# Can We Predict Churn Rate?:

## The Lifeline of Startups

2023 Sep 5th

森島 晴也 Hal Morishima

# TABLE OF CONTENTS

# 01. Overview: Recap of my Capstone Project

Here is a recap of my project:

| | |
|---|---|
| **Overview** | • To predict whether a user will discontinue their subscription after it expires.<br>• Users to make a new subscription within 30 days after their current membership expiration date. |
| **Solution** | • Machine learning: Classification<br>• Logistics Regression, SGD Classifier, XGBoost, Decision Tree, Random Forest |
| **Potential Impact** | • Better business performance: Retaining existing customers usually comes at a lower cost than acquiring new ones.<br>• Implication to IT side: Product Enhancement/ Data-Driven Decision Making/CDP |

# 02. Data Set and Preprocessing

The project involves data integration, which is quite complicated. There are also a significant number of outliers and null values present in the data.

**1: Dataset & EDA**

| members | train | transactions | engagement |
|---|---|---|---|
| •Basic customer data.<br>•6M<br>•msno(unique id), city, bd(age), gender | •ID and Churn status<br>•970K<br>•msno, is_churn(class labels). | •User transactions data<br>•1.4M<br>•msno(unique id), payment_method_id, plan | •User engagement<br>•18M<br>•msno(unique id), date, num_25, num_50 |

**2: Preparation**

**2-1: Initial Data Cleaning**
Aggregation

**2-2: Merger**
Necessary to make ID unique and consolidate all columns.

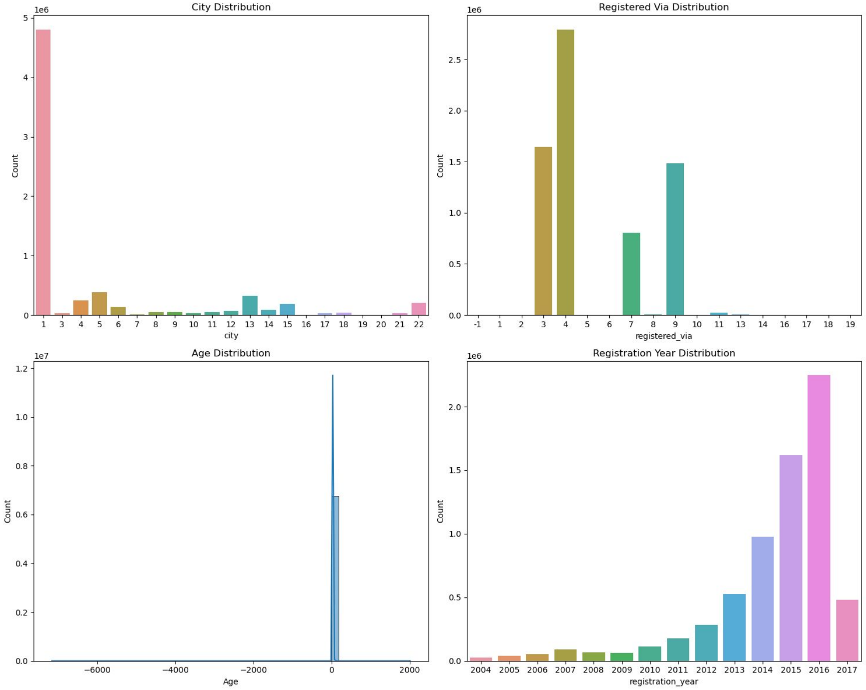
**3: Preprocess**

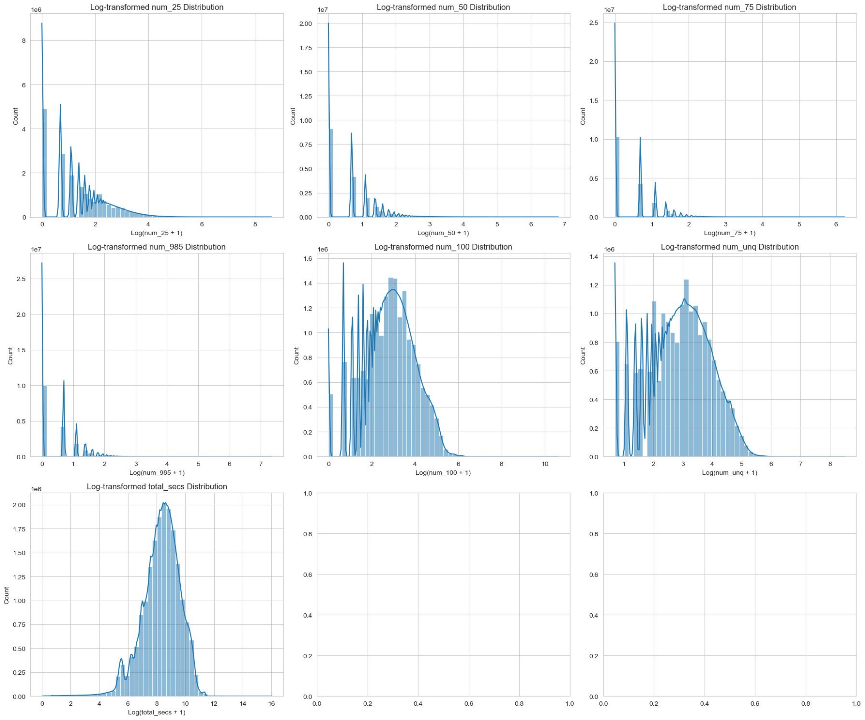
**3-1: After Pre-Processing**
Outlier, null-value(Imputation, Drop)
Data type conversion, One-hot encoding

