Project Life in Red

Capstone Project Executive Summary for fulfilling the requirements of the Master of Sciences in Business Analytics at the Central European University

Student: András Ákos Somkuti
As of: June 24, 2020

Project Background

Project **'Life in Red'** is an attempt at solving an unsupervised learning problem combined with general data analytics on the transaction chat log data set from the website livejasmin.com. LiveJasmin.com is one of the most visited websites on the internet - currently ranked #37 by Alexa - and offers live video streams of performers from around the world, who are available to engage - with the promise of nudity - in private video chats for a fee per minute. Members can buy credits on the website after registration, each performer can set their fees per minute and members can use their credit balance to enter these private chat rooms. As the chat progresses, credits are progressively deducted and the members can also send the performers surprises, which are not unlike tips in the real world. Surprises and the time of the chat multiplied by the rate make up the total spend by the member on the transaction, which is then shared between the site and the performer.

The project scope is focused on answering the following questions:

1. How much of the time that is spent chatting is sexual in nature?
2. Are there identifiable topics and if so, what are members and performers talking about?
3. Is there a relationship between the content of the topics and the success of a performer on the site or the members' spending patterns?
4. Are there any other practical observations about the emotional content of the conversations?

Answering these questions could inform the public at large about what this community is and what it is not, further the company's understanding of the drivers behind demand and the best ways to grow it and educate performers on best practices to maximise their potential in the business.

Methodology

A random five per cent sample of all the livejasmin chat logs from the past year was created, which was then subsequently cleaned and enriched. The chat logs had to be broken up into lines and classified if the transaction was in writing (at least by one participant throughout the entire chat) or at least partially exclusively over audio (meaning that neither participant used typing exclusively throghout the chat, both at some point used audio).

Language was detected and metadata from the chats was extracted, like number of words and lines, sentiment per line and per chat, bag of words, timestamps, length of chat and many others.

Term frequencies were analysed and latent dirichlet allocation was used to identify topics in the chats. Two different approaches were taken, the first was looking to identify sexual vs non-sexual topics of

conversation, while the other looked at larger number of topics, finally identifying 10 for analysis. The topic probabilities per document were added and topic probabilities were multiplied by the length of the chats in terms of time to understand the weight of each. All of these features were then aggregated by member and by performer to do in-depth analysis of them using three fold cross validation with linear regressions, LASSO, random forests and xgboosting.

Results

The analysis has shown that the vast majority of chats is very short. In fact this could provide the impression that the site is primarily about the short in personal interactions.

The more in-depth look is exactly the opposite. These short interactions, albeit numerous, account only for a small percent of the time spent chatting on the platform, five minute or shorter chats represented half of all chats, but not even 17% of the billed minutes. To understand the site and the ecosystem one mustn't look at the median transaction and averages are meaningless. In fact, the distributions on the site be it around spending patterns, length of chat and others follow Pareto law and have fat tail distributions with large variances.

Chat length is obviously important. Longer chats mean more revenue. Interestingly, longer chats mean less surprises by the minute, however this is mostly made up by the surprises being larger in size. Also as the length of the chats increases, communication intensity does so as well, words per minute increase from 5 to 20, which indicates strong cadence in longer chats. More intense communication turned out to positively correlate with spending.

In longer chats, performers type more and faster as well, while members while typing faster also use audio more frequently. Longer chats had more intense emotions across all verticals, with joy, anticipation and trust leading the pack. Longer conversations had disproportionately higher ratings for trust, anticipation and fear, while lower proportional values (but higher absolute values per minute) for joy, anger and disgust. Indicating deep, intense conversations and less vulgarity and sexuality.

This is confirmed by the topic split. While 58% of the time is focused around sexuality as chats get longer this goes down from an initial 80% for short chats, to sub 20% for long ones, while members start spending progressively more credits per minute on those interactions, indicating the highest value-add that a performer can provide is the building and maintaining of deep, meaningful and lasting human relationships, fostering genuine feelings of mutual belonging, bonding, trust and love.

The more sex is center stage in the conversation, the less money is made. Not only does the data show a strong negative correlation between the length of chats and the topic of sex, the reduction in rates shows that average transaction size can be up to 80% less as sexual topics get introduced.

A detailed break down of topics with building ten instead of just two provides a more granular understanding of the ecosystem. While the breakdown still shows that sexual topics have a large share of time spent, it also shows that sleeping is the highest priced activity. It shows that not only do the most successful models get paid for sleeping, they get a higher rate for sleeping than others get for performing hardcore sexual roleplaying.

Transaction sizes show correlation with general topics. However small talk could be identified as a separate topic and the analysis shows that while talking about non sexual topics is great, it must be meaningful. Small talk in fact had a stronger correlation with small earnings than sexual topics did. Deep emotions, general topics like music, movies and family and sleep were the clear winners in terms of earnings, while smalltalk, BDSM, hardcore sex chat and striptease shows were predicting lower earnings.

In fact, a 10% higher proportion of the time that a member spends chatting about general non-sexual topics, correlates with them spending up to 1500 USD more on the platform on average.

Topics also showed clear and discernable emotional splits. Hardcore sex chat had a lot of disgust and anger, while compliments were nearly exclusively about joy and deep conversations about anticipation, trust and joy.

Modelling these emotions and topics was done on the aggregate data for models and performers and showed that boosting had the lowest root mean squared error, with about 5000 USD and 4993 USD respectively. This is really large, meaning that prediction is not working very well based on the content of the chats only, meaning that the fat tails are very hard to predict. Classification models could significantly improve these results as the biggest value is in identifying the large spenders early and maximising the wallet share, thus predicting exact spending is not so important.