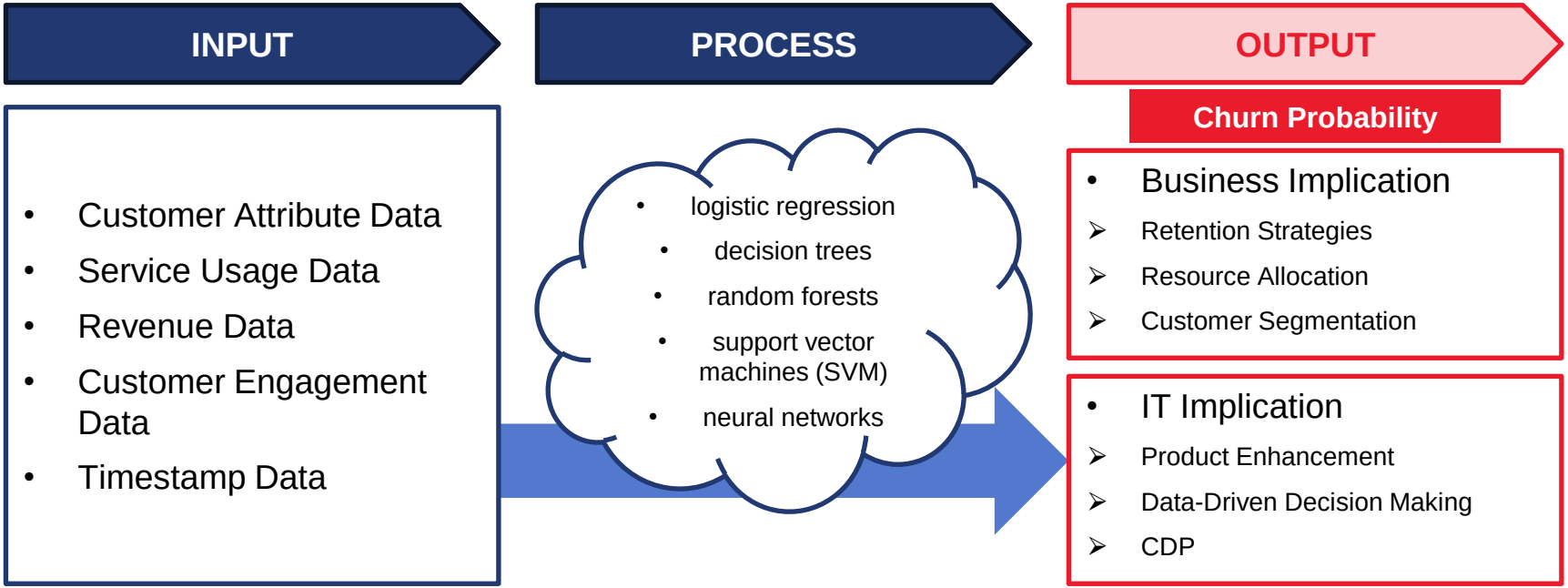
(#) Total customers at the start of this time period]

X 100 = (%) Customer Churn Rate

- KKBOX is a leading music streaming service based in Taiwan
  ≒ Spotify
- Monthly subscription-base app

03. IMPACT

By inputting data related to customer and product usage into the model, we can analyze and calculate Churn Probability. This valuable insight can be leveraged to derive implications for both business and IT aspects.



04. DATASET INTRODUCTION

The data is split into four CSV files and requires merging. The dataset is somewhat organized and manageable, yet it contains a substantial amount and variety of data.

members_v3.csv

- Basic customer data.
- Contains 6769473 number of rows and 6 features
- msno(unique id), city, bd(age), gender, registered_via and registration_init_time.

train_v2.csv

- ID and Churn status
- 970960 number of rows
- msno(unique id) and is_churn(class labels).

transactions_v2.csv

- User transactions data
- 1431009 number of rows and 9 features
- msno(unique id), payment_method_id, payment_plan_days, plan_list_price, actual_amount_paid, is_auto_renew, transaction_date, membership_expire_ date and is_cancel

transactions_v2.csv

- User engagement
- 18396362 number of rows and 9 features
- msno(unique id), date, num_25, num_50, num_75, num_985, num_100, num_unq and total_secs.

04. DATASET INTRODUCTION

By inputting data related to customer and product usage into the model, we can analyze and calculate Churn Probability. This valuable insight can be leveraged to derive implications for both business and IT aspects.

No.	Category	Column Name	Description	Concerns
1	User Identification	msno	Unique ID for each user.	n/a
2	Target Variable	is_churn	Indicates whether a user has churned or not.	n/a
3	Transaction Info	payment_method_id	The method used for payment.	n/a
4	Transaction Info	payment_plan_days	The length of the membership plan.	Need to check the unrealistic value 0. Unsubscribed?
5	Transaction Info	plan_list_price	The list price of the plan.	Need to check the unrealistic value 2000.
6	Transaction Info	actual_amount_paid	The amount actually paid for the plan.	Need to check the unrealistic value 0. Unpaid?
7	Transaction Info	is_auto_renew	Indicates whether the plan renews automatically.	n/a
8	Transaction Info	is_cancel	Indicates whether the user canceled the membership in a transaction.	n/a
9	User Engagement	date	Date of user log.	n/a
10	User Engagement	num_25, num_50, num_75, num_985, num_100	Songs played less than XX% of the song length.	n/a
11	User Engagement	num_unq	Number of unique songs played.	n/a
12	User Engagement	total_secs	Total seconds of music played.	n/a
13	Demographic Info	city	Same as the column names.	n/a
14	Demographic Info	bd (age)	Same as the column names.	Outlier values ranging from -7000 to 2015. Insert 0 or average bd to these numbers?
15	Demographic Info	gender	Same as the column names.	65% of data are null-value. Drop the column or change the column name(unknown) and leve?
16	Demographic Info	registered_via	Registration method.	Outlier value -1.00
17	Demographic Info	registration_init_time	The time of registration.	n/a

05. NEXT STEP

Promptly address outliers and missing values while improving data visualization. Then, proceed to select models with a hypothesis-driven approach.

1. Data Prep

Identification and handling of

- Extensive missing data
- Null values
- Outliers
- Dummy variables

Further visualization

- Consider hypothesis for modeling

2. Modeling

- Model list-up and evaluation
- Model selection

05. NEXT STEPS

Given the large data size and diverse data types in the dataset, a model that can accommodate various variables while reducing runtime would be preferable.

Possible Solutions

- **Data Size: BIG**
  18,396,362 rows
- **Diverse Factor**
  3-4 major ctg
- **Time Stamp**
- **Individual ID**

- Logistic Regression: Simple, Runtime
- Decision Trees: Multiple factors
- Random Forest: Less overfitting risk compared to decision trees?
- Neural Networks: non-linear relationships/ effective for large datasets