Middle Eastern Video on Demand

Kaiwen Wu kw1820

Background

Mevod

Dubai based video on demand company trying to expand and become dominant regional player

Subscription business - customers pay a monthly fee for access to the service.

Moving into the **OTT** space (a new delivery method for video and audio over the Internet,)

Since this is a new business, marketing team is piloting several pricing schemes
- **No trial fee / Discounted trial fee**
- **14 day trial period / 7 day trial period**

What we are going to do

AB testing to understand what marketing strategies have been most effective to date

Customer segmentation to help the marketing team design acquisition strategies supporting the Executive team's growth objective

Build a churn model and develop recommendation(s) on an alternative product pricing structure as well as a distribution of expected CLV, representing the uncertainty of future payments given the current customer base.

Data Overview

CSR
2208643 rows
1031 representatives
136930 customers

Information regarding whether a current subscribers, trial and account

Subscribers
227628 rows
227459 customers

Self-reported / Null values

Engagement
2585724 rows
135019 customers

Stickiness / loyalty (how many time opened the app, number of video started etc.)

Data & Findings

Age - Most of customers are **female** in the age range of **35-60,** with an average of **45**, and we had almost **89%** of female customers.

Most customers (more than 99%) chose **base_usa_14_day_trial** plan with a **monthly price** of **4.7343** and a **discounted price** of **4.5141**

Most customers had a **14 days trial** and **4.065%** chose to refund after the trial, **⅔** of them uses ios system

Most customers has a relatively **low stickiness**, most of them does not rate videos or send messages to customer rep

Most customers watched **4-5** videos on average

Data & Findings

Now, let's take a closer look at the three channel

	Rep	Cust	Cancel	Renew
OTT	1031	1209872	758007	396657
google	953	17235	0	0
itunes	1007	142253	34	0

55280 (4.56%) missing

Requires manual check

Proposed analyses

Data preprocessing
Get rid of outliers and do some manual checking (including outliers of the ages, account cancel before created, [creation until cancel day] does not match trial period etc) and deal with imbalanceness.

AB Testing
Conduct AB testing twice, once regarding the length of the trial period and the other regarding trial fee.
H0: 14 Days Trial is better than 7 days
H1: 7 Days Trial is better than 14 days
H0: Free Trial is better than discounted trial fee
H1: Discounted trial fee is better than free trial

Segmentation
Conducting clustering and get customer segmentation, then prepare a customer profile to help the marketing team design acquisition strategies.

Proposed analyses

Churn

Mainly use **logistic regression** and tree based models like **decision tree and GBDT.**

logistic regression
Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting.
It can be updated easily with new data using stochastic gradient descent.
Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships.

Tree based
We can get feature importance, use it as a reference and visualize the branches and results.
Fairly robust to overfitting, doesn't require careful normalization or scaling of features.

Analyses for AB testing

Conduct AB testing twice, once regarding the length of the trial period and the other regarding trial fee.

H0: 14 Days Trial is better than 7 days
H1: 7 Days Trial is better than 14 days
H0: Free Trial is better than discounted trial fee
H1: Discounted trial fee is better than free trial

What does a converted customer looks like?

[current_sub_TF]: T
[trial_completed_TF]: T
[payment_period]: != 0

[paid_TF]: T
[refund_after_trial_TF]: F

```
]: df_converted=pd.merge(converted,converted1[['subid']],on='subid', how='left')
]: df_converted
]:
```

	customer_service_rep_id	subid	current_sub_TF	cancel_date	account_creation_date	num_trial_
0	31856201	27800927	True	NaT	2020-03-27 23:59:04	
1	39331506	27089117	True	NaT	2020-03-27 23:57:48	

Analyses for AB testing

H0: 14 days = 7 days (14 days is better than 7)
H1: 7 days > 14 days (7 days is better than 14)

	Converted	Total	Conversion Rate
14 Days	523596	1281127	0.408699
7 Days	35314	64043	0.55141

```
z = diff /np.sqrt(((p_B * (1-p_B)/64043)+(p_A * (1-p_A)/1281127)))
```

```
z
```

```
70.90511117706859
```

