Gartner HackElite Submission

Name: R. SHRI PRASAD
Institute Name: IIT - MADRAS
Roll No.: NA15B043

Gartner®

Client Retention

What's better than acquiring a new customer?

Although a little counter-intuitive, answer isn't "acquiring another customer". It's actually retaining an existing one.

But, why?

It costs 5x to acquire a new customer, than it does to retain one [1]

It’s 50% easier to sell to existing customers than it is to new prospects. [2]

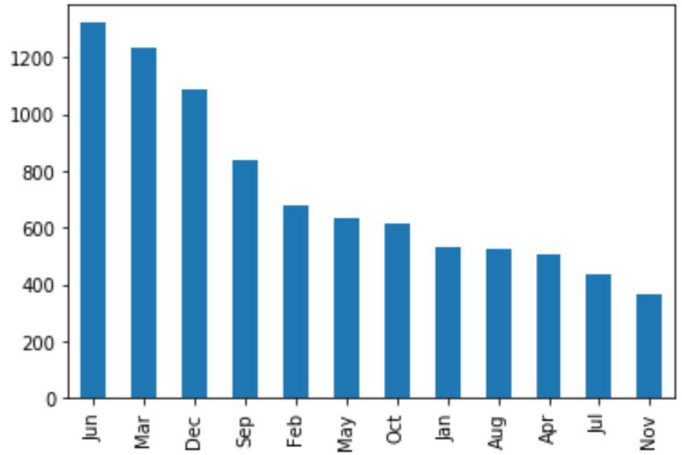
A 5% increase in customer retention can increase a company’s revenue by up to 95% [3]

1- https://www.invespcro.com/blog/customer-acquisition-retention/
2- https://www.amazon.com/Marketing-Metrics-Definitive-Measuring-Performance/dp/0137058292
3- https://hbswk.hbs.edu/archive/the-economics-of-e-loyalty

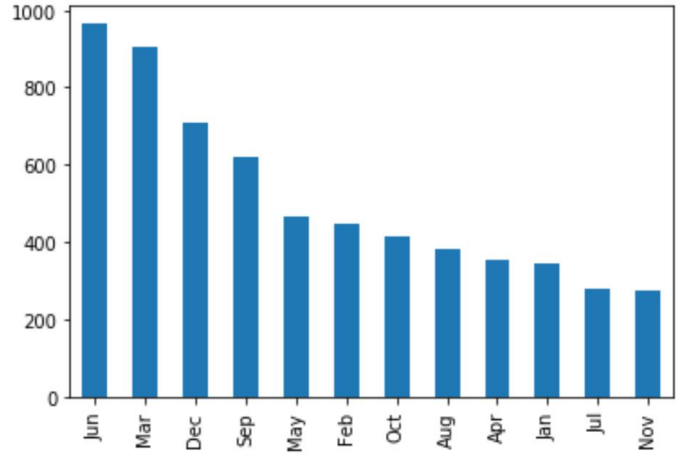
Gartner

Analysis and Feature Engineering

Its pretty interesting to know that Clients start their subscription usually at the end of fiscal year quarters i.e June, March, Dec & September :



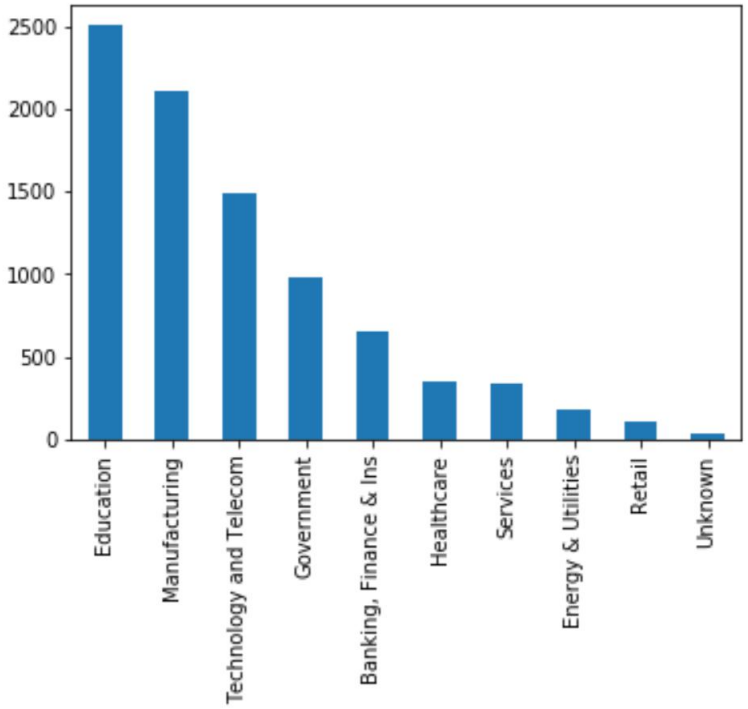
Clients who were retained at the end of the year also displayed a similar trend :