**3-2: Feature Engineering**
8 new columns about order value, engagement, lifespan

# 03. EDA Summary

Through the merging process, the presence of Null values, outliers, and data skewness has become evident. It seems that around 30 instances need to be addressed and corrected.



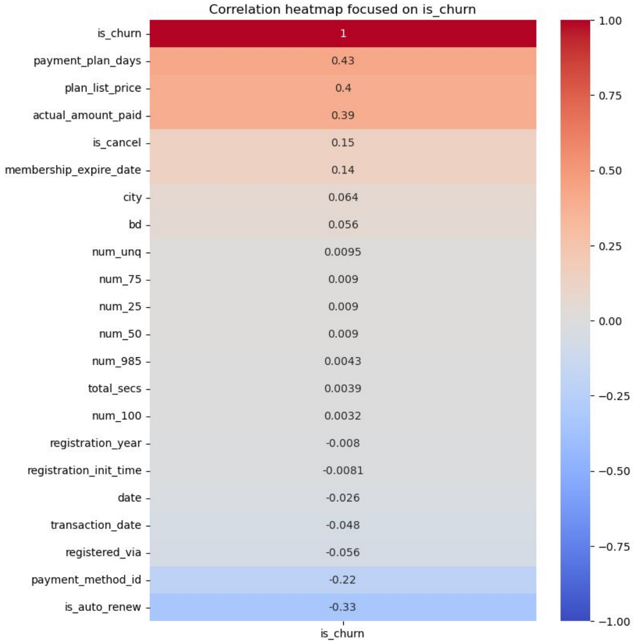
# 04. Feature Engineering

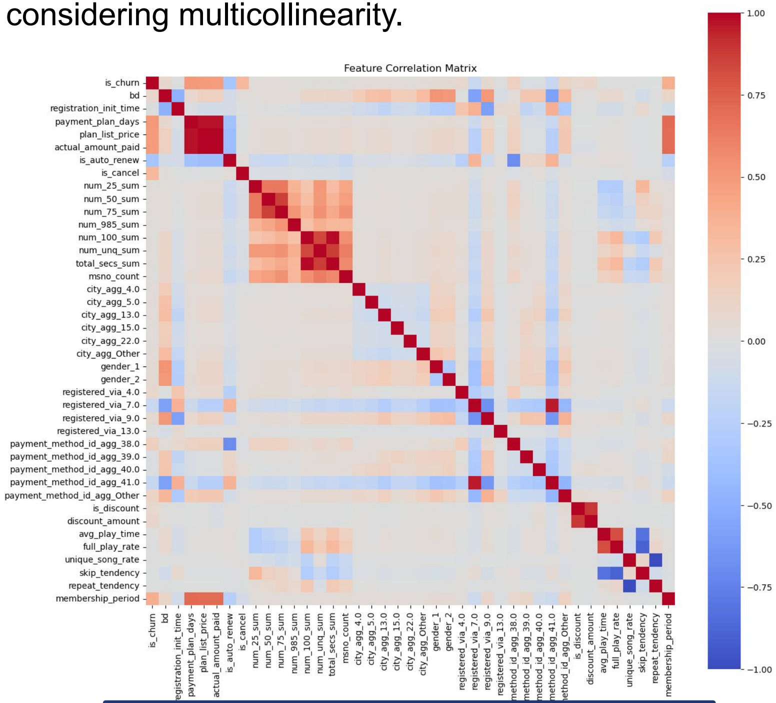
Created eight features related to customer Order Value, Engagement, and Lifespan, which are components of Customer Lifetime Value (LTV).