According to the Z table, Z 0.05 = 1.644854, 70.905 > 1.644854, so we reject the null hypothesis

We would recommend **7 days trial** for better conversion rate

Analyses for AB testing

H0: high = low (high is better than low)
H1: low > high(low is better than high)

	Converted	Total	Conversion Rate
High	97	325	0.29846
Low	82558	227096	0.36353

```
z = diff1 /np.sqrt(((p_B1 * (1-p_B1)/227096)+(p_A1 * (1-p_A1)/325)))
```

```
z
```

```
2.561838752550662
```

According to the Z table, Z 0.05 = 1.644854, 2,56183 > 1.644854, so we reject the null hypothesis

We would recommend **lower trial fee** for better conversion rate

Analyses for AB testing

When conducting the optimal sample size, we found that the ab testing for trial length is larger than the optimal sample size, while the ab testing for trial fee is smaller than the optimal sample size

```
opt_sample_size(p_A, p_B - p_A, 0.8, 0.05)
192.38433493505244
```

```
opt_sample_size(p_A1, p_B1 - p_A1, 0.8, 0.05)
6781.632720282813
```

Analyses for Customer segmentation

Used k-means to conduct clustering

First check correlation between features to see what feature to use

e.g.

	num_weekly_services_utilized
num_weekly_services_utilized	1.000000
weekly_consumption_hour	0.407255
num_ideal_streaming_services	0.844068
age	0.004579
months_per_bill_period	NaN
monthly_price	0.002487
discount_price	0.002170
age_group_Elderly(50-70)	0.009531
age_group_Mid-aged(35-50)	0.025959
age_group_Others	-0.021079
age_group_Teenagers(<18)	-0.003127
age_group_Youth(18-35)	-0.030795
package_type_base	0.015292
package_type_economy	0.038315
package_type_enhanced	-0.032840
preferred_genre_comedy	-0.125202
preferred_genre_drama	0.112720

Dropped

Analyses for Customer segmentation

Used k-means to conduct clustering, and printed out the cluster centers to see characteristics of each group

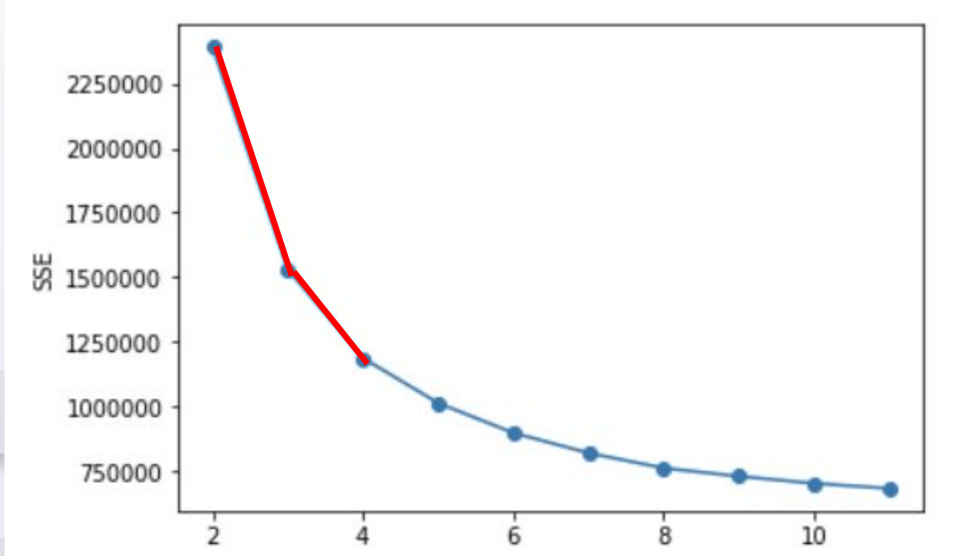
Features used :

- 'Num_weekly_services_utilized',
- 'Weekly_consumption_hour',
- 'age_group',
- 'months_per_bill_period',
- 'monthly_price',
- 'discount_price',
- 'package_type',
- 'Preferred_genre',
- 'Intended_use',
- 'plan_type'

