Ben Le Chi
June 28, 2020

Capstone Project Report

1. Problem statement

The goal of the project is to use users generated data to build a model that predicts if a user is going to leave the game after a specific period of time the user stays in the game.

2. Background on the subject

Mobile gaming industry is one of the most competitive market out there with approximately 7000 games being released on Google Play and 2000 games being released on Apple iTunes on daily basis. Users can easily switch from one game to another of the same genre, it is important for game developers to know when a users is going to leave or what makes the users leave so that measures can be taken to prevent users from leaving and potentially increase revenue from existing users before they actually leave. Therefore, I wanted to use the data science techniques to help game developers classify if a users is leaving the game or not after four days in the game.

The issue is not new, however, I could not find any specific project that works on the mobile game especially social casino type. There were other similar projects on casual games.

3. Data Set

The data set are two csv files namely users play log and transactions records of a social casino game called LengBear - published exclusively for Cambodia market. The game has more than 1.2M downloads. The play log is collected from 01 to 13 May 2020 which has 1,768,639 rows from 241,713 unique UserID. The transaction detail file has 35,680 transactions from 1st March 2020 to 15th May 2020.

Player details have the following columns: 'Sequence','UserID', 'GameID', 'Level', 'WinNo', 'DrawNo', 'LostNo', 'WinAmt', 'LostAmt', 'Date', 'Currency_Type1', 'Currency_Type2'

Transaction detail has the following columns: 'UserID','Amount','Chips','Date', ‘Channel’.

4. Summary of the preprocessing, feature engineering and any other data cleaning/transformation, and exploratory data analysis (EDA)

I spent a bulk of my time on exploratory data analysis in order to understand the data completely. Here are the things I found.

4.1 Exploratory Data Analysis

Firstly, Saturday is the day where users spend the most, it can be that they have more time to spend during weekend and play more, thus ended up paying more. Promotion should be shown heavily on Friday night and Saturday to increase revenue.

Secondly, the game revenue started declining when people went into lockdown due to Covid-19 pandemic as shown in figure 1. This was highly counter-intuitive as more people were forced to stay indoor they should be spending more money online. Other large game company such as Electronic Arts or Sony has been reported record sales, why would it be different for people in Cambodia?

I went on to investigate the payment channels and found something very interesting. In developed world, credit cards are widely available users can pay for in-game items so easily. However, in many parts of the world (Cambodia is one of those), there is a huge unbanked population who do not have a way of paying to Google or Apple. These users leverage direct carrier billing and local e-wallets to pay into the game. This is very interesting as the game owners can save some profits under the profit sharing scheme with Google Play Store or Apple Store. And the e-wallets have much higher average transaction size compared to the direct carrier billing ones.

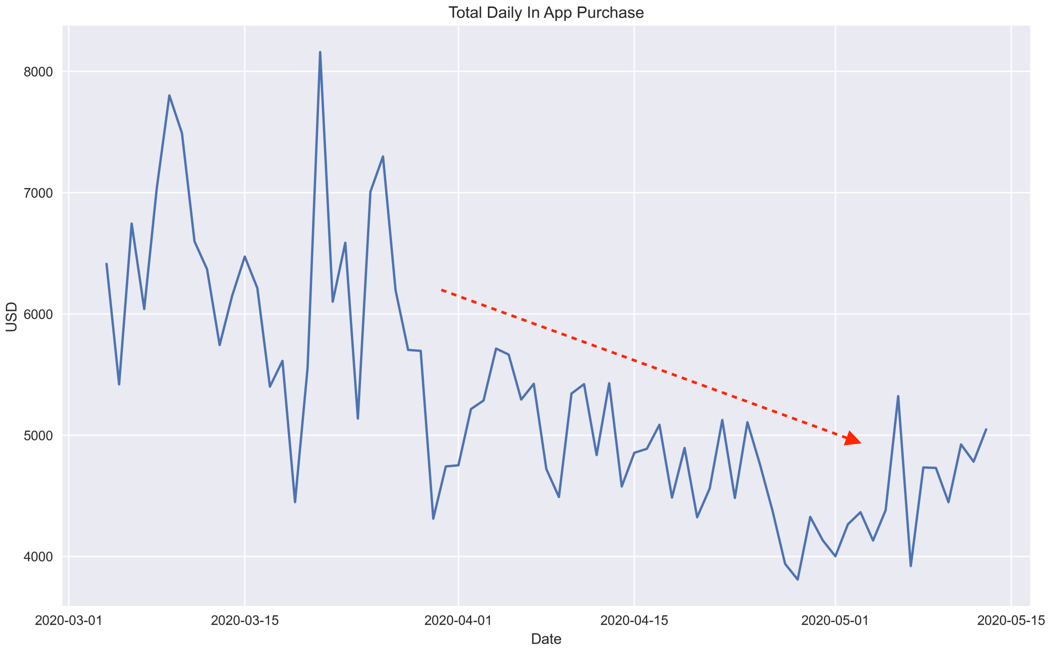

Figure 1. Dropping in revenue since the lockdown

4.2 Feature engineering and modelling

I first summing up a lot of user attributes for t0 to t4 such as number of game played by unique users, number of winning matches, losing matches, winning chips, losing chips and the total in-app purchased amount. And then fit the data into three models, Logistics Regression, Random Forest and Multi-layer Perceptron, I ended up having 37 features. Then I went to make each day user data into features and ended up with 148 features. The accuracy also improved by 6% which is great.

5. A summary of all the modelling completed

01. Logistic Regression Classifier :

```
              precision    recall  f1-score   support

           0       0.80      0.83      0.81     28416
           1       0.80      0.77      0.78     25418

    accuracy                           0.80     53834
```

```
   macro avg       0.80      0.80      0.80     53834
weighted avg       0.80      0.80      0.80     53834
```

02. Random Forest Classifier :

```
              precision    recall  f1-score   support

           0       0.81      0.81      0.81     28416
           1       0.79      0.79      0.79     25418

    accuracy                           0.80     53834
   macro avg       0.80      0.80      0.80     53834
weighted avg       0.80      0.80      0.80     53834
```

03. Multi-layer Percepton Classifier :

```
              precision    recall  f1-score   support

           0       0.78      0.80      0.79     28416
           1       0.77      0.75      0.76     25418

    accuracy                           0.78     53834
   macro avg       0.78      0.78      0.78     53834
weighted avg       0.78      0.78      0.78     53834
```

6. Conclusion

In the end, Logistics Regression and Random Forest have about the same accuracy. However, the company can choose Logistic Regression Classifier for its quickest run time. After running the models, we can get the list of users who are highly likely to churn we can approach these users in a few way to improve the chance of them staying in the game.

Firstly, the company can segment the users into multiple groups based on their chances of leaving. For example: 0-20%, 21-40%, 41-60%, 61-80% and the rest.

Secondly, the company can deploy A/B testing on each of the segment to check how they respond to certain promotion packages and choose the best ones that work.

Thirdly, the company can also batter matched the users of the same winning and losing rate so that they can spending time with the people of the same skills and do not feel exploited by highly skillful users.