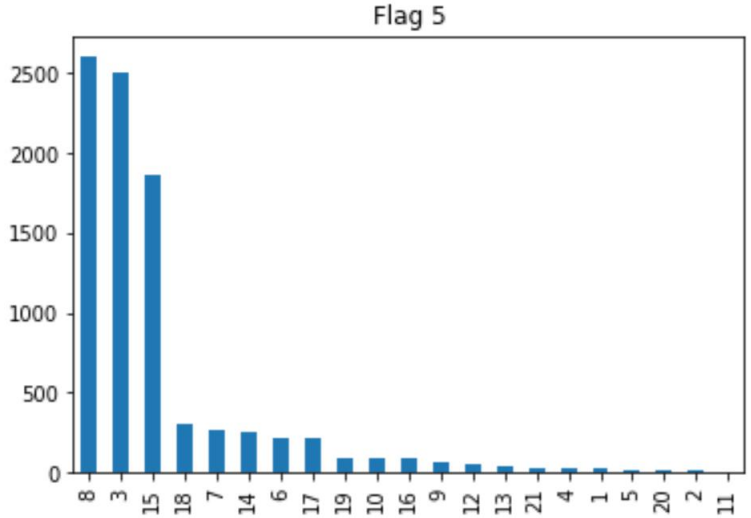


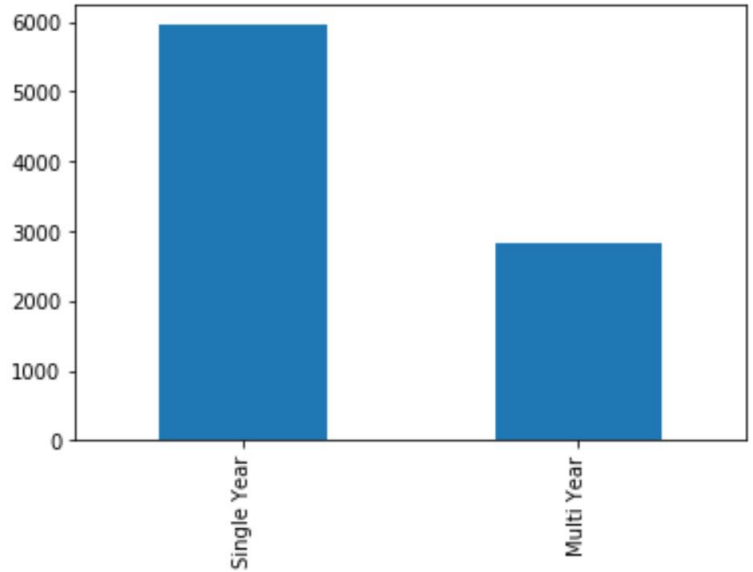
Analysis and Feature Engineering

Other notable Observations:

- Most of Gartner's clients are in the Education, Manufacturing and Technology & Telecom sector
- Clients are majorly from Regions 8, 3 and 15
- Clients usually go for Single Year subscriptions, rather than long term Multi Year subscriptions