	0	1	2
num_weekly_services_utilized	2.847	3.327	3.027
weekly_consumption_hour	23.659	36.057	29.171
months_per_bill_period	4.000	4.000	4.000
monthly_price	4.727	4.735	4.735
discount_price	4.508	4.515	4.515
age_group_Elderly(50-70)	0.378	0.348	0.368
age_group_Mid-aged(35-50)	0.322	0.366	0.352
age_group_Others	0.068	0.044	0.047
age_group_Teenagers(<18)	0.000	0.000	0.000
age_group_Youth(18-35)	0.231	0.241	0.234
package_type_base	0.457	0.440	0.456
package_type_economy	0.081	0.094	0.082
package_type_enhanced	0.263	0.311	0.281
preferred_genre_comedy	0.516	0.469	0.515
preferred_genre_drama	0.199	0.252	0.201
preferred_genre_international	0.027	0.040	0.033
preferred_genre_other	0.018	0.023	0.021
preferred_genre_regional	0.037	0.057	0.046
intended_use_access to exclusive content	0.353	0.387	0.367
intended_use_education	0.028	0.022	0.027
intended_use_expand international access	0.067	0.064	0.069
intended_use_expand regional access	0.074	0.062	0.076
intended_use_other	0.041	0.026	0.035

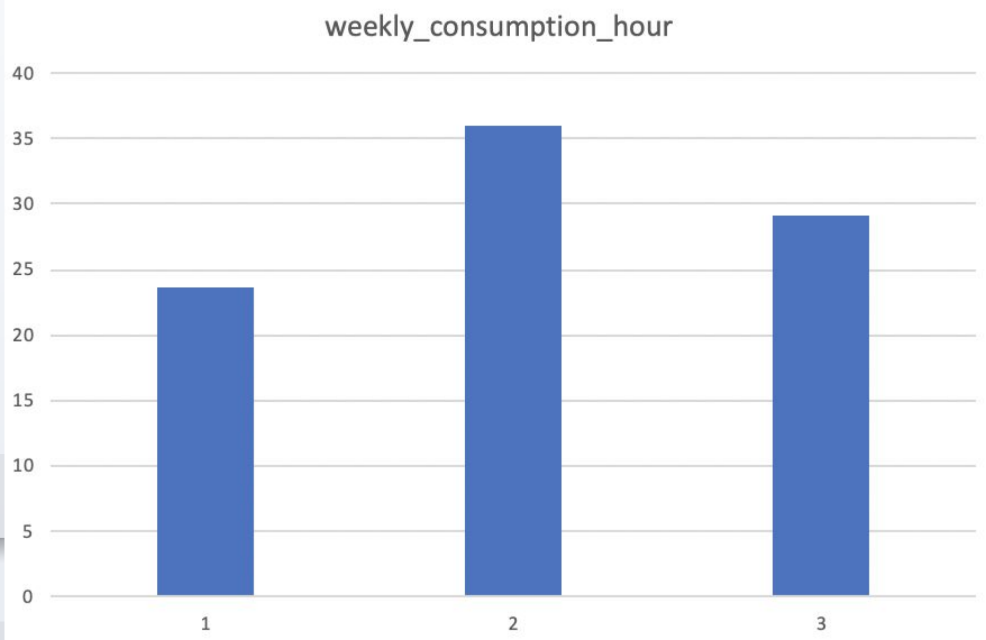
Analyses for Customer segmentation

Using elbow method and append the SSE, we found that **<u>3</u>** clusters is the best. Then we print out the cluster center for each cluster to gather insights for recommendations



Analyses for Customer segmentation

Then we print out the cluster centers to see characteristics of each group, the most distinguished feature is weekly consumption hour.



