Mevod Marketing Analysis

Rico Ma

Github Repo: https://github.com/Dirodzz/Final_Case.git

01 Case Introduction

02 EDA; Data Cleaning & Feature Engineering

03 Customer Segmentation

04 Attribution & CAC

05 Churn Modeling; Revenue Modeling; CLV

01

Case Introduction

Business Background

Mevod is a Dubai based video on demand company which is trying to move into providing OTT services to the local market.

The strategy of Mevod is to be the dominant regional OTT services provider by tailoring their services to Middle Eastern/Arabic-speaking market and provide superior customer services.

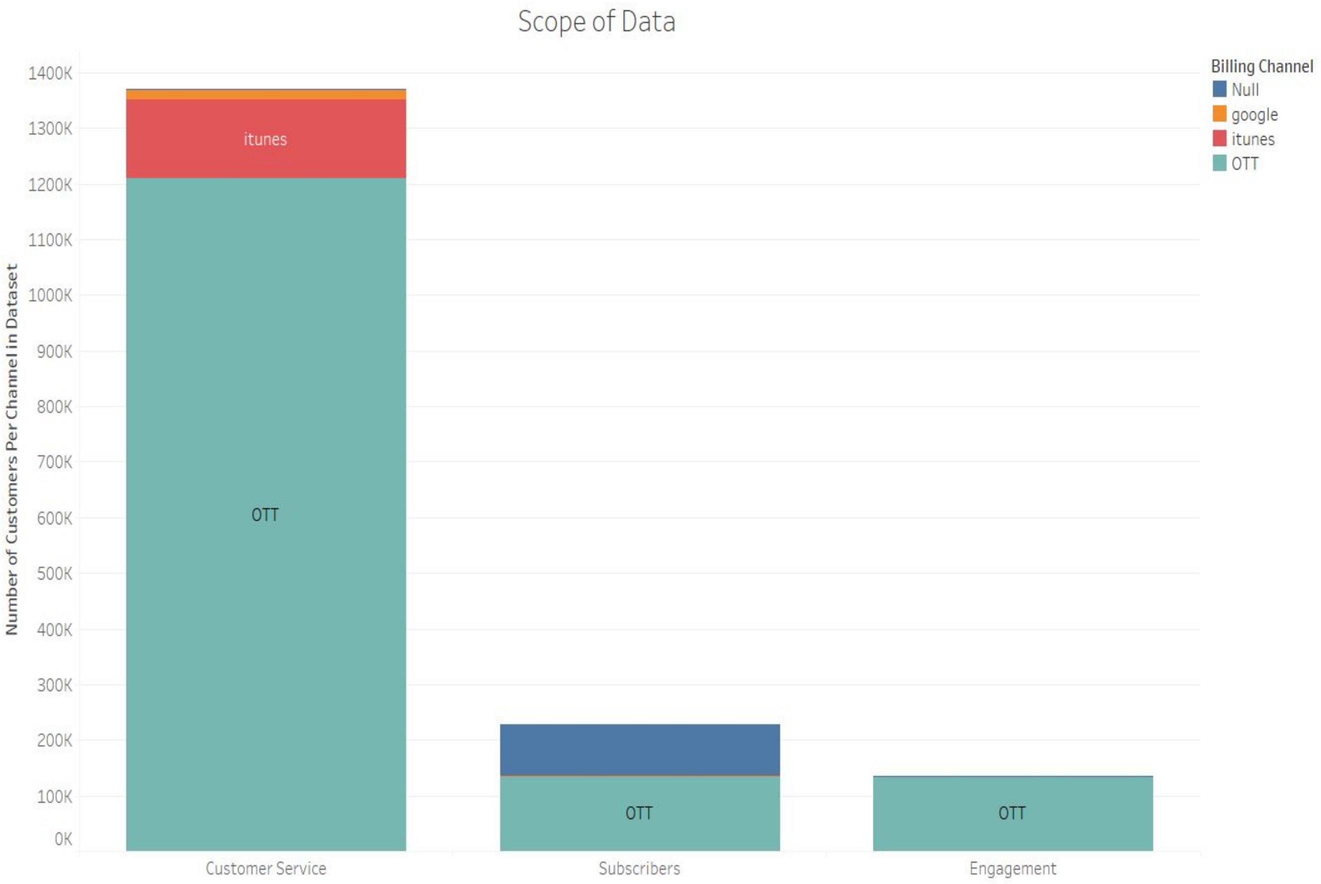
Key Objectives

1. Acquire customers, penetrate the local market
2. Increase the conversion rate for customers in trial period
3. Optimize marketing efficiency
4. Reduce churn rate

02

EDA; Data Cleaning & Feature Engineering

Number of Subscribers in Datasets



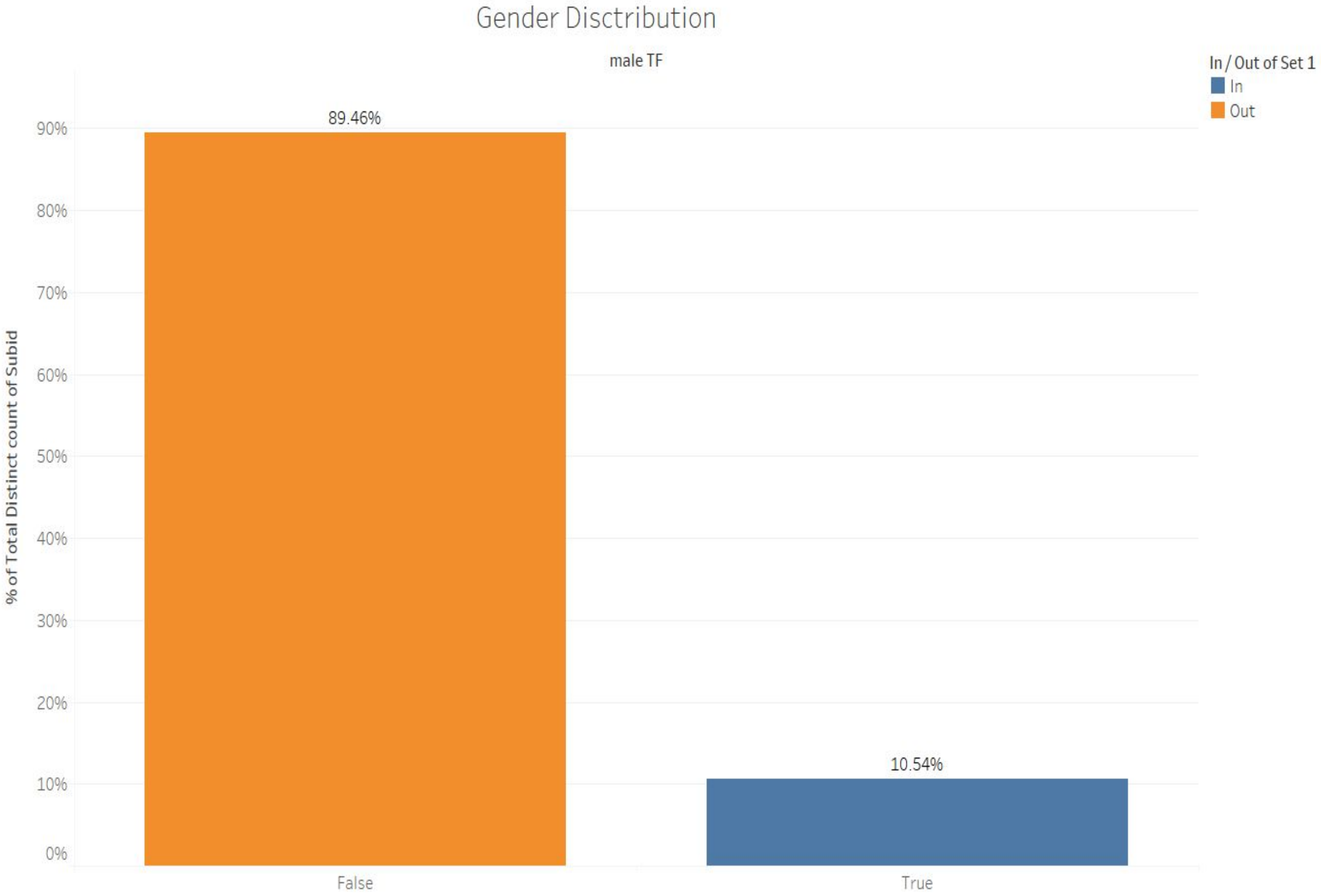
Customer Service Reps dataset contains records for 1.37M subscribers from various channels such as OTT, google, and iTunes

