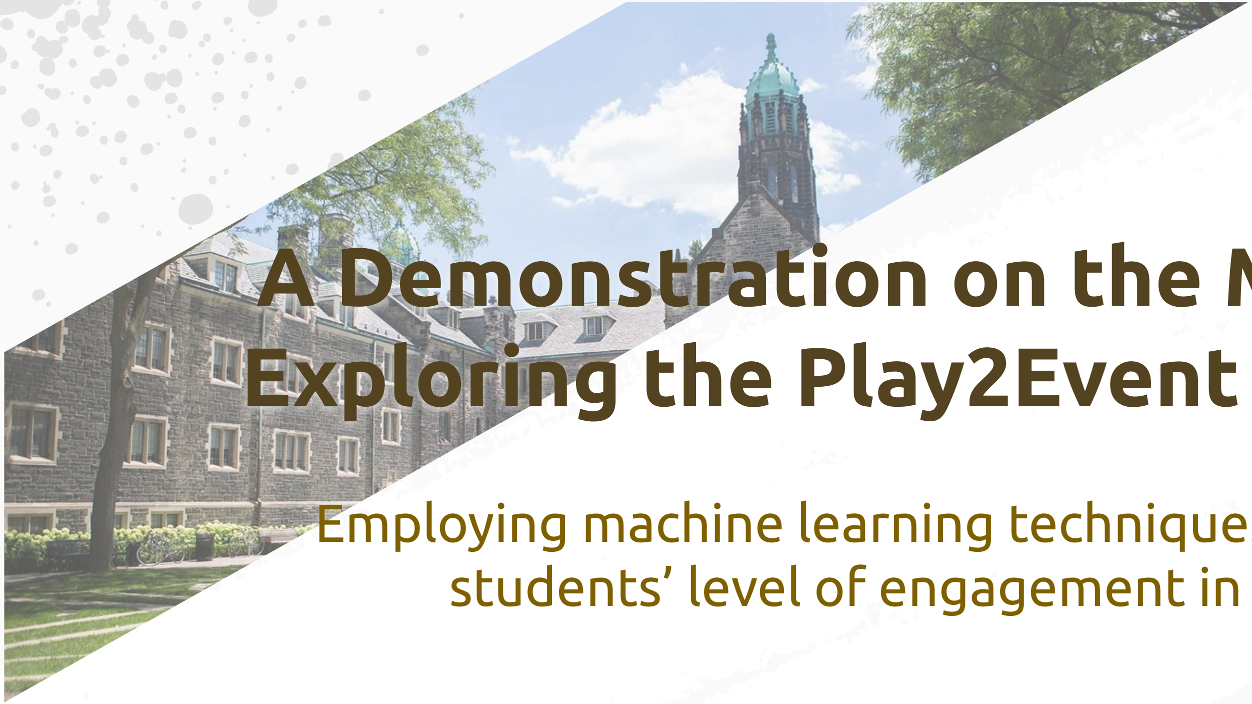


A Demonstration on the Methods of Exploring the Play2Event Game Data

Employing machine learning techniques to investigate students' level of engagement in the game

Presented by We R :

Pan Chen, Siqi Zheng,
Xiaoxuan Han, Yini Mao

TABLE OF CONTENTS

1 Introduction

2 Data Cleaning & Wrangling

3 Principal Component Analysis

4 Survival Analysis

5 Results

6 Limitations & Future work

Introduction of the Problem

Background

Elm City Stories, an educational video game, aims to increase middle to high school students' perception of risk and therefore prevent their risky behaviours. The experiment records the logs of 166 players (11-14 years-old) with their actions and time spent in each event.

Goal

Investigate and improve the engagement level of participants using machine learning techniques

Data Cleaning & Wrangling

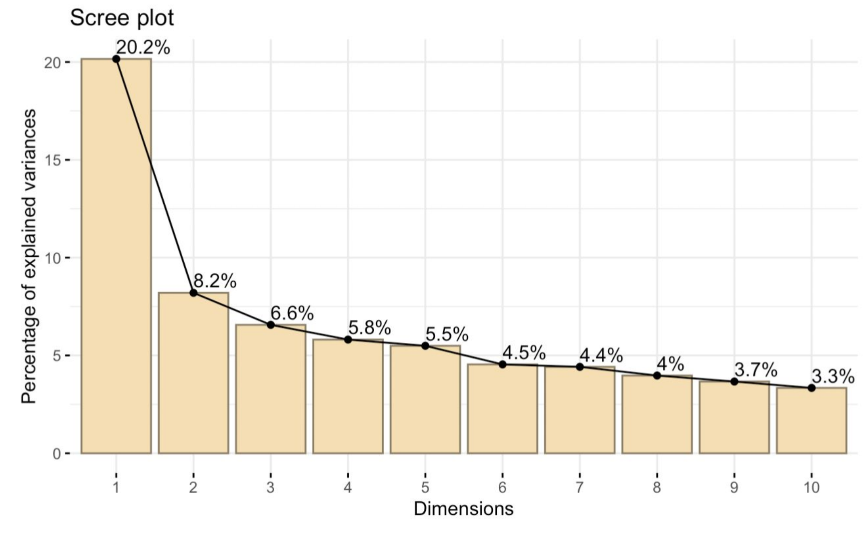
1. Based on “event_id”, kept the *“proportion_complete”* that had the the largest time elapsed.
2. Cleaned the data by removing and replacing NA values, and deduplicating observations that had the same *“player_id”*, *“event_id”*, and *“event_time_dbl”*.
3. Removed outliers for potential explanatory variables.
4. Split *“proportion_complete”* into binary categories by using its mean as the threshold.
5. Created a new variable *“event_new”* to store the binary outcomes obtained from the previous step.

