Can We Predict Churn Rate?:

The Lifeline of Startups

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. • Users to make a new subscription within 30 days after their current membership expiration date.
Solution	• Machine learning: Classification • Solutions: Data preprocessing, feature engineering, and applying and evaluating the model.
Potential Impact	• Better business performance: Retaining existing customers usually comes at a lower cost than acquiring new ones. • 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. •6M •msno(unique id), city, bd(age), gender	•ID and Churn status •970K •msno, is_churn(class labels).	•User transactions data •1.4M •msno(unique id), payment_method_id, plan	•User engagement •18M •msno(unique id), date, num_25, num_50

2: Preparation

Initial Data Cleaning
For merger

Merger
Necessary to make ID unique and consolidate all columns.

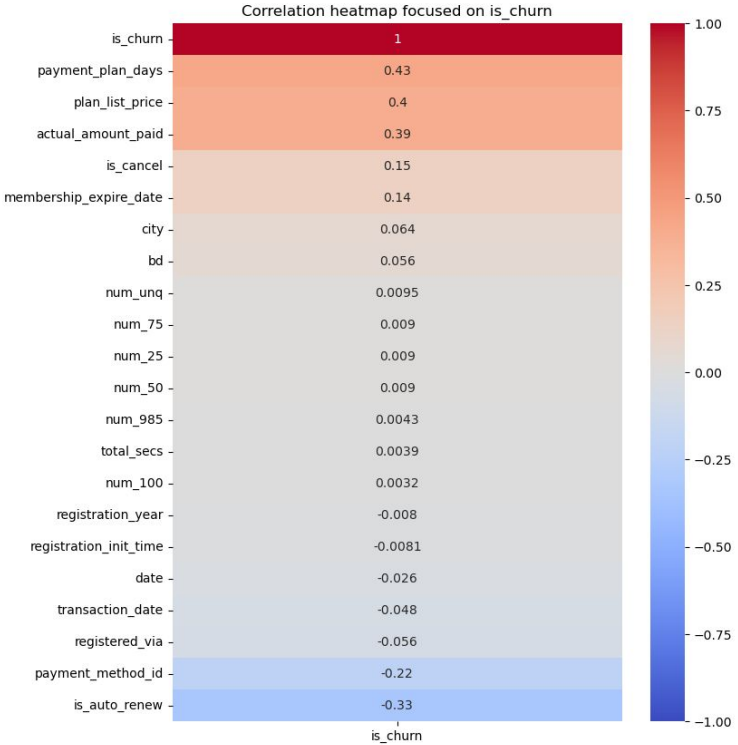
3: Preprocess

After Pre-Processing
Outlier, null-value, duplication

Feature Engineering
New column; long_time_user

03. EDA Summary

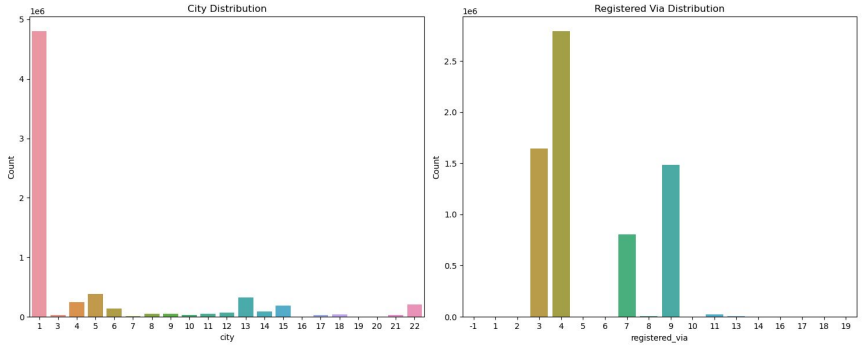
The relationship between the explanatory variables and the target variable is getting clear through correlation analysis.

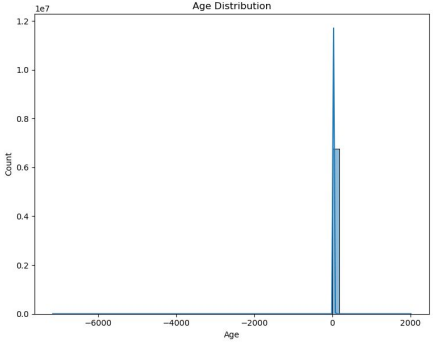


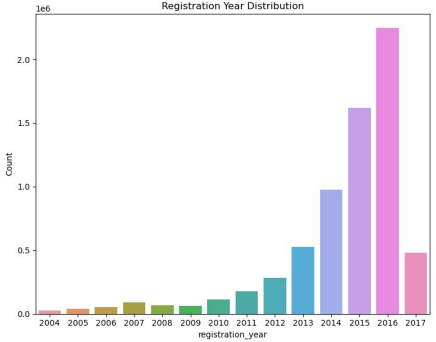
- Strong negative correlation with is_auto_renew.
- registered_via and payment_method_id show weak negative correlations with is_churn.
- Slight positive correlation between city, bd (age), and is_churn.
- Other features have very weak correlations.