Subscribers dataset contains records for 0.22M subscribers, mostly from OTT channel

Engagement dataset contains records for 0.13M subscribers, all from OTT channel

For the purpose of this analysis, only **subscribers showed in Engagement** dataset will be used (total count is 135,019 subscribers).

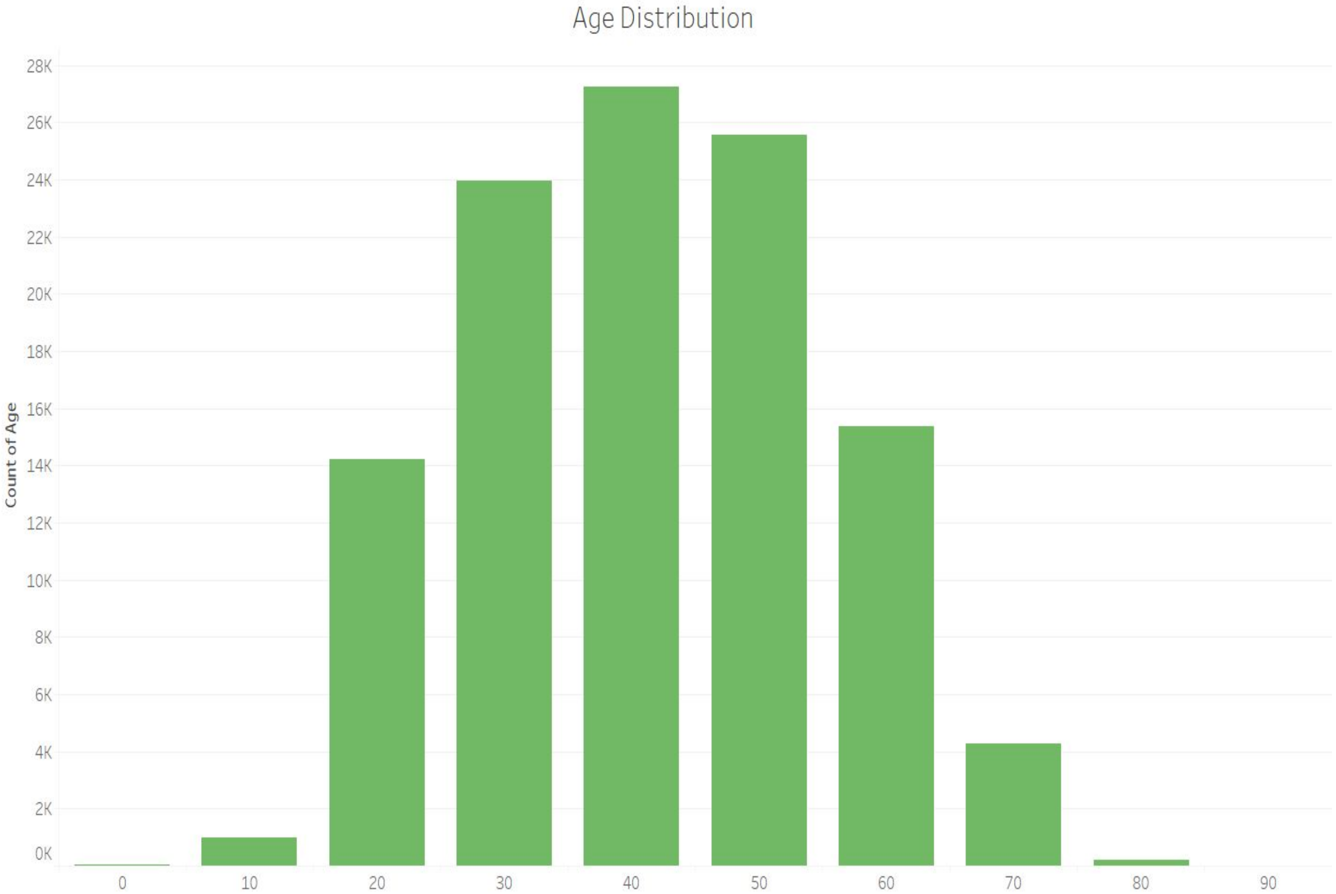
Gender Distribution



In terms of **Gender Distribution**, the current data showed that there is a significant gender difference. Almost 90% of the subscribers are Female.

