Nathan Esau, Steve Kane
September 2015

Case Study Competition

Analysis of a Uken Games mobile app

SFU

Objectives

1. Come up with a good way to visualize the data that helps to provide explanatory insights
2. Use the demographic and event features to predict revenue, engagement and player retention
3. Suggest control-treatment experiments for follow up analysis

Introduction

Size training set: 250,000 players
Size validation set: 50,000 players
Types of variables: Country, gender, dates of in-game events, in-game purchases, prizes won; about 40 variables

Revenue measures the amount of money spent, while **Engagement** measures the amount of time spent in game. **Retention** indicates whether a player plays the game 30 days after installation.

- `Revenue` and `engagement` are **heavily** skewed
- `97.8%` of players don't spend money
- Given that players pay, the average is `$3.33` with only `5%` paying more than `$13.20`
- The average time spent playing is `32.75`, but the median time spent playing is only `7`.
- `Retention`: `9%` of players returned 30 days later

Figure 1: Important features for predicting revenue

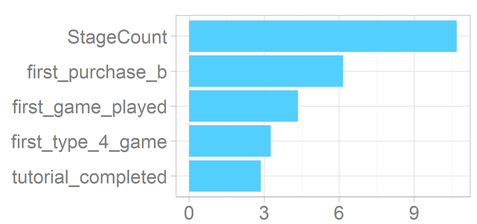


Figure 2: Important features for predicting engagement

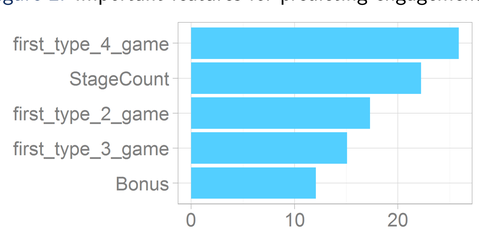


Figure 3: Stages played, `engagement`, `revenue`, player skill (`bonuses` completed)

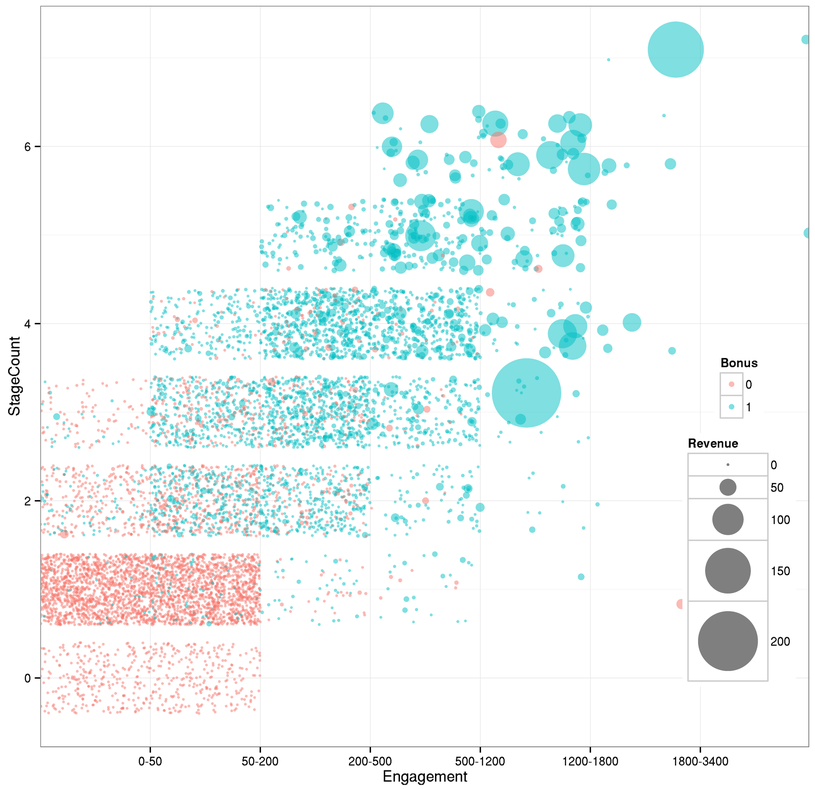
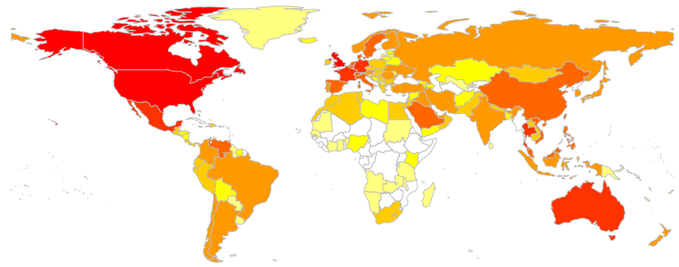