That is our
response variable!

Principal Component Analysis

What is PCA?

- ➢ A technique for reducing dimensionality
- ➢ Aim to explain most of the variability in a large dataset with a small number of factors that are uncorrelated with each other



What did we do?

- ➢ Compute the principal components for numerical variables only
- ➢ Find important factors that explains most of the variation
- ➢ Visualize eigenvalues (scree plot).

Results

- ➢ We found that the following ten variables are the most important:
  - ○ Elapsed time (in seconds)
  - ○ Player’s current “know” ,”priority”, “people”, “refusal”, and “me” skill level
  - ○ Advancement level
  - ○ Previous and current points at this level
  - ○ Amount of time watched for animation

Survival Analysis - Survival Random Forest (Ishwaran et al., 2008)

- Draw 500 bootstrap samples called out-of-bag data .
- Grow a survival tree for each bootstrap sample.
    - Survival trees: binary trees grown by recursive splitting of tree nodes.
    - Using a survival criterion, the root node is classified into two daughter nodes.
- At each node of the tree, randomly select 4 candidate variables.
    - The node is split using the candidate variable that maximizes survival difference between daughter nodes.
- Grow trees to full size.
- Calculate the ensemble cumulative hazard function (CHF) for each tree.

Results & Variable Importance

skill_level_refusal	0.1575
skill_level_people	0.1391
skill_level_know	0.1387
skill_level_priority	0.1369
avatar_age	0.1075
animatic_time_elapsed	0.0854

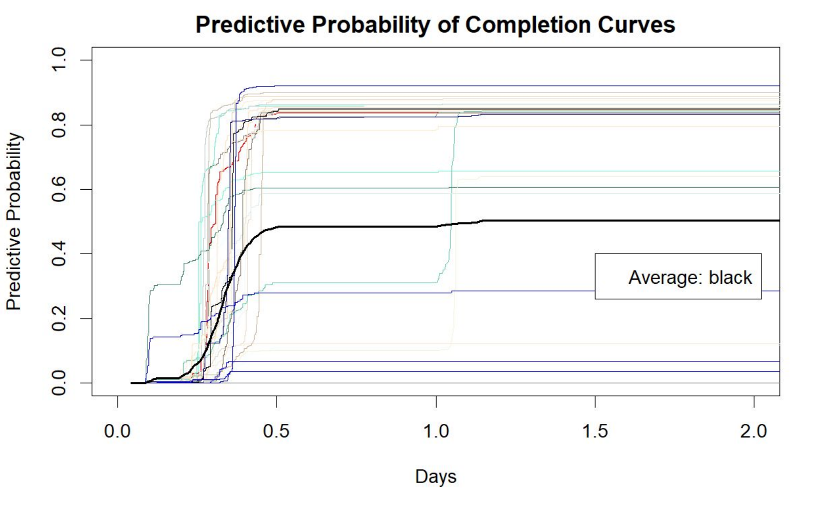
Results

➢ Besides the variables from PCA, we also add a new categorical variable age

➢ Player's current "know", "priority", "people", "refusal" skill level, age and animation time are the most important variables

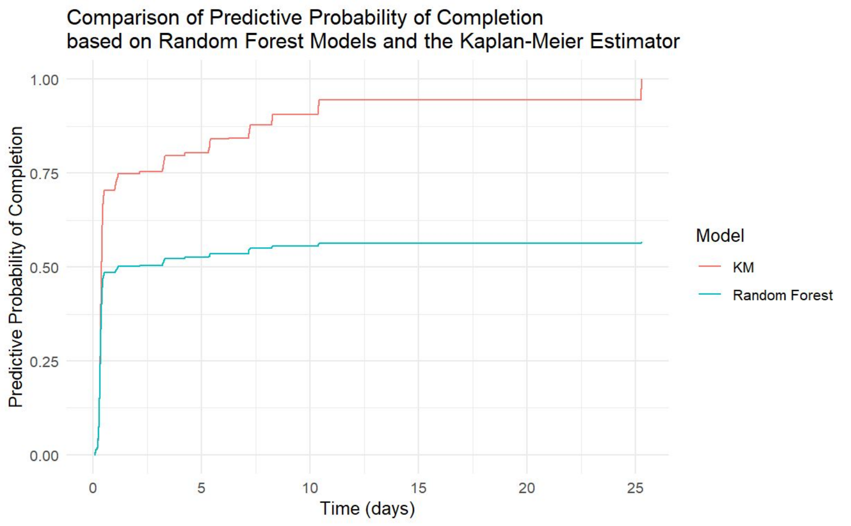
○

Predictive Probability of Completion Curves



- ➢ The average predictive probability is around 0.5
- ➢ The probability of completion varies much by users