This result indicates that the current OTT services and/or marketing strategy lacks the ability of acquiring male customers

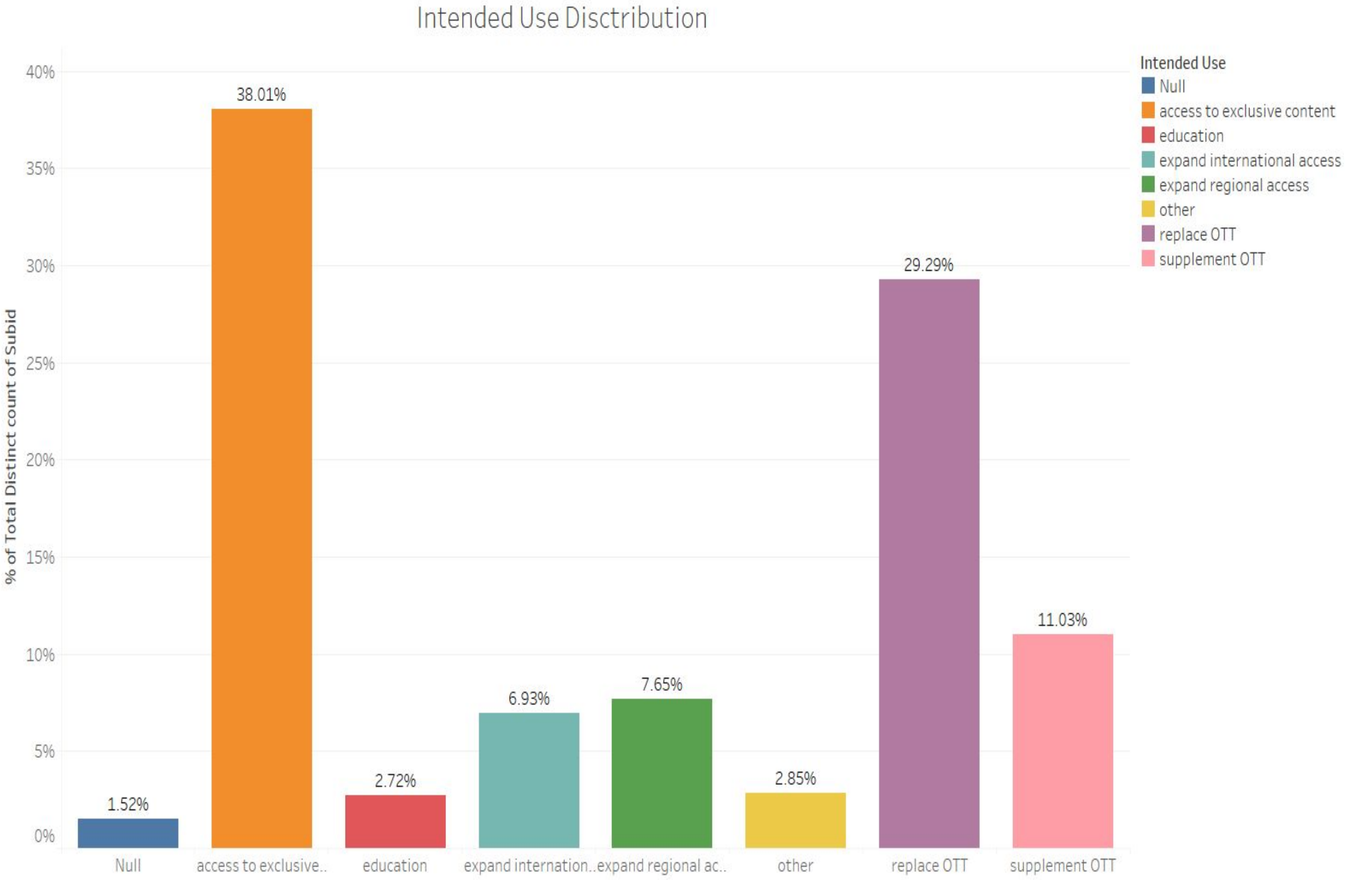
Age Distribution



Age Distribution is almost a normal distribution with an average of 45.74 and standard deviation of 13.65

Note: Age has many invalid inputs, the system should limit the input range to improve the reliability of data (e.g. range in (0, 150))

Intended Use Distribution



Almost 40% of subscribers value the ability to access to exclusive content and 30% of subscribers are targeting on removing existing OTT.

This result shows that the company should continue to focus on **(exclusive) content creation and inclusion**, in addition to the customer services aspect

Other Findings

- 47.34% of customers have internet package level as “base” , and 26.83% of customers use “enhanced” level of internet.
- 68.12% of customers are using IOS system and the rest of them are using Android
- All recorded subscriber have language setting as “ar” (Arabic), country as “UAE” and paying discounted price for their subscription
- Most subscribers (99.78%) received a base 14-day trial, paying in “uae” (Emirati dirham)
- Over 75% subscribers do not have information about “payment type”, “num_ideal_streaming_services”, and “num_weekly_services_utilized”.

Feature Processing

Based on the previous EDA findings, some of these features in provided datasets are not useful or contains invalid information for the purpose of this analysis. Data cleaning processes have been done sequentially as following

1. **Too many missing values**: features with too many missing values (over 70%) may cause bias or missing leading results. These features will be **dropped** (e.g. “payment_type”, “num_ideal_streaming_services”, “num_weekly_services_utilized”, etc)
2. **Conflict Data**: data from certain features conflict with each other. Some values are corrected to ensure the overall consistency (e.g. “trial_completed”, “current_sub_TF”, “last_payment”, “next_payment”, “cancel_before_trial_end”, etc)
3. **Low variance**: features with low variance (one value dominants) provides limited information to reflect the differences among customers. These features will also be **dropped** (e.g. “language”, “country”, “plan_type”, etc)
4. **Invalid value**: Invalid inputs in certain features (e.g. age over 100) will be **marked as missing information**.
5. **Missing values**: Missing values in certain features are **filled by average value** of that feature (i.e. “weekly_consumption_hour”, “revenue_net”, “age”, “join_fee”)[1]

[1] Unable to use KNN to fill missing values due to the intensive system memory requirement

New Features Created

In addition to the existing features in the datasets, new features are created to better reflect customers' situations and usages of the services

1. **Daily Average Engagements Per Period**: Since the engagement records for customers are given in a daily format, for each customer, the degree of their engagement will be reflected by calculate their daily average in each period. (e.g. "daily_app_opens_period_0", "daily_num_series_started_period_1", etc)

2. **Delta to Previous Period**: After calculate the daily average engagement information, for each customer, create "delta to previous period" feature to reflect the change of degree of their engagement by using a certain period daily average minus previous period's daily average (i.e. "delta_to_previous_daily_app_opens_period_1" = "daily_app_opens_period_1" - "daily_app_opens_period_0")

3. **Churn and Churn-related**: To be explained in later pages

Churn Identification

From the data provided, there are multiple ways to flag whether a customer is churned

Optimistic

Identify churners only if their accounts have been cancelled (voluntarily or failed to fulfill payments)

Conservative

Identify churners by using engagement record. If a customer stop to engage with the company's services at a certain time, they may still be a subscriber (e.g. forget to cancel or simply remove their bill information, waiting for the system to cancel), they will be considered to churn at their last engaged period

Churn Identification - Details (part 1)

To better reflect the real-life situation, churn will be determined using the conservative method, based on the account creation date and engagement period information. The detailed determination process as following

1. **Maximum Period**: By knowing that the timeframe of engagement dataset is up to 2020-04-30, under the assumption that 1) every customers will receive a 14-days trial period and sign a 4-month contract for paid period if they decide to; 2) no time gap between any consecutive payment period, the maximum payment period any customer could reach could be calculated.