Figure 4: Heatmap for distribution of players from around the world



Predictions

- We used the random forest model to predict `revenue` and `engagement` since this type of model works well with skewed data (Huang, 2005).
- The gradient boosted model was used to predict `retention`. This model performed better than linear models we tried.

Table 1: Comparison of prediction mean and training mean

	`Revenue`	`Engagement`	`Return Player`
Validation Mean	3.31 | Revenue > 0	33.33	6.5%
Training Mean	3.33 | Revenue > 0	32.75	9.0%

Note that the random forest model predicts 0 revenue for most of the new players, which is what we would expect.

Random Forest Model

Figure 5: Simple decision tree

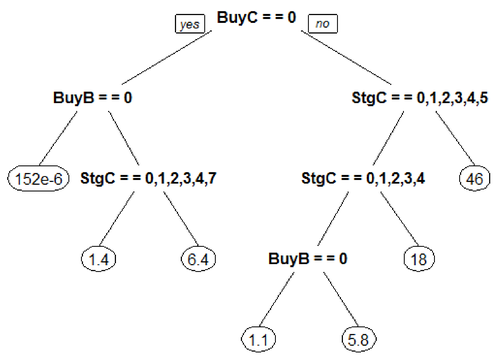
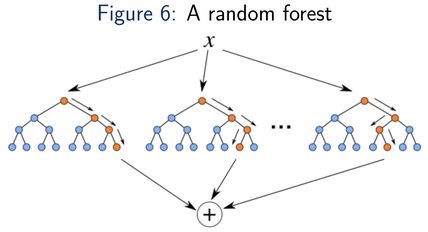


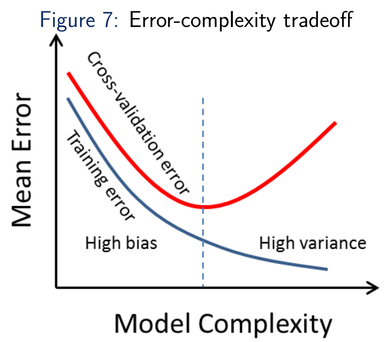
Figure 6: A random forest



We create an `ensemble` of many unstable models to form a stable model.

Gradient Boosted Model

Figure 7: Error-complexity tradeoff



We update our model, $F(x) = 0.5 \log\left(\frac{1+\bar{y}}{1-\bar{y}}\right)$ by continually adding a new basis function to minimize the loss function,

$$l(y_i, \hat{y}_i) = y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i) \ln(1 + e^{\hat{y}_i})$$

The gradient boosted model uses a `greedy` algorithm.

Conclusions

A/B tests to improve key metrics

- Figure 3 shows that completing bonus objectives is likely to increase revenue. We expect that *creating additional stages* or *adjusting the difficulty* of the bonus objective could increase revenue.
- It would be interesting to analyze the impact of providing unlockable features when a player the game connects to Facebook or another device.

Relationship of key metrics

- Stage Count is highly significant to both `engagement` and `revenue`
- The types of games a player tries out is highly significant to `engagement`
- The type of purchase a player made is very significant to `revenue`

References

Analysis was done using `R`. `http://www.R-project.org`. The `ggplot2`, `randomForest`, `data.table`, `xgboost`, `readr`, `qdapTools`, and `Matrix` packages were used. For the **models** see *Methods to Extract Rare Events* by Weihua Huang. 2005.