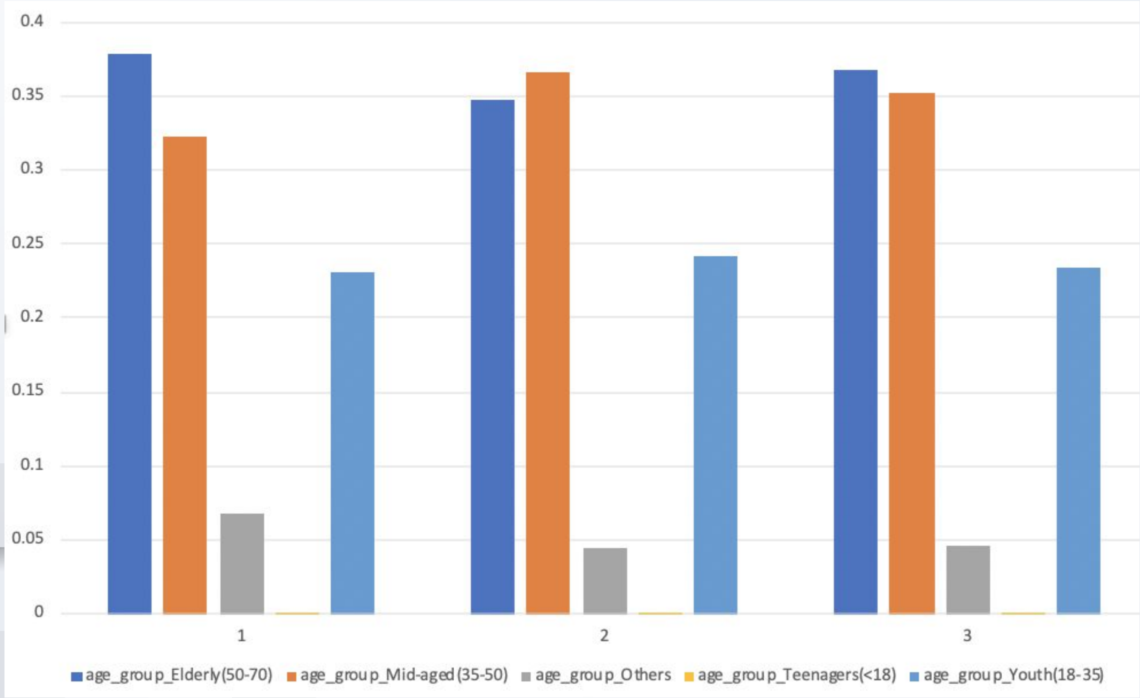
<u>Group1:</u>
23.6588219239869

<u>Group2:</u>
36.0574977111519

<u>Group3:</u>
29.171247787106

Analyses for Customer segmentation

Then we looked in to the age distribution of each group and found that mid-aged could be our major target group.



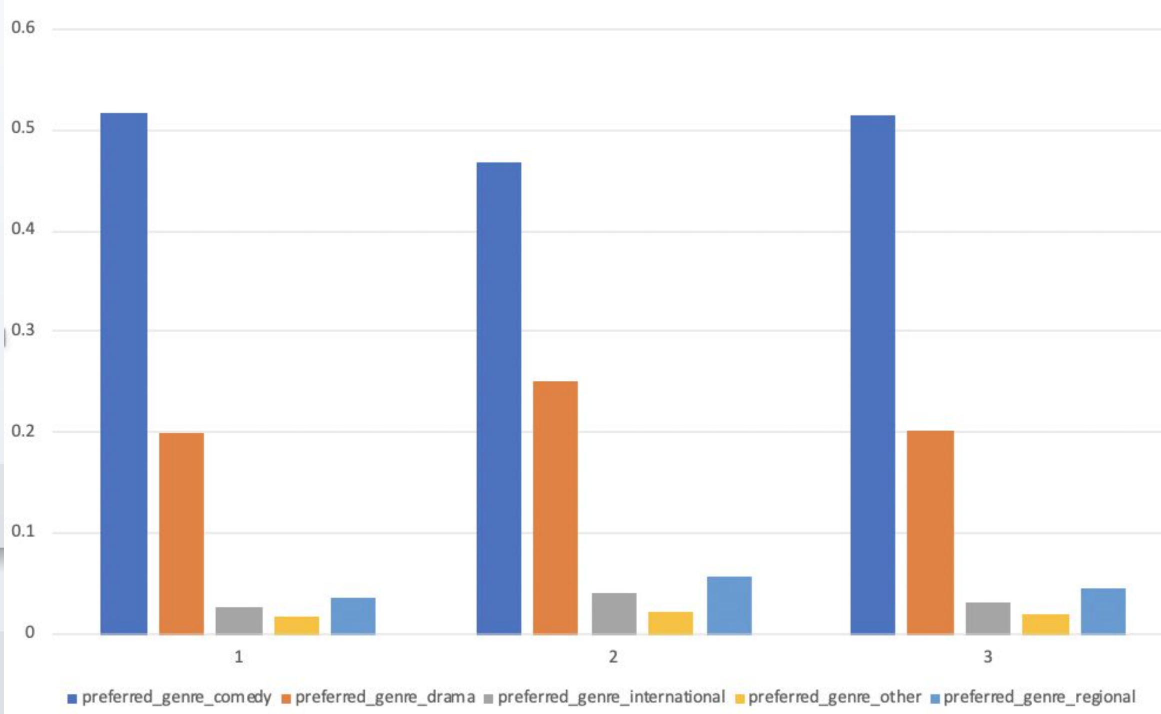
Group1:
Proportion of elderly group is higher than mid-aged and all the other group