2. **Last Engaging Period**: From the engagement period, assuming the latest engagement record is the last time for a customer to use this service. The payment period of this latest engagement record is the last engaging period for this customer.

Churn Identification - Details (part 2)

3. **Churn**: Churn status is determined by checking whether the Maximum Period and the Last Engaging Period for a customer are the same. If they are the same, it indicates that the customer subscribing the services constantly. Otherwise, the customer will be consider as a churner and the Last Engaging Period is the period that customer left.

Example: A customer created their account on 2020-04-02, the maximum period this customer could reach is payment period 1. If this customer's engagement record ends at some day in period 0, his or her Last Engaging Period will be period 0 and the customer will be considered as churned at period 0.

After this churn identification process, out of 135,019 customers, 104,270 (77.23%) of them are marked as churn.

03

Customer Segmentation

Scope and Feature Selection

With the consideration real life implementation, the scope of this analysis is to use customers’ self-reported data, trial period information, and churn status to segment customers and help to design targeted acquisition strategy

Features used are

Age; Gender; Intended Use; Internet Package Type; Operating System; Preferred Genre; Weekly Consumption; Cancel before Trial Ends; Refund after Trial; Join Fee; Number of Engagement Activities in trial period (daily average); Maximum Period; Last Engaging Period; Churn

Segmentation Result - General (part 1)

Clusters	latest_eng_record_period	Churn	cancel_before_trial_end	revenue_net	male_TF_True
Cancel_in_Trial	0.00	1.00	0.92	0.06	0.00
Male	0.55	0.70	0.47	2.45	1.00
Paid_User	1.21	0.46	0.00	3.88	0.00

Using K Means model, customers could be separated into three groups: 1) “Cancel_in_Trial”, 2) “Male”, 3) “Paid_User”

1. **“Cancel_in_Trial”**: people who never reach to paid period. Among these customers, 92% of them canceled this service before trial ends and they generate almost no revenue to the company
2. “**Male**”: All male customers are in this group. Although 55% of them also not reach to their paid period, compared with the 60% population drop rate after trial period, male customer showed a higher conversion rate
3. **“Paid_User”**: People who never cancel during their trial period and have usually reach to paid period. These customers generate most revenue to the company

Segmentation Result - Usage (part 2)

Clusters	weekly_consumption_hour	access to exclusive content	replace OTT	daily_app_opens_period_0	daily_num_videos_completed_period_0
Cancel_in_Trial	26.92	0.36	0.30	1.64	2.58
Male	34.77	0.42	0.31	1.78	2.55
Paid_User	27.48	0.39	0.27	2.25	3.33

Customers in different segments varies in terms of self-reported information and number of engagement activities

1. **"Cancel_in_Trial"**: These people reported the lowest weekly consumption and less interesting in access to exclusive content. They are also less engage to the service in their trial period
2. "**Male"**: Male customers have the highest interest to the access of exclusive content and replace OTT. They also reported the highest weekly consumption. But this self-reported information is not reflected in their number of daily engagement activities
3. **"Paid_User"**: These people showed the lowest interest to replace their OTT but their number of engagement activities are the highest, compared with other groups

Segmentation Result - Acquisition Channel (part 3)

Clusters	technical_facebook	survey_facebook	survey_referral	survey_tv
Cancel_in_Trial	0.39	0.57	0.07	0.16
Male	0.25	0.37	0.11	0.24
Paid_User	0.33	0.52	0.13	0.16

In terms of acquisition channel, there are also differences from these groups

1. **“Cancel_in_Trial”**: These people have a higher probability from facebook advertising
2. **“Male”**: Male customers have a lower probability from facebook advertising and are more likely be influenced by tv and referral
3. **“Paid_User”**: Customers who reached to their paid period are affected by referral more often

Future Recommendation

1. Despite the segmentations, at least 70% of customers intended to use the service to access to exclusive content or replace OTT. So, the current strategy of acquiring and/or creating regional content is valuable for customers indeed.
2. To acquire male customer, facebook is not as effective as to people in other segments. TV, email and/or other channel could be considered if targeting customer in this group.
3. Customer acquired by referral are more likely to reach their paid period. The company could introduce incentives for referring to expand the referral effect. An example could be that if a new customer is referred by an existing customer, both of them could receive a discount for their next paid period.
4. For people who are clustered into “Cancel_in_Trial” group, the company could have customer representative contact them during/after their trial period, trying to ask the reason behind their low number of engagement activities. These data could be further analyzed and create marketing/product strategies based on these information. If their demand are not feasible, the company should consider to lower the marketing investment for people in this group.

04

Attribution & CAC

Assumption, Scope and Feature Selection

To perform this attribution and CAC analysis, multiple assumptions are made.