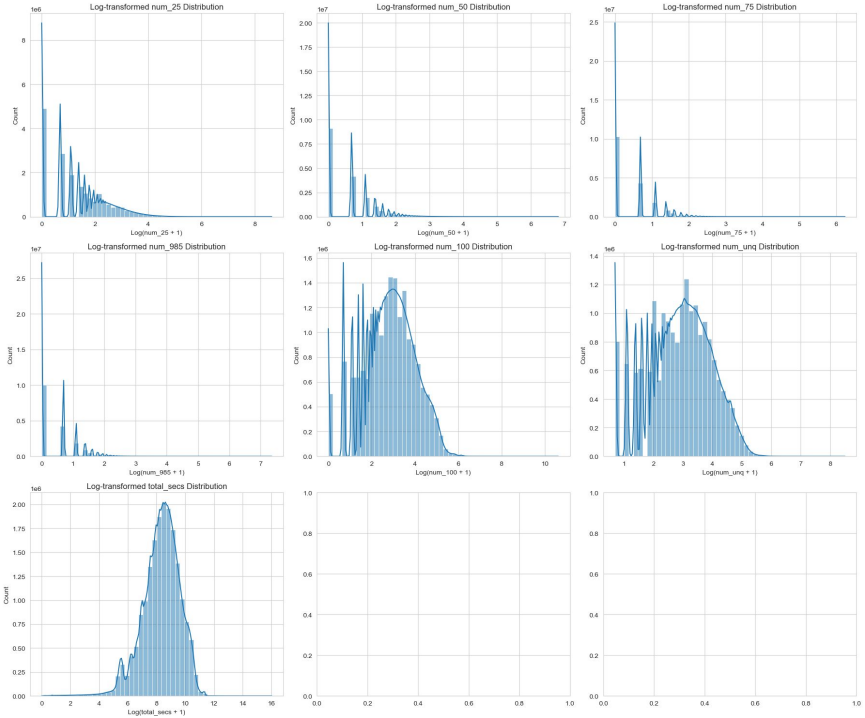
03. EDA Summary

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.



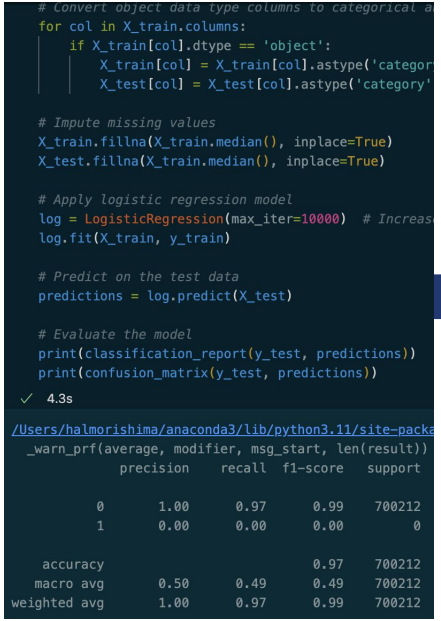






04. Baseline Model & Next Action

The results of the baseline model are not meaningful, prompting a need to iterate back to data cleansing and feature engineering in an iterative manner.



```
# Convert object data type columns to categorical a
for col in X_train.columns:
    if X_train[col].dtype == 'object':
        X_train[col] = X_train[col].astype('categor
        X_test[col] = X_test[col].astype('category'

# Impute missing values
X_train.fillna(X_train.median(), inplace=True)
X_test.fillna(X_train.median(), inplace=True)

# Apply logistic regression model
log = LogisticRegression(max_iter=10000)  # Increas
log.fit(X_train, y_train)

# Predict on the test data
predictions = log.predict(X_test)

# Evaluate the model
print(classification_report(y_test, predictions))
print(confusion_matrix(y_test, predictions))
```

✓ 4.3s

```
/Users/halmorishima/anaconda3/lib/python3.11/site-packa
  _warn_prf(average, modifier, msg_start, len(result))
              precision    recall  f1-score   support

           0       1.00      0.97      0.99    700212
           1       0.00      0.00      0.00         0

    accuracy                           0.97    700212
   macro avg       0.50      0.49      0.49    700212
weighted avg       1.00      0.97      0.99    700212
```

Initial Data Cleaning

Outlier, null-value, duplication

Feature Engineering & Normalization

Create basic statistical features, deal with skewness

Modeling

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier
- Neural Network Model