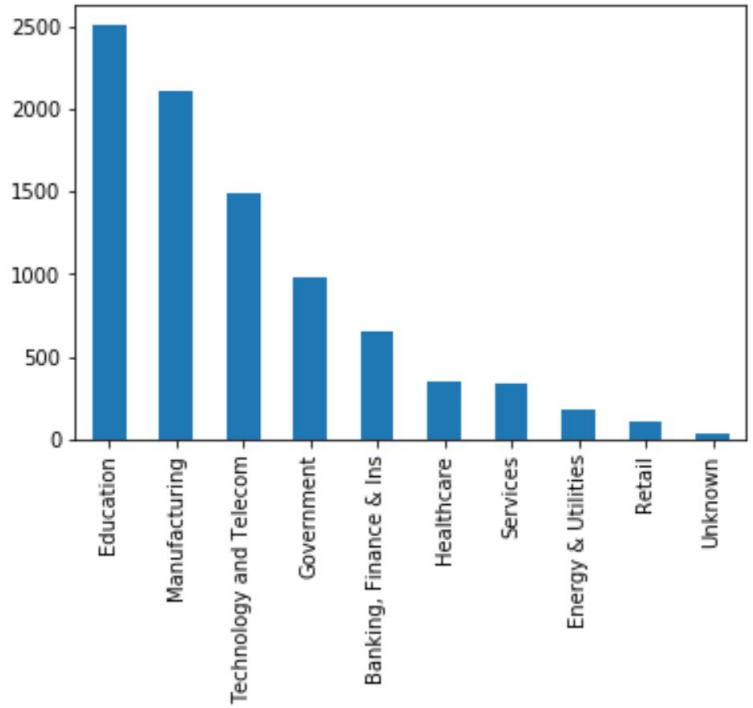


Analysis and Feature Engineering

Dataset at first glance had no explicit missing values. But upon closer inspection, it was clear that they had been masked as 'Unknown'

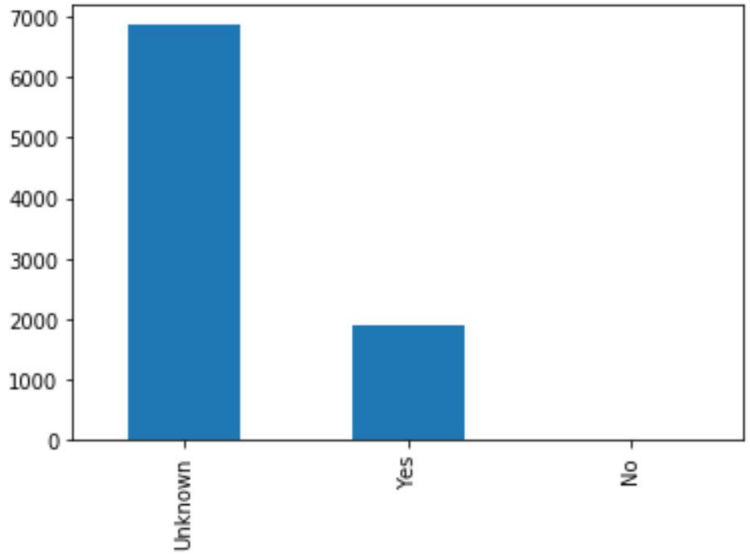
```
data1['Flag 6'].unique()
```

```
array(['Manufacturing', 'Government', 'Technology and Telecom',
       'Banking, Finance & Ins', 'Education', 'Energy & Utilities',
       'Healthcare', 'Services', 'Retail', 'Unknown'], dtype=object)
```




```
data1['Flag 4'].unique()
```

```
array(['Unknown', 'Yes', 'No'], dtype=object)
```



Analysis and Feature Engineering

An interesting observation was that all the 'Unknown' labels in 'Flag 6' have one thing in common: their Region (aka) 'Flag 5'

```
data1[data1['Flag 6']=='Unknown']
```

	Client ID	Company ID	Client Contract Starting Month	Flag 1	Flag 2	Flag 3	Flag 4	Flag 5	Flag 6
1520	10004520235	6293775	Jun	Single Year	2	Yes	Yes	21	Unknown
1691	10004520565	6294502	Mar	Single Year	3	No	Yes	21	Unknown
1788	10004511236	6296172	Dec	Single Year	3	Yes	Unknown	21	Unknown
2176	10004517346	6292443	Dec	Single Year	2	No	Unknown	21	Unknown
2391	10004512563	6295159	Dec	Single Year	3	No	Unknown	21	Unknown
2590	10004521834	6292814	Mar	Multi Year	6	Yes	Yes	21	Unknown
2800	10004519928	6294416	May	Multi Year	6	No	Yes	21	Unknown
3442	10004513813	6293470	Jun	Multi Year	6	Yes	Unknown	21	Unknown

