Hikeathon Solution

Team Name - Kamakal Vectors

Summary

Problem Statement : Predicting the probability of two users chatting

Data Size: 80 million + records with more than 8 million users

AUC numbers:

- 0.9415 on the public LB
- 0.9409 on the private LB
- 0.9402 local CV

Undersampling is the way to go ….

- Our Solution heavily depends on Negative Undersampling
- We ended up using around 5% of the negative examples (is_chat==0)
- We didn’t notice any drop in performance
- This approach helped us on two fronts:
  - *Rapid Experimentation*, we were able to test out many different sets of features because of the reduced training time
  - Final Solution was a *blend* of multiple models trained of different 5% of the negative samples. This helped us in getting a more *robust and better model*

Secret Sauce ----> Create More Features

Feature creation played a very important role in our climb up the leaderboard, we used three major sets of features in our final model:

- User Activity Features
  - We were given a user activity matrix in 13 dimensions, which was used to create metrics on similarity of users based on activity
- Graph Features
- User Social Circle Features

Graph Features

We created three graphs from the data set:

- Undirected Contact Graph
- Directed Contact Graph
- Chat Graph (capturing the data with is_chat=1)

Some of the metrics used included:

- Jaccard coefficient
- Resource allocation index
- Degrees of nodes(in case of directed graph, in/out degrees)

User Social Circle Features (1/2)

- These variables proved to the **most crucial for boosting model performance**
- ***Number of mutual nodes*** was calculated between the node pairs. Higher this number, more the chances they interact
- ***With how many mutual nodes does each node pair interact?*** Let's take an example:
  - We have to find the chat probability between A&B. X,Y,Z are the mutual contacts between A&B
  - Now if A&B are chatting with all three then there is a higher chance they will chat with each other
  - If they are not chatting with anyone, then it's highly likely that X,Y,Z are customer care numbers :p

User Social Circle Features (2/2)

- ***How many time is each node involved in a chat?*** For A and B to have a higher chance of chatting, they need to be chatting with other people too

- ***How chatty is the neighbourhood?*** If this number is high for both the nodes, then it indicates they are part of a more talkative neighbourhood and hence higher the chances of chatting.

- ***Inverse Links***, this features captured the inverse relationships present in the data shared across train and test.

Final Model

- The final model is an ensemble of 10 LightGBM Classifiers with each model fitted over a five-fold random stratified CV
- Each of the above model was built on a different subset of data with negative undersampling