| Category | Feature | Description |
|---|---|---|
| Order Value | is_discount | Binary feature indicating whether a discount was applied on the transaction (1 if actual_amount_paid < plan_list_price, otherwise 0). |
| | discount_amount | Calculated discount amount for each transaction as the difference between plan_list_price and actual_amount_paid. |
| Engagement | Average Play Time per Song | Average time a user spends listening to a song, calculated by dividing total seconds played by total number of songs played (sum of 'num_25', 'num_50', 'num_75', 'num_985', and 'num_100'). |
| | Full Play Rate | Proportion of songs played over 98.5% of their length ('num_100') to the total number of songs played. |
| | Unique Song Play Rate | Proportion of unique songs played ('num_unq') to the total number of songs played. |
| | Skip Tendency | Tendency of a user to skip songs before they reach 25% of their length, calculated as the ratio of 'num_25' to the total number of songs played. |
| | Repeat Tendency | Tendency of a user to repeat songs, calculated as the difference between total number of songs played and the number of unique songs played ('num_unq'), divided by total number of songs played. |
| Life Span | Membership Period in Days | Derived feature representing the membership period in days by calculating the difference between 'membership_expire_date' and 'transaction_date'. |

# 05. Feature Selection

Identified and decided to drop features with low coefficients or considering multicollinearity.




•Dropped columns with low relevance or
•Replaced them with other columns through feature engineering."

•Removed columns with high priority by setting a correlation threshold of 0.8

# 05. Feature Selection

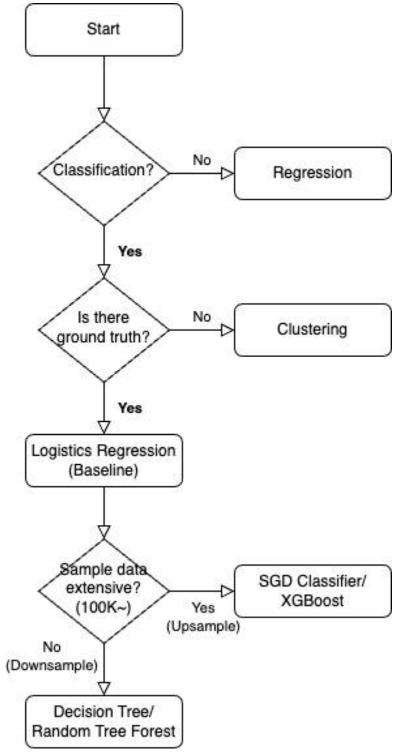
Including the dummy-encoded features and the additional ones, we ended up with a total of 31 columns.

Data columns (total 32 columns):

| # | Column | Non-Null Count | Dtype |
|---|---|---|---|
| 0 | msno | 725722 non-null | object |
| 1 | is_churn | 725722 non-null | int64 |
| 2 | bd | 725722 non-null | int64 |
| 3 | registration_init_time | 725722 non-null | float64 |
| 4 | payment_plan_days | 725722 non-null | int64 |
| 5 | actual_amount_paid | 725722 non-null | float64 |
| 6 | is_auto_renew | 725722 non-null | int64 |
| 7 | is_cancel | 725722 non-null | int64 |
| 8 | msno_count | 725722 non-null | float64 |
| 9 | city_agg_4.0 | 725722 non-null | int64 |
| 10 | city_agg_5.0 | 725722 non-null | int64 |
| 11 | city_agg_13.0 | 725722 non-null | int64 |
| 12 | city_agg_15.0 | 725722 non-null | int64 |
| 13 | city_agg_22.0 | 725722 non-null | int64 |
| 14 | city_agg_Other | 725722 non-null | int64 |
| 15 | gender_1 | 725722 non-null | int64 |
| 16 | gender_2 | 725722 non-null | int64 |
| 17 | registered_via_4.0 | 725722 non-null | int64 |
| 18 | registered_via_9.0 | 725722 non-null | int64 |
| 19 | registered_via_13.0 | 725722 non-null | int64 |
| 20 | payment_method_id_agg_38.0 | 725722 non-null | int64 |
| 21 | payment_method_id_agg_39.0 | 725722 non-null | int64 |
| 22 | payment_method_id_agg_40.0 | 725722 non-null | int64 |
| 23 | payment_method_id_agg_41.0 | 725722 non-null | int64 |
| 24 | payment_method_id_agg_Other | 725722 non-null | int64 |
| 25 | is_discount | 725722 non-null | int64 |
| 26 | discount_amount | 725722 non-null | float64 |
| 27 | avg_play_time | 725722 non-null | float64 |
| 28 | full_play_rate | 725722 non-null | float64 |
| 29 | skip_tendency | 725722 non-null | float64 |
| 30 | repeat_tendency | 725722 non-null | float64 |
| 31 | membership_period | 725722 non-null | int64 |

# 06. Modeling

Selected a model that takes into account the characteristics of the dataset.