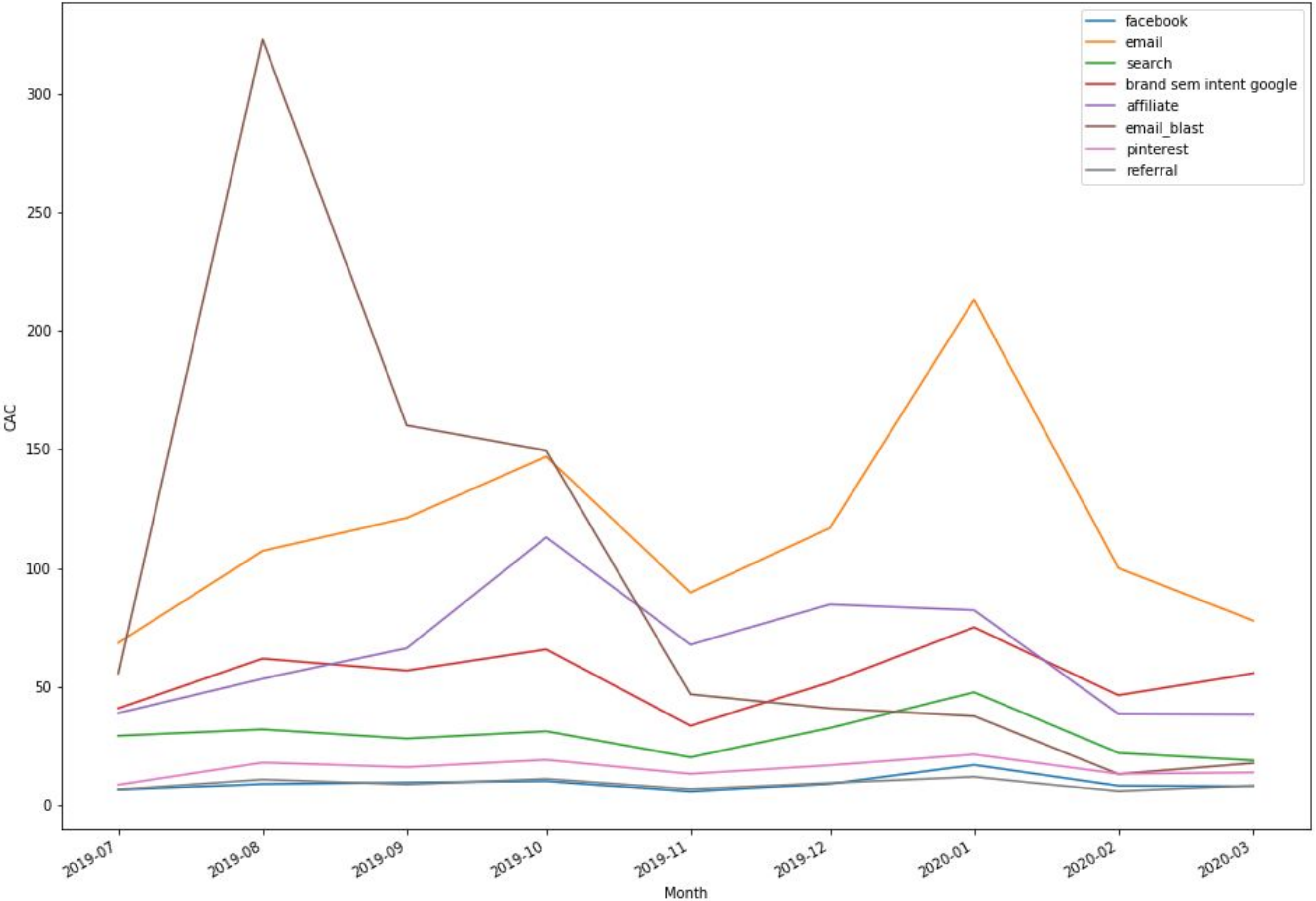
1. Advertises have no delay or carryover effect. Credit from a new customer acquisition will only be assigned to the advertises happened in that month without considering advertises in previous months.
2. Advertises is continuous in any given month (not happened occasionally in particular days of a month).
3. Customer account creation date is the date a customer been acquired.
4. Customers' self-reported attribution channel are at least partially true (e.g. one should not have reported attribution as "facebook" while not having facebook account at all).

Features used are
Account Creation Date; Attribution_Technical; Attribution_Survey; Advertising Monthly Spending Data

Approach

1. **Filling missing value in "Attribution_Survey"**: Some customers did not report their attribution channel, these missing value are filled using value "other".
2. **Obtain Customer Acquisition Month**: Using information provided in Account Creation Date, obtain the month of acquisition for each customer
3. **Obtain Number of Customer Acquired by Channels for Each Month**: Using acquisition months, attribution_techinical and attribution_survey, count the number of customers acquired by channels for in a monthly basis. For every customer, attribution_technical and attribution_survey are considered to have the same weight. Channels reported in attribution_technical and attribution_survey will get 0.5 credit. One exclusion is where channel "tv" appears in attribution_survey. Because "tv" channel is not traceable to individual customer level, tv will not appear in attribution_technical inputs. Therefore, if customers reported "tv" in attribution_survey, their attribution_technical input will be ignored and a full credit (1) will be assigned to "tv".
4. **Calculated Monthly CAC**: After having number of customer acquired by channels for each month, use advertising spending amounts reported for each month divide by the number of customer acquired by channels for that month to get monthly CAC for each channel. Note that not all channels have the spending information reported. Therefore, this analysis will only include CAC results for channels with spending data.

CAC Result and Recommendation



Because the **month of 2019-06** have a lower than normal number of customer, this month's CAC **is removed** in CAC analysis.

Email and Email Blast advertising usually have a higher CAC. Referral, facebook, and pinterest are usually the cheap options.

Email blast showed a decreasing CAC overtime. This may due to the better customer targeting algorithms or the improved popularity of the company.

Combining with the segmentation results, over 50% of customers are from facebook but these customers are harder to keep as well. The company should investigate further to improve the targeting strategy of facebook advertising.

05

Churn Modeling
Revenue Modeling
CLV

Churn Modelling – Goal and Feature Selection

To **predict** whether **a customer** who passed his or her trial period (currently in the first paid period) **will churn in the next period** at the end of paid period 1.

Under this setting, **features** are available at that time and used in the prediction model are

Age; Gender; Intended Use; Internet Package Type; Operating System; Preferred Genre; Weekly Consumption; Cancel before Trial Ends; Refund after Trial; Join Fee; Paid TF; Number of Engagement Activities in trial period and paid period 1 (daily average); Delta of Number of Engagement Activities to Previous Period (paid period 1 minus trial period); Churn

Approach

1. **Select Qualified Dataset**: Customer who could reach 2 paid period at maximum and did not cancel services at trial period are selected. Out of 135,019 customers, 22,852 customers are selected. 14,636 were churned and the remaining are still with the company.
2. **Convert Categorical Data to Binary Value**
3. **Split Dataset into Train and Test**: 60% are used to train model and 40% for testing. The train_test_split is performed by stratified sampling which ensures train dataset and test dataset have the same churn distribution as the population (Churn:Non-Churn = 16:9).
4. **Standard Scale**: Use standard scalar to normalize continuous variables for models which may be sensitive to the differences of scales of data
5. **Tuning Hyperparameters**: GirdSearchCV and RandomSearchCV (5-fold cross validated) are used to tuning hyperparameters for models. Parameters for models are selected based on model accuracy.