Group2:
Proportion of mid-aged is higher than elderly and all the other group

Group3:
Almost the same but the proportion of elderly group is slightly higher

Analyses for Customer segmentation

Comedy and drama continues to be the most popular genres as we have found in the EDA. Generally, comedy is higher than drama in all groups.

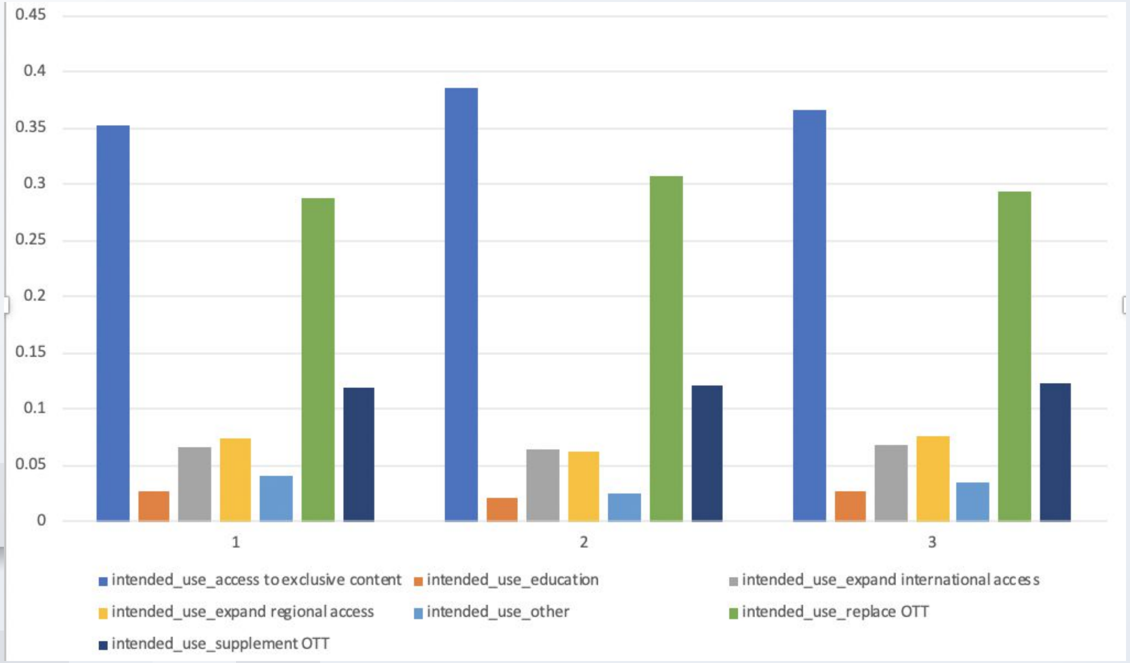


<u>Group2:</u>
Consume more drama and regional than the other groups

<u>Group1 & Group3:</u>
Behaves similarly and consumes comedy followed by drama.