Analysis and Feature Engineering

There also seems to be a solid correlation between Region and the Industry of the Client:

```
print(dataset_1[dataset_1['Flag 5']==17]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==16]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==15]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==14]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==13]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==12]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==11]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==10]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==9]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==8]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==7]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==6]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==5]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==4]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==3]['Flag 6'].unique())
print(dataset_1[dataset_1['Flag 5']==2]['Flag 6'].unique())
```

```
['Services']
['Government']
['Technology and Telecom' 'Healthcare' 'Manufacturing' 'Retail']
['Banking, Finance & Ins' 'Energy & Utilities' 'Healthcare'
 'Manufacturing']
['Retail' 'Services' 'Technology and Telecom']
['Energy & Utilities']
['Technology and Telecom']
['Banking, Finance & Ins']
['Banking, Finance & Ins']
['Manufacturing' 'Government' 'Banking, Finance & Ins']
['Government']
['Government']
['Technology and Telecom']
['Banking, Finance & Ins']
['Education']
['Technology and Telecom' 'Manufacturing']
```

Analysis and Feature Engineering

There are some categories of 'Flag 2' in Train dataset that aren't there in Test - and vice versa :

```
print(train['Flag 2'].unique())
print(test['Flag 2'].unique())
```

```
[ 2  7  6 10  3  4  1  5 13  9]
[10  2  6  7  4  3  8 13 12  5 11]
```

- Presence of unseen / novel categories in Test set might confuse the model that has been trained on the Train set.
- Also, these Outlier categories (8, 11 & 12 in Test) only have a single entry under each of them.
- So it's safe to impute them with the 'Mode' of Flag 2 i.e category 2.

'Flag 3' had a couple of mislabelled entries :

- Imputed them with 'Yes' by observing other entries that had similar Company ID and Flag values

```
data1['Flag 3'].unique()
```

```
array(['Yes', 'No', 'C'], dtype=object)
```

```
dataset_1[dataset_1['Flag 3']=='C']
```

	Client ID	Company ID	Client Contract Starting Month	Flag 1	Flag 2	Flag 3
7444	10004512577	6292260	Oct	Multi Year	13	C
306	10004519806	6295793	Oct	Multi Year	13	C

2 rows × 106 columns

Analysis and Feature Engineering

Dataset has Activity values for every month of the year. Intuitively, the dataset would be more interpretable if we knew the recency and consistency of these values over the subscription period.

Having this in mind, I've engineered features to display the client's total activity in the 1st, 2nd, 3rd & 4th quarters of their subscription.

This would tell us: 1) How consistent they have been over the year

2) How active they have been at the start & end of the subscription period,and if that can explain the retention of client

To confirm the intuition, these features had high levels of Feature_importance in the prediction model

```
In [218]: X_train_extra
```

Activity 1 4th Quarter	Activity 2 4th Quarter	Activity 3 4th Quarter	Activity 4 4th Quarter	Activity 5 4th Quarter	Activity 6 4th Quarter	Activity 7 4th Quarter	Activity 8 4th Quarter	Activity 1 3th Quarter	Activity 2 3th Quarter	Activity 3 3th Quarter	Activity 4 3th Quarter	Activity 5 3th Quarter	Activity 6 3th Quarter	Activity 7 3th Quarter	Activity 8 3th Quarter	Activity 1 2nd Quarter	Activity 2 2nd Quarter	Activity 3 2nd Quarter	Act 4 Qu
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	1	0	
0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	14	3	0	1	0	0	0	0	12	3	0	
12	0	0	0	0	0	0	0	10	0	0	1	0	0	0	0	62	0	0	
2	0	0	1	0	0	0	0	6	1	0	1	0	0	0	0	2	0	0	

...and so on

Approach for building model Structure

I've used solid Gradient Boosting algorithms for the prediction, namely XGBoost & CatBoost.

Gradient boosting does very well as it is a robust, sequentially built, out-of-the-box classifier that can give tremendous results on a dataset with minimal effort spent on data cleaning & scaling.

Other advantages include : 1) Learns complex non-linear decision boundaries with ease.

2) Offers a lot of flexibility on optimizing different loss functions

3) Provides a wide range of hyperparameter tuning options that make the function fit very flexible.