Models tested are Logistic Regression, Decision Tree, GBDT, and Random Forest

Churn Modeling Result

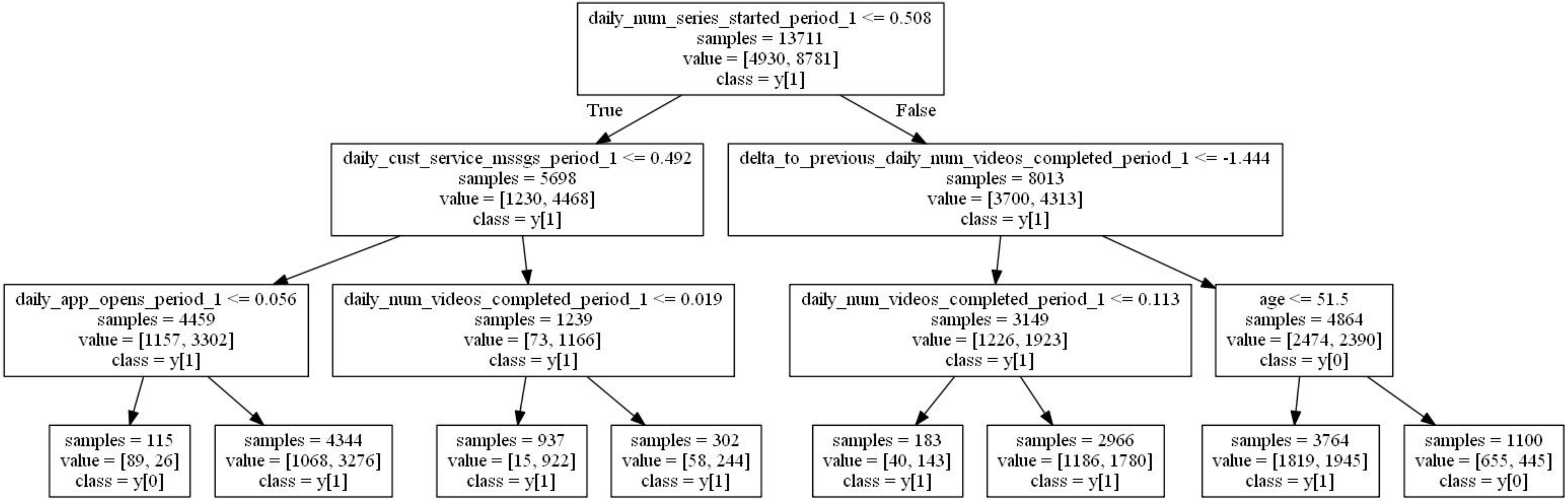
	Logistic Regression	Decision Tree	GBDT	Random Forest	Baseline Model (Dummy Classifier)
Test Accuracy	0.6677	0.6577	0.6706	0.6726	0.6405
Train Accuracy	0.6730	0.6718	0.6934	0.7509	0.6404
AUC	0.69	0.68	0.71	0.71	0.5

A dummy classifier which predicts every customer as churner is created for comparison purpose.

Ensembled models like GBDT and Random Forest **outperform other basic models**. However, random forest (with max_depth = 10 and min_samples_leaf = 2) showed a tendency of overfitting. After manually turning hyperparameters to reduce overfitting will also cause a reduction in model performance.

Since GBDT has similar AUC and test accuracy as Random Forest, **GBDT is chosen to be the prediction model**.

Decision Tree Display



daily_num_series_started_period_1 <= 0.508
samples = 13711
value = [4930, 8781]
class = y[1]

True

False

daily_cust_service_mssgs_period_1 <= 0.492
samples = 5698
value = [1230, 4468]
class = y[1]

delta_to_previous_daily_num_videos_completed_period_1 <= -1.444
samples = 8013
value = [3700, 4313]
class = y[1]

daily_app_opens_period_1 <= 0.056
samples = 4459
value = [1157, 3302]
class = y[1]

daily_num_videos_completed_period_1 <= 0.019
samples = 1239
value = [73, 1166]
class = y[1]

daily_num_videos_completed_period_1 <= 0.113
samples = 3149
value = [1226, 1923]
class = y[1]

age <= 51.5
samples = 4864
value = [2474, 2390]
class = y[0]

samples = 115
value = [89, 26]
class = y[0]

samples = 4344
value = [1068, 3276]
class = y[1]

samples = 937
value = [15, 922]
class = y[1]

samples = 302
value = [58, 244]
class = y[1]

samples = 183
value = [40, 143]
class = y[1]

samples = 2966
value = [1186, 1780]
class = y[1]

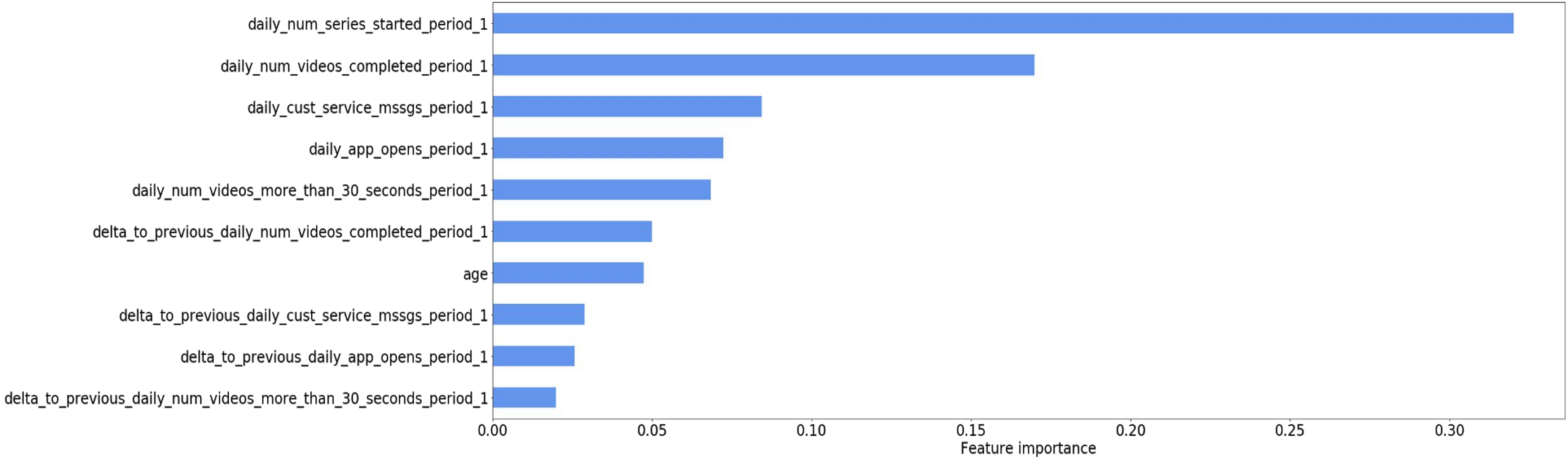
samples = 3764
value = [1819, 1945]
class = y[1]

samples = 1100
value = [655, 445]
class = y[0]

Even though decision tree is not chose for this analysis, it is still useful for show the decision tree graph to understand how tree-structured model made decisions.

By limiting the tree to have depth of 3, the decision is made as shown above.

GBDT Feature Importance



Because **GBDT model** are a combination of decision trees, there is **hard to illustrate the decision process**. But important features of the model could be obtained (showed above). From this feature important graph, conclusion can be drawn.

1. **Daily average number of engagement activities and changes of these numbers after the trial period are major indicators** for churn status. So, company should keep monitoring customers' engagement activities and may introduce other measurements to better reflect the degree of customer involvements.
2. **Age** should also be taken into consideration. As the current age data have many missing and invalid inputs, company should improve the data collection process (e.g. Have age as a necessary input and ensure input type is integer while range is within 0 to 100).