Analyses for Customer segmentation

Exclusive content and replace OTT continues to be the most reasons for using our service and as we have found in the EDA.

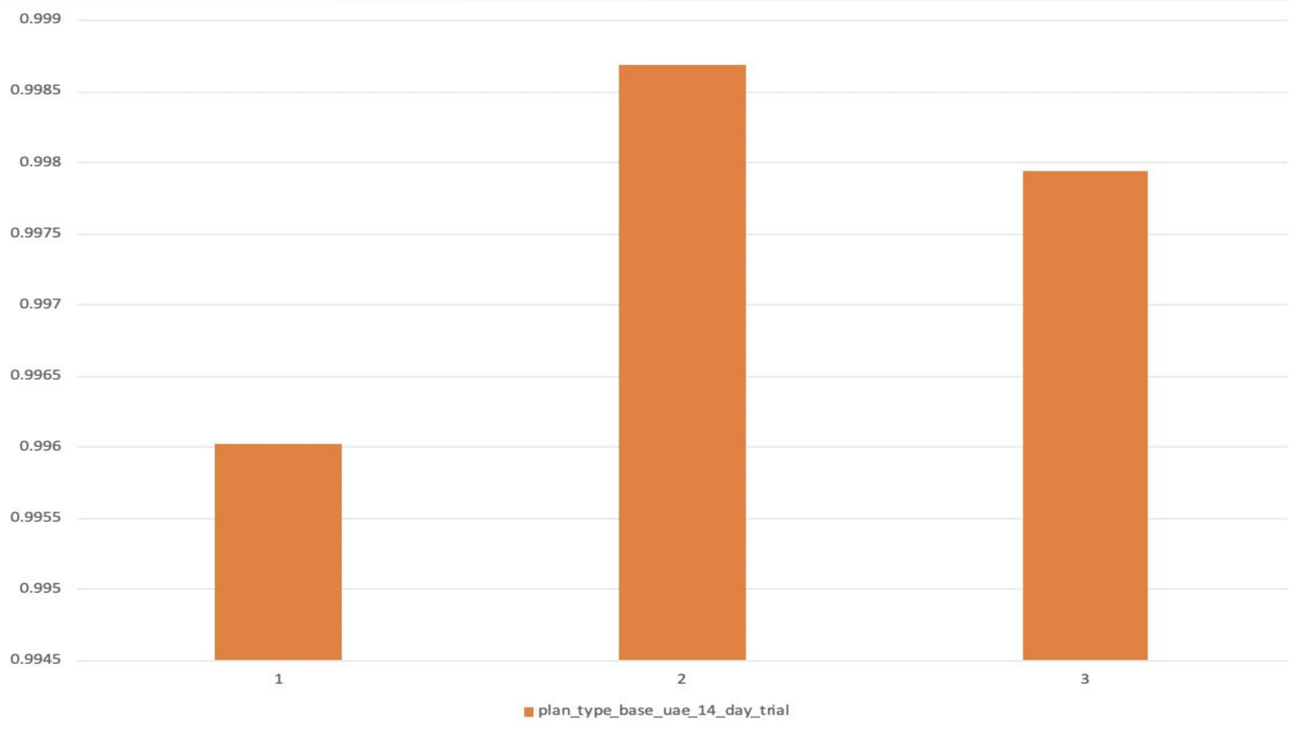


<u>Group2:</u>
Consumers came for the replacement of OTT more than all the other groups

<u>Group1 & Group3:</u>
Behaves similarly and the most important for using the service is exclusive content followed by replace OTT.