APPROACH:

First tried **XGBoost** with extensive hyperparameter tuning using GridSearchCV. Plugged in the best combination and obtained an F1 score of 88.52 on the portal.

Next tried **CatBoost** - again with extensive hyperparameter tuning. Plugged in the best combination and obtained an F1 score of 92.90 on the portal.

Proceeded with **CatBoost**.

Approach for building model Structure

```
accuracy_score on train dataset :  0.9704791785510554

accuracy_score on test dataset :  0.9321550741163056
```

Snapshot of the accuracy score of the optimum CatBoost model on train and test set

	precision	recall	f1-score	support
0.0	0.92	0.86	0.89	546
1.0	0.94	0.96	0.95	1208
micro avg	0.93	0.93	0.93	1754
macro avg	0.93	0.91	0.92	1754
weighted avg	0.93	0.93	0.93	1754

Snapshot of the classification metrics of the optimum CatBoost model on the test set

NOTE : In binary classification, recall of the positive class is also known as “sensitivity”; recall of the negative class is “specificity”

What are the most important activities that will impact client retention?

The most important activities that impact client retention significantly :

ACTIVITY #1 : More the documents the client reads, the better he recognizes the ingenuity and committment of Gartner

ACTIVITY #2 : Social media views- More the views, more is the exposure clients get through Gartner.

ACTIVITY #3 : Frequent dedicated sessions with Analysts can build a strong customer loyalty and relationship

ACTIVITY #4 : Only if a client is updated with all the perks & features of a product or a service, he can get the most out of it. If a client isn't using certain modules that is tailor suited for them, we must show them how it works or provide the support they need.

ACTIVITY #5 : Again, higher the interaction is with the client, stronger is the trust and commitment they have towards Gartner. Also,personal 1:1 meetings let the clients know Gartner takes every client seriously

ACTIVITY #8 : Testimonials are given by satisifed clients telling how Gartner had helped solve their problems. More the testimonials given indicate higher satisfaction which is ultimately an indication of client retention

NOTE : These conclusions are made from the Machine Learning model's Feature_importance attribute

What should be the order of these activities in client contract life cycle?

Link to the screenshot of Feature importances : https://bit.ly/2HEI8Yh

Which months are most important for client engagement for driving higher retention?

ACTIVITY#	SIGNIFICANT MONTHS	NOTE
1	0,1,2,3,4,5,6,7,8,9,11	Activity #1 is the activity with the most impact. All the months are important for client engagement
2	9,4,8,10,5,3	Boost in the social media views in the last few months can compel the client to continue the subscription
3	2,9,1,0,8,3,10	Frequent interaction with the client throughout the period can have a strong impact on retention
4	0,9,6,11,3,4,8,1,5	Updating the client with all the perks & features in the first month itself (month 0) can enable the client get the most out of Gartner
5	2,10	Personal 1:1 meetings in the first few months and the last few months strengthens the possibility of retention. But ideally, it should be done as frequent as possible,
6	2	Not so significant in determining retention
7	1,6	Not so significant in determining retention
8	11,2,1,7,4	Testimonials given at month 11 implies that the client ended the subscription with satisfaction. This would've compelled him to continue being a part of Gartner

NOTE : MONTHS ARE IN THE ORDER OF SIGNIFICANCE

Gartner

What activities should service associate recommend in first month to drive higher engagement in subsequent months?

Activities to lay emphasis on in the first month :

ACTIVITY #1 : Make the client reads enough relevant documents in the first month

ACTIVITY #4 : Make sure the client is up-to-date with all the perks and services

ACTIVITY #3 : Generate an Inquiry with an Analyst

Recommendations

The activity data that is available at present is already elaborate enough. But the following recommendation may help :

- Loyalty programs seem simple, but they can have a huge impact on customer retention.
- Set up automated emails to different types of customers when certain conditions are met, such as when they've gone 30 days without contact.
- Try to get as much product usage data as we can so that we can be more informed about our customers' needs. If a customer isn't using certain modules that would be super useful for them, we can then show them how it works or provide the support they need.
- **Perform A/B tests**:

  There's nothing wrong with making an assumption.We hypothesize that sending out weekly emails to existing customers with in-depth content will increase customer retention rates. Maybe the customer retention rates increase. Maybe they don't. There's only one way to find out.

Thank you!

Gartner®