**Unbalanced Class**

- Data is skewed
  no churn: 679,119
  churn: 46,603

No excessive data manipulation

- Decision Tree

**Extensive Dataset**

- 725,722 data entries

Tolerant of large-scale data

- SGD Classifier
- XGBoost

**Overfitting**

- Again, unbalanced data…

Mitigate overfitting

- Random Tree Classifier

# 06. Modeling

Selected and compared scores that are aligned with the objectives of this study.

| Model | ROC-AUC | Recall | F1 Score |
|---|---|---|---|
| Logistic Regression | 0.76 | 0.52 | 0.66 |
| SGD Classifier | 0.82 | 0.74 | 0.48 |
| XGBoost | 0.79 | 0.74 | 0.26 |
| Decision Tree | 0.84 | 0.70 | 0.65 |
| Random Forest | 0.96 | 0.89 | 0.89 |

Best F1 Score but low recall

Best ROC-AUC score

Overfitting?

※F1 scores are for Class1(Churning)

# 07. Implication

From logistic regression, we didn't obtain particularly valuable information about the determining factors of churn.

**Coef: Logistics Regression**

```
Coefficients in descending order of influence:
payment_plan_days: 1.9736433823147177
is_auto_renew: -0.6538205588134377
is_cancel: 0.5705074495066044
membership_period: 0.29058935671740005
registration_init_time: -0.21272637738007666
discount_amount: 0.20373647816209695
payment_method_id_agg_Other: -0.17420849910405678
msno_count: -0.12510542340170558
registered_via_4.0: 0.09757443629735633
registered_via_9.0: -0.09213480319000811
registered_via_13.0: 0.07091278008556431
city_agg_Other: 0.0653510651768192
city_agg_5.0: 0.06113470171760278
```

**payment_plan_days**

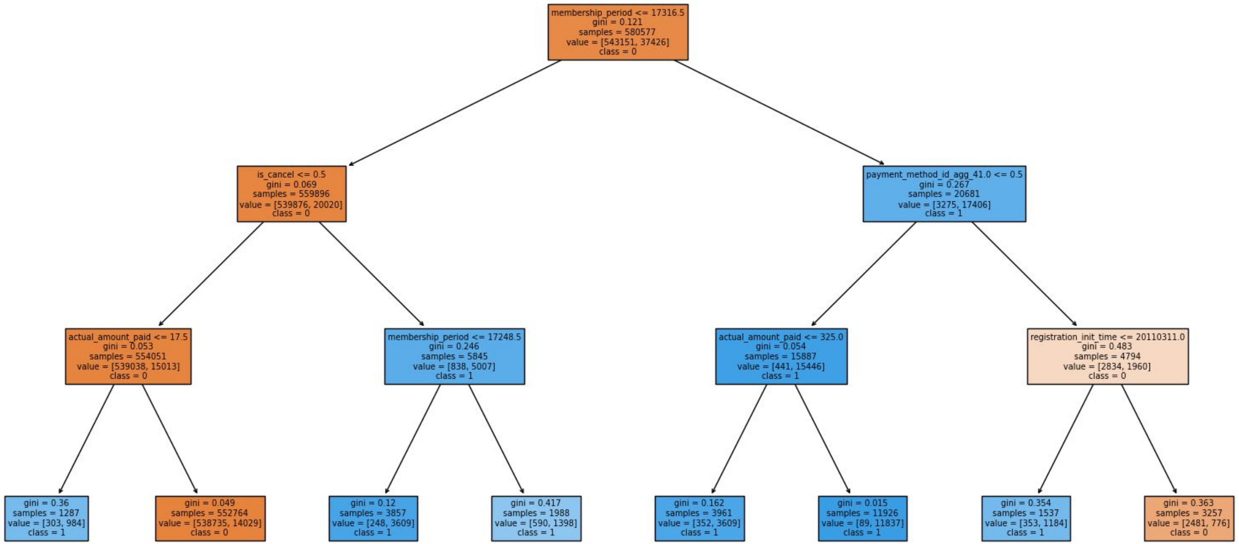
**is_auto_renew**

**is_cancel**

# 07. Implication

From the decision tree analysis, it was revealed that a specific payment method is associated with the churn rate.

## Decision Tree




**membership period**

**payment_method_agg_41**

**is_cancel**