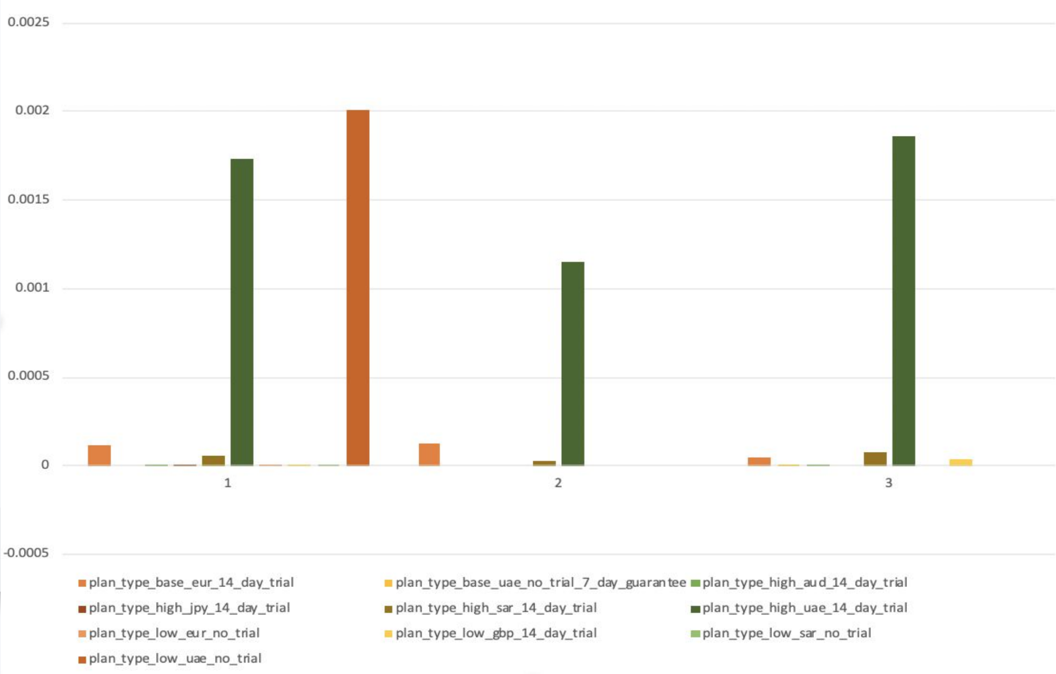
Analyses for Customer segmentation

Base_UAE_14_Days continues to be the most popular plan type as we have found in the EDA.



Analyses for Customer segmentation

As for the rest of the plan types, High_UAE_14_Days is the second popular plan type and it's negatively correlated to the consumption of the Base_UAE_14_Days. group 1 is distinguished by using Low_UAE_no_trial



Analyses for Customer segmentation

Our customer profile looks like.

- In age range of 35-70
- Came for exclusive content and the replacement of OTT
- Hold a base UAE plan
- Like drama and comedy
- Has an average consumption hour of 29.62
- Use IOS system

Analyses for Customer segmentation

Recommendations includes:

- Create packages consists exclusively drama and comedy
- Create packages for 30-hours-consumption
- Extend partnerships with exclusive content providers.
- Build exclusive marketing channel for IOS system
- Create female and male exclusive package (most customers are female but male generates more revenue)

Churn analysis

Data used:

Revenue_net_1month
Payment_period
Months_per_bill_period
Creation_until_cancel_days
Revenue_net
monthly_price
Discount_price
op_sys_iOS
Package_type_economy
Package_type_enhanced
Gender
Cancel
Paid
Refund

Churn analysis

Compersian between models
Logistic regression

```
accuracy_score(y_test,y_pred)
```

]: 0.8760705911390843

```
]: from sklearn.metrics import classification_report
   print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

       False       0.93      0.90      0.92     30648
        True       0.72      0.80      0.76      9867

    accuracy                           0.88     40515
   macro avg       0.83      0.85      0.84     40515
weighted avg       0.88      0.88      0.88     40515
```

Churn analysis

Compersian between models

Decision Tree

```
clf = DecisionTreeClassifier(max_depth = gsearch.best_pa
clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
accuracy_score(y_test,y_pred)
```

```
0.8767123287671232
```

```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

       False       0.98      0.85      0.91     30648
        True       0.68      0.95      0.79      9867

    accuracy                           0.88     40515
   macro avg       0.83      0.90      0.85     40515
weighted avg       0.91      0.88      0.88     40515
```

Churn analysis

Compersian between models
Random Forest

```
accuracy_score(y_test,y_pred)

0.8783907194866099

print(classification_report(y_test, y_pred))

              precision    recall  f1-score   support

       False       0.97      0.86      0.91     30648
        True       0.69      0.92      0.79      9867

    accuracy                           0.88     40515
   macro avg       0.83      0.89      0.85     40515
weighted avg       0.90      0.88      0.88     40515
```

