Sample Project Report

Problem statement

The project aims to create a model using user data to predict whether a user will quit the game after a certain amount of playtime.

Background on the subject

The mobile gaming industry is characterised by intense competition, with approximately 7000 games released on **Google Play** and 2000 on **Apple iTunes** daily. Users frequently switch between games of the same genre, underscoring the importance for developers to **predict** when users might disengage. By understanding the drivers of **user churn**, developers can proactively implement strategies to retain users and enhance **revenue**. This study aims to apply **data science** techniques to assist game developers in predicting whether a user is likely to **leave a game** after **four days**. While similar research exists for casual games, there is a noticeable gap in studies focusing on social casino games.

Data Set

The dataset consists of two CSV files: **"User Play Log"** and **"Transaction Records"** from a social casino game called **LengBear**, available exclusively in the Cambodia market. This game has been downloaded more than **1.2 million** times. The play log covers the period from May 1st to May 13th, 2020, containing **1,768,639 rows** from **241,713** unique UserIDs. The transaction record file includes **35,680** transactions recorded between **March 1st** and **May 15th, 2020**.

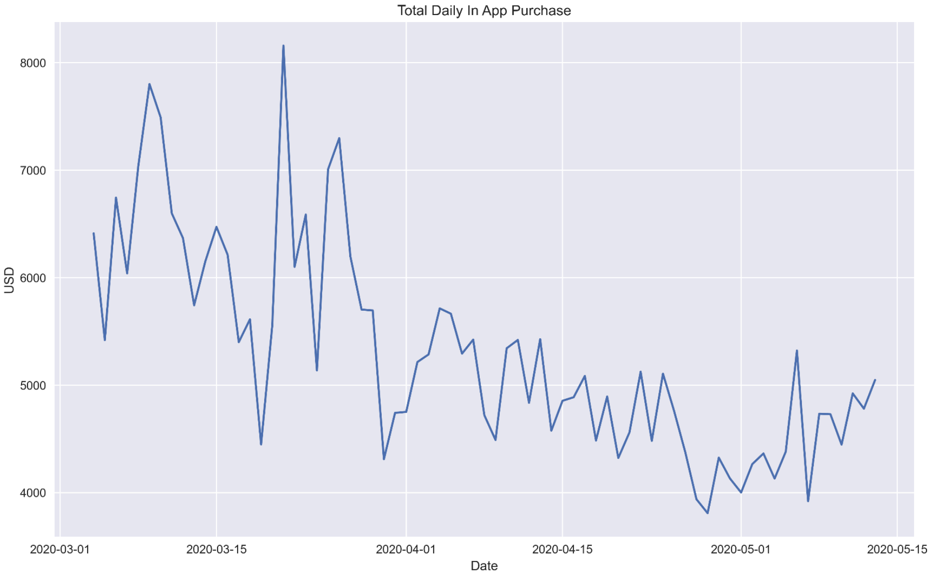
Player details include columns like **'Sequence'**, **'UserID'**, **'GameID'**, **'Level'**, **'WinNo'**, **'DrawNo'**, **'LostNo'**, **'WinAmt'**, **'LostAmt'**, **'Date'**, **'Currency_Type1'**, and **'Currency_Type2'**. Transaction details consist of **'UserID'**, **'Amount'**, **'Chips'**, **'Date'**, and **'Channel'**.

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

Exploratory Data Analysis

Saturday is when users spend the most money. This might be because they have more free time on weekends and play more, which leads them to spend more. To make more money, we should advertise promotions heavily on Friday nights and Saturdays.

When I looked into how people pay, I found something interesting. In rich countries, many people use credit cards to buy stuff in games, which is really easy. But in lots of places, like Cambodia, many people don't have bank accounts. Instead, they use phone bills or local digital wallets to pay in games. This is cool because game makers can keep more money since they don't have to share profits with big stores like Google or Apple. Also, people who use digital wallets tend to spend more money compared to those who use phone bills.



Feature engineering and modelling

First, I combined several user stats from time t0 to t4, like games played, wins, losses, chips, and purchases. Then, I used three models—**Logistic Regression**, **Random Forest**, and **Multi-layer Perceptron**—with a total of **37 features**. After breaking down daily user data into features, I ended up with **148 features**. This boosted accuracy by **6%**, which is pretty cool.

A summary of all the modelling completed

```
1.Classification report for Logistic Regression model:
              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
```

```
2.Classification report for 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
```

3.Classification report for Multilayer Perceptron Classifier:

```
              precision    recall  f1-score   support

           0       0.80      0.80      0.80     28416
           1       0.77      0.78      0.78     25418

    accuracy                           0.79     53834
   macro avg       0.79      0.79      0.79     53834
weighted avg       0.79      0.79      0.79     53834
```

Conclusion

At the end, both **Logistic Regression** and **Random Fores**t had similar accuracy. However, the company can opt for **Logistic Regression Classifier** due to its **faster** runtime. After running the models, we can identify users highly likely to churn and take steps to retain them.

Firstly, the company can group users based on their likelihood of leaving, like 0-20%, 21-40%, etc. Then, they can test different promotion packages on each group using A/B testing to see which ones work best. Lastly, they can match users with similar winning and losing rates, so they can play with others of similar skills and not feel overwhelmed by highly skilled players.