Gartner HackElite Submission Template

Name: Pranav Pawar
Institute Name: IIT Madras
Roll No.: MM17B003

Gartner®

Recommendations

According to the facts, If you're in the subscription business, you're in the customer service business. Strong customer relationships are at the core of the subscription business model. Without them, there can be no sustainable recurring revenue growth. As a company's customer base gets bigger, this becomes one of the most important keys in the entire framework. Acquiring new subscribers is critical, but in the Subscription Economy the vast majority of customer transactions consist of changes to existing subscriptions: renewals, suspensions, add-ons, upgrades, terminations, etc.

Based on my predictive analysis on available data I believe that following are some important recommendations which should be considered by Gartner -

1. Document reading should be recommended the most, read count directly correlated with the retention rate
2. Second aspect of retention is views on the social media, the more interactive the social media pages Gartner develop the more people will join the culture and might get benefited by the service, such customers are most likely to get retained
3. Number of calls done by Gartner service help desk - The more you care about the customer the more the increase in customer satisfaction rate. Many customer struggle with the service available because of the compelxities, but a help desk call might help out such customers and which will lead to maximum customer satisfaction
4. Gartner should focus on these activities mostly in the first three months of subscription period, because this is when the customer is in a raw state, impression made in initial period might last long, and we need to make sure that we deliver our best!

Analysis and Feature Engineering

Analysis Pipeline:

1. Hypothesis generation
2. EDA (Exploratory Data Analysis)
3. Feature Engineering
4. Building Model
5. Cross Validation
6. Parameter tuning

Hypothesis Generation: - I have designed a hypothesis at starting phase without seeing the data that Client will be retained if client was engaged in more activities like reading the documents, activity on social media, discussion with consultant, analyst of Gartner team in the starting phase of his subscription rather than the presence of clients in conference meetings. Conference meetings and symposium meetings comes after certain time period so it will not be significant feature for retention of client.

Exploratory Data Analysis: -

- Initially summary statistics of dataset (count of numerical variables and categorical variables)
- skewness and kurtosis of numerical variables (to see whether the variables are skewed or not) and cross tabulation of categorical variable with respect to response variable (to see counts of levels in “”Yes” or “No”)
- plotted the distribution of numerical and categorical variables
- missing values handling (No missing values were present in the dataset)
- to check multicollinearity plotted the correlation plot and calculated the Variance Influence Factor (VIF)
- Feature Scaling

Analysis and Feature Engineering

Feature Engineering -

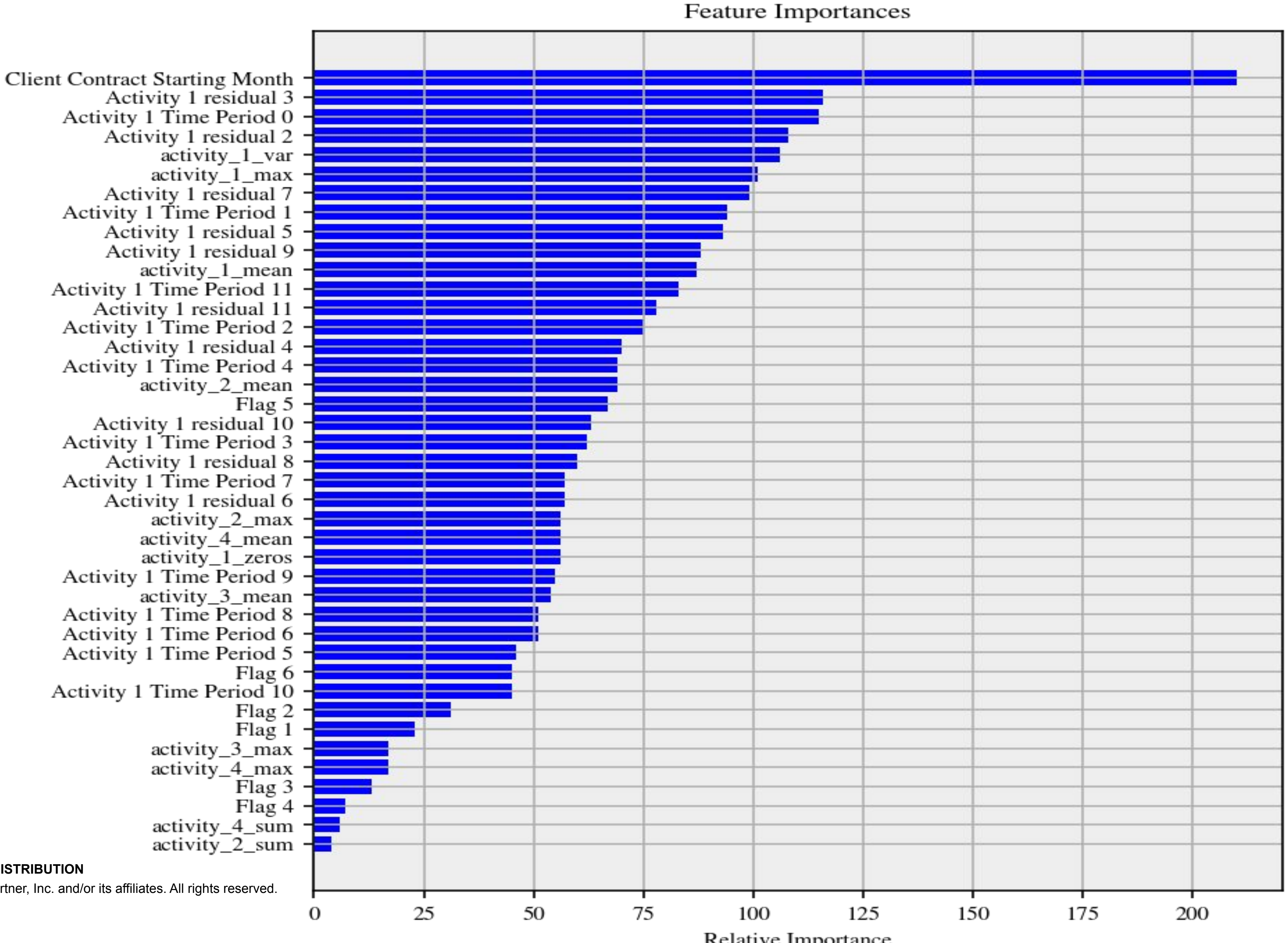
1. Statistical Features -
   a. I created several statistical features from the monthly activities. These feature includes, sum, max, min, variance of particular activity across the month spectrum, turned out to be most important features
   b. Later it was observed that stats of only activity no. 1,2,3,4 were making an significant impact on the model, so I considered only these features, the feature pool is as follows
   c. 'activity_1_mean', 'activity_1_max', 'activity_1_var', 'activity_1_zeros', 'activity_2_mean', 'activity_2_max', 'activity_3_mean', 'activity_3_max', 'activity_4_mean', 'activity_4_max', 'activity_2_sum', 'activity_4_sum',