Churn analysis

Compersian between models
GBDT

```
0.8936
```

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
False	0.95	0.91	0.93	30648
True	0.75	0.84	0.79	9867
accuracy			0.89	40515
macro avg	0.85	0.88	0.86	40515
weighted avg	0.90	0.89	0.90	40515

Churn analysis

Decided on GBDT because of the higher model accuracy and we can get feature importance, use it as a reference and visualize the branches and results. Fairly robust to overfitting, doesn't require careful normalization or scaling of features.

Revenue_net_1month	0.748
Payment_period	0.057
Months_per_bill_period	0.023
Creation_until_cancel_days	0
Revenue_net	0.083
Monthly_price	0.066
Discount_price	0
op_sys_iOS	0.001,
Package_type_economy	0.001,
Package_type_enhanced	0.009,
Gender	0.012,
Cancel	0
Paid	0
Refund	0

Price related features played an important role in terms of making predictions

Lowering the price can be effective when trying prevent customer churn given the current results based on modeling.

Churn analysis

Then we generate the churn probability of the selected model to prepare for the calculation of CLV.

```
pd.DataFrame(churn_prob_gbdt).desc
```

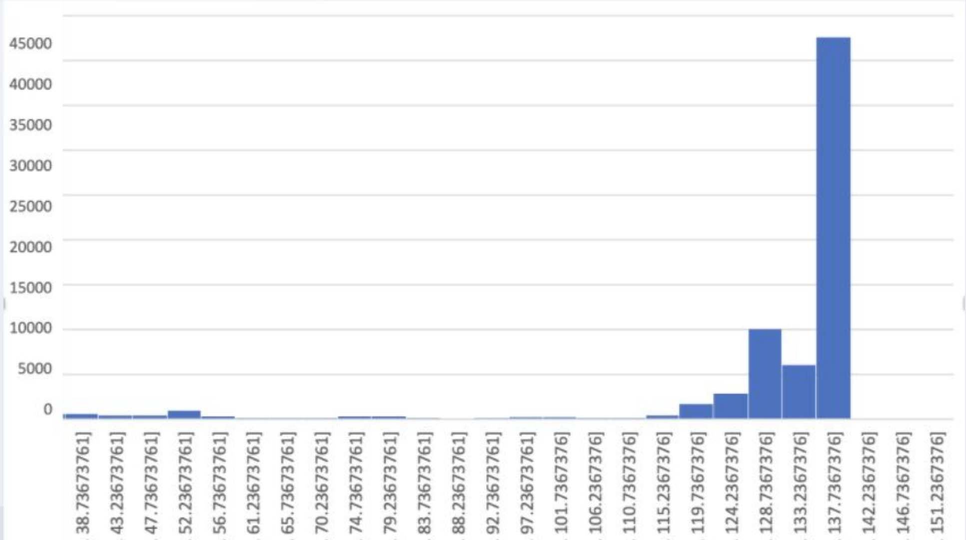
	0
count	135050.000000
mean	0.239082
std	0.335396
min	0.000422
25%	0.000583
50%	0.000598
75%	0.558605
max	0.998970

As we can see from the predictions generated from GBDT model, the customer has a relatively low churn probability.

CLV analysis

Group the customer by entrance to sign-up form captured by product and also the month they signed to get the cost of different channels.

Then use 'monthly_price'* ((1+r)/(1+r-(1-'probofchurn'))) - ('monthly_price') to calculate future price and use net revenue+ future revenue - cac to calculate clv. Distribution is as follows, the clv has a mean of 85.50084918636045 and a median of 128.64801477068605



Given the low churn probability and a relatively high CLV, we can conclude that the uncertainty of future payments given the current customer base is low.

Conclusion

- AB Testing and model provides merely predictions, actual decision should be made based on the reality of the business and future developments.

- This result can give us a guidance on resource allocation in our business campaign , we can target customers with less uncertainty and avoid missed opportunities.

THANKS!

Any questions?