Churn Modeling Result

	Logistic Regression	Decision Tree	GBDT	Random Forest	Baseline Model (Dummy Classifier)
Test Accuracy	0.6677	0.6535	0.6706	0.6726	0.6405
Train Accuracy	0.6730	0.6603	0.6934	0.7509	0.6404
AUC	0.69	0.66	0.71	0.71	0.5

A dummy classifier which predicts every customer as churner is created for comparison purpose.

Ensembled models like GBDT and Random Forest **outperform other basic models**. However, random forest (with max_depth = 10 and min_samples_leaf = 2) showed a tendency of overfitting. After manually turning hyperparameters to reduce overfitting will also cause a reduction in model performance.

Since GBDT has similar AUC and test accuracy as Random Forest, **GBDT is chosen to be the prediction model**.

Revenue Modelling – Goal and Feature Selection

Based on the churn probability obtained from churn prediction model, introduce a discount offer with certain discount rate to churners to revert their churn decision and maximize overall revenue.

Dataset are the same as the one used in churn modelling (22,852 customers)

Under this setting, **features** used in this analysis are

Revenue_net_1month; Churn_Actual; Churn_Predicted; Churn_Predicted_Probablity

Assumption

Several **assumptions** are made for this revenue modelling

1. Customers paid materially no fee for their trial period
2. Revenue_net_1month inputs equal to the net revenue of customer generated for every paid period divide by their number of months in each paid period (in this case is 4-month)
3. Due to the low variance of discounted price, the churn prediction model result does not influenced by changes in subscription fee. But to illustrate the changes in revenue, assuming the churn probability will be affected by the change of prices.
4. Offer acceptance rate is inversely related to the degree of discount offered (offer acceptance rate = 1 - discount rate)

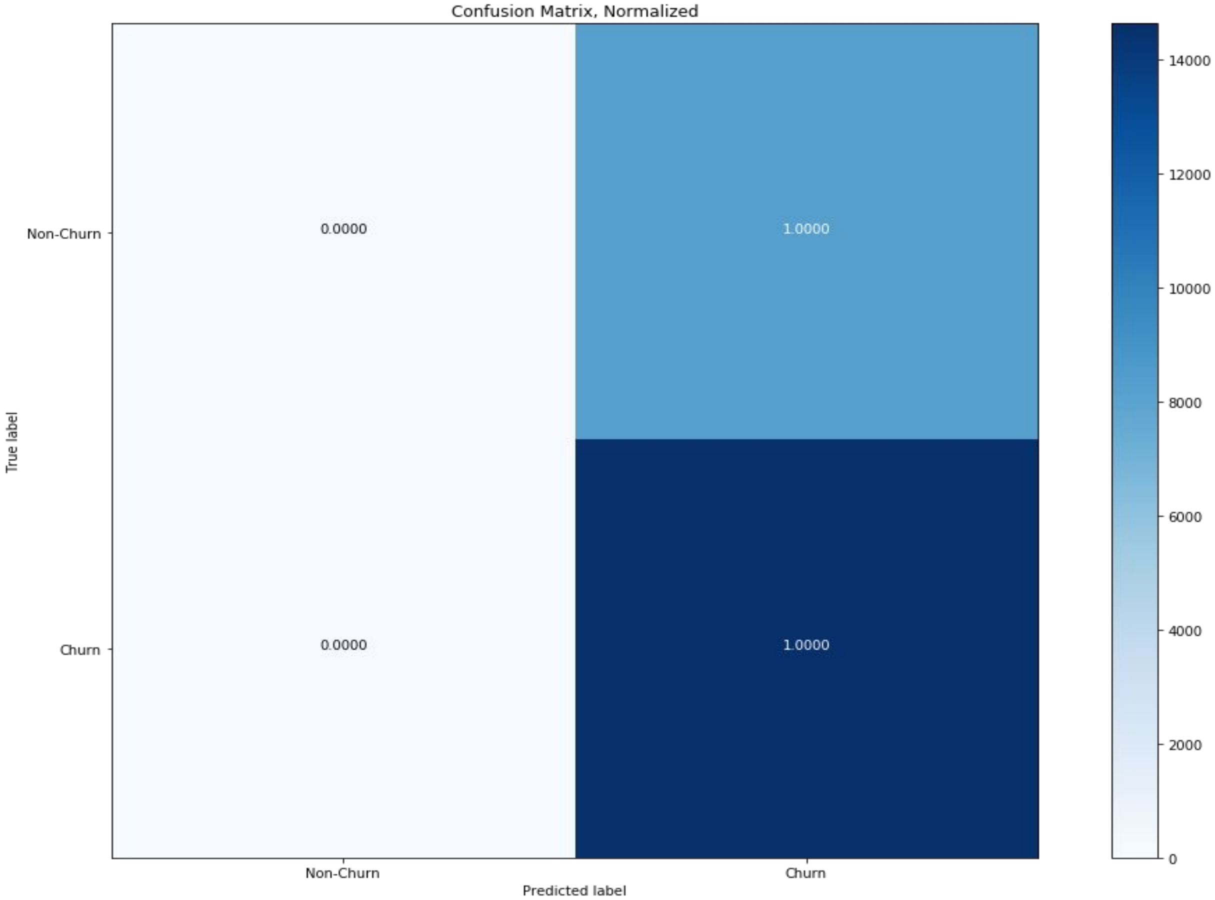
Optimization Process

Based on the information available, there are **two decision variables** to be determined in order to optimize the revenue

1. **Decision Boundary**: Given the churn probability from the prediction model, what is the lowest probability required that make the company to give offers?
2. **Discount Rate**: If certain customers have a churn probability that higher than or equal to the lowest churn probability required to receive the offer, what is the discount rate the company should offer?

Under the assumption that offer acceptance rate = 1 - discount rate, this revenue model **search over all possible combination of decision boundary and discount rate**, with the increment of 0.01 and maximum of 0.99 for decision boundary and the increment of 0.05 and maximum of 0.95 for discount rate.

Revenue Modelling Result



If the company **do nothing** to respond (a discount offer is not introduced at all), the total revenue generated from this customer in next period is **$144,844.48**

After the revenue optimization process, the best result is to use decision boundary of 0.06 and discount rate of 5%. Under the decision boundary of 0.06, the result is shown on the left. It implies that all customers are treated as churners. The improved revenue is **$347,809.66**.

Recommendation

The revenue modelling result indicates that if assumptions could reflect the real situation, the company should consider to **give a 5% discount to every customer** who are in their first paid period no matter they are a churner or not, the reduction of revenue caused by non-churners paying less will be offset by the additional gain from churners that become non-chuners.