2. Residual Features -
   a. Residual features basically calculates the difference in activity of sequential months,
   b. And these were one of the most important features identified by the model
   c. So for example activity_1 residuals will be differences between its values in consecutive months. And Activity 1 residual 2 means (activity1_month_2 - activity1_month_1)
   d. Only activity 1 residuals were significant, and they were considered for model buidling
   e. The final feature pool considered is as follows
   f. 'Activity 1 residual 2', 'Activity 1 residual 3', 'Activity 1 residual 4', 'Activity 1 residual 5', 'Activity 1 residual 6', 'Activity 1 residual 7', 'Activity 1 residual 8', 'Activity 1 residual 9', 'Activity 1 residual 10', 'Activity 1 residual 11'

3. Further statistical features (max, min, mean, var) of residuals were considered but they turned out to be overfitting the model, so this idea was dropped
4. Feature importance can be seen in next slide

RESTRICTED DISTRIBUTION

Analysis and Feature Engineering



RESTRICTED DISTRIBUTION

Approach for building model Structure

Model Experiments -

1. **Baseline Model - 5 Fold LightGBM**
   a. **Validation scores - (sklearn.model_selection.train_test_split, train_validation_ratio = 0.2)**
      i. Training accuracy = 0.9851682829435254
      ii. Validation accuracy = 0.8774230330672748
      iii. Validation F1 Score = 0.9166989538938396 (Sensitivity = 0.8974, specificity = 0.7379)
      iv. Validation ROC AUC Score = 0.9331437034250922
      v. Test public leaderboard F1 Score = 0.8803

2. **Weighted ensemble of 3 LightGBM + 3 XGBoost**
      i. Validation accuracy = 0.8945267958950969
      ii. Validation F1 Score = 0.9276729559748428 (Sensitivity = 0.8981, specificity = 0.7378)
      iii. Test public leaderboard F1 Score = 0.8847

3. **Why LightGBM -**
   a. I had used random forest initially, which gave poor results because of a bit sparse nature of data, and it was not able to minimize the error, hence decided to use boosted decision trees which mainly focuses on misclassified data points, leading to the lowest error
   b. LightGBM is an advanced version of boosted decision trees,
      i. Faster computing wrt XGBoost, GBM
      ii. Penalize misclassified data points more to obtain the precise predictions.
      iii. Also, presence of only few decisive parameters makes the model easy to tune.
      iv. **Lower memory usage:** Replaces continuous values to discrete bins which result in lower memory usage. Considering my machine’s hardware specs, LightGBM seems to be the best option to use.

Gartner®

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

The most important activities are

1. Number of documents read by client,
2. Number of views on social media of Gartner
3. Number of inquiries generated by client in starting period of contract.
4. If Client is involved in such activities in the first three months of subscription then there is better chances is that he will be retained.

In terms of business perspective, Gartner should focus on these activities mostly in the first three months of subscription period, because this is when the customer is in a raw state, impression made in initial period might last long, and we need to make sure that we deliver our best!

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

This is the descending order of activities (Max at top)

- ➔ Number of documents read by client
- ➔ Number of views on social median of Gartner
- ➔ Number of inquiries generated by client
- ➔ Number of calls done by Gartner service help desk
- ➔ Number of symposiums attended by client
- ➔ Number of 1:1 meetings done by Gartner Consultant
- ➔ Number of testimonial shared by client
- ➔ Number of conferences attend by client

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

First 4 months from the subscription start are the most important months in terms of driving higher retention. As it seems obvious that any client will lead to judge the service very quickly and to transform them into happy customers we need to deliver best service in initial months

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

Most important activities -

1. Activity 1 (Documents read)
2. Activity 2 (Views on social media)
3. Activity 4 (Inquiries generated by Client)

Reasoning -

1. According to the data analysis, these are highly correlated features with the target
2. Considering activity 1, it sounds legible that more the read count and views on social media, the more the customer interaction on the platform. This directly relates that customer is pretty happy with the service and is exploring more everyday
3. The 4th Activity which is inquiries generated is also important, because the client might be facing issues with the service and resolution of such issues is the key factor in the customer’s subsequent engagement