However, the **current revenue model is more likely to be naive** due to that the acceptance rate is set to be (1 - discount rate). **Further investigation** could be done **in the area of customers' price elasticity** (e.g. customer survey, AB Testing, etc). Then the reliable offer acceptance rate information in respect of various degrees of discount will result a more accurate result.

CLV – Goal and Data Selection

Calculate their **lifetime value for customers** who created account recently and are currently **in their first paid period**.

Customers used for this analysis are selected based on the maximum period customers could reach (max_period = 1) and whether they churned in trial period (last engaging period = 1). 20,917 customers are selected for this analysis.

Features used in this analysis are those used churn modeling and month CAC information calculated above.

Assumption, Setting & Approach

1. The **probability of churn** for a customer **is constant**. The customer will not change the churn probability at any point in time in the future.
2. The **market interest rate is 10%** annually and **also will not change** at any point in time in the future

Under these environment, information required for CLV calculation for each customer is calculated as followed

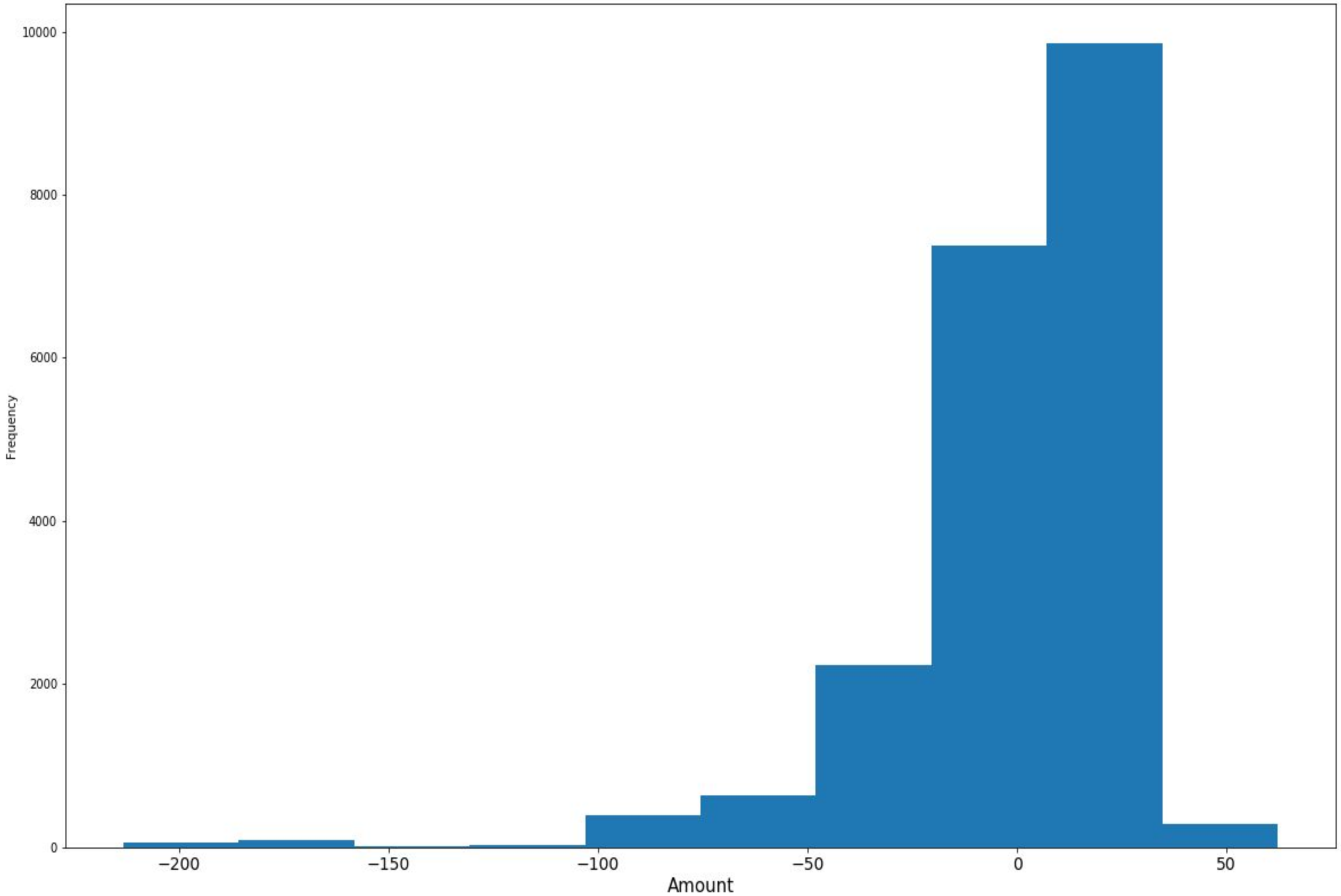
1. **Revenue (nominal)**: The revenue of each customers is calculated by multiplying the revenue_net_1month of this customer by 4.
2. **Probability of payment**: Use the churn prediction model to output the probability of churn. The probability of payment is 1 - the probability of churn.
3. **Acquisition Cost**: The acquisition cost for customers are calculated based on the customers’ account creation date, attribution_technical and attribution_survey. Use the account creation date to identify the month in the monthly CAC result table. Then the average of two CACs for channels reported in attribution_technical and attribution_survey in that month will be the acquisition cost for this customer. Missing cost will be filled by using the average cost of other customers.

CLV Formula

After having information required for each customer, CLV is calculated by using this formula

$$CLV = revenue(nominal) \times \left(\frac{1 - \frac{0.1 \times 4}{12}}{1 - \frac{0.1 \times 4}{12} - Probability\ of\ Payment} \right) - CAC$$

CLV Result



60% of customers have a positive CLV which implies that **over half of the current customers are generating revenues** for the company.

However, the **average CLV** amount all these customers **are still negative** (-0.53). This implies that these customers on average are not making money for the company.

The company should consider to **reduce the number of customers who have a higher CAC to become churners**.

Overall Conclusion

- Keep implementing the existing strategies of acquiring and/or creating exclusive content because these are the one of the main reasons for all kinds of customers to choose the OTT services.
- Improve the current data system in terms of data completeness, validity, and diversity could result more accurate and comprehensive marketing analysis in the future.
- Consider to have targeted marketing plan for customers in different segments. Customers from various segments may have different interest and the effective channel to reach them may also be different.
- Introduce a marketing budget allocation plan to optimize the advertising efficiency (i.e. allocate less budget to channels with higher CAC)
- Number of engagement activities (actual amount and delta to previous period) are good indicators for customers' loyalty and the probability of churn.
- Consider to provide offers proactively to customers who have high probability of churn to prevent reduction on revenue

THANKS!

If you have any questions, please contact

Rico Ma

bm2785@nyu.edu