Predictive Probability



- ➢ Kaplan-Meier Estimator predicts probability based on time only
- ➢ We want to quantitatively assess the contribution of each variable - Cox Proportional-hazards Model

Survival Analysis - Cox Proportional-hazards Model

- The Cox proportional hazards model is a regression model for investigating the association between the playing time of participants and the event happening.
- We want to examine how specific factors influence the event happening at a particular point in time i.e. the hazard ratio (HR).

Survival Analysis - Cox Proportional-hazards Model

$$h(t|x(t)) = h_0(t)\exp((x(t) - \overline{x})'\beta)$$

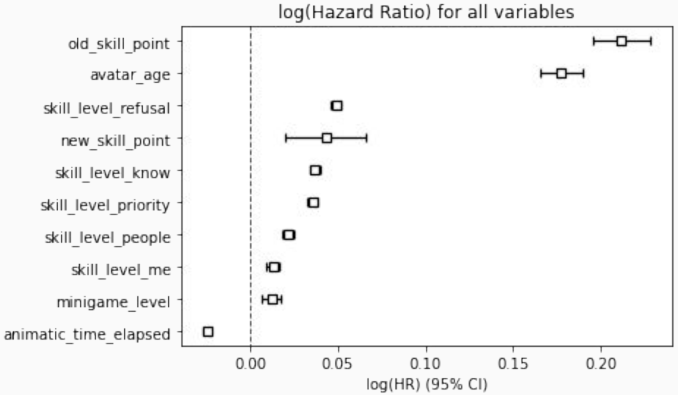
$h_0(t)$: the baseline hazard function.

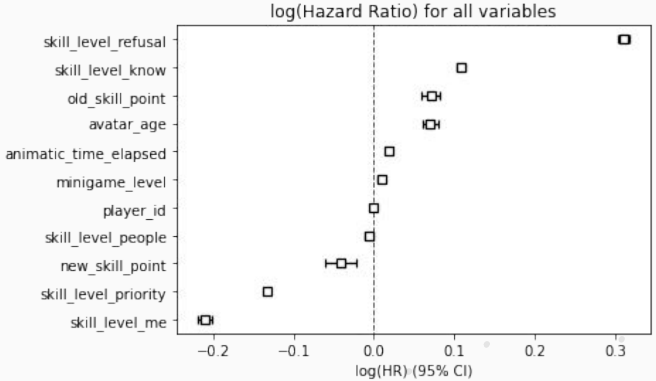
β : vector of regression coefficients.

$x(t)$: covariates at time t.

$\bar{x}$: the mean of each covariate.

95% Confidence Interval of log(HR)





- ➢ A higher Hazard Ratio (HR) indicates a higher completion rate, in this context.
- ➢ old_skill_point (previous points at this level of the game), avatar_age (years), skill_level_refusal (player's current skill level) -> most positive for the event completions.
- ➢ The game should be designed to help them get more sense of achievement.

Conclusion

- By applying Principal component analysis, we found Player’s current “know” ,”priority”, “people”, “refusal”, and “me” skill level affect the event completion rates most. Therefore, the game designers could think about how they can train these skills of participants to help them better utilize this game.

- The Survival Random Forest Model shows that the probability of completion is predicted to be only around 50%, but the predicted probability varies much depending on the player. Therefore, there is still much room to adjust the focuses and the overall design of the game.

- The game designer should consider lowering the game difficulty, to help the participants get a sense of achievement, which is shown a way to increase the event completion rate by the Cox Proportional hazard model.

Limitations & Future Work

Limitations

- ➢ Lack of health outcome data from players to build a prediction model
- ➢ Only analyzed the log dataset instead of incorporating more information
- ➢ Training models with a subset due to computational constraints

Future Work

- ➢ Include a more holistic range of categorical variables
- ➢ Train models with GPU
- ➢ Investigate the mean score of students in the treatment groups to analyze the effects of the game

Bibliography

1. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2 (3). doi: 10.1214/08-aoas169
2. Jolliffe, Ian T., and Jorge Cadima. “Principal Component Analysis: A Review and Recent Developments.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016, p. 20150202., https://doi.org/10.1098/rsta.2015.0202.
3. “R PCA Tutorial (Principal Component Analysis).” *DataCamp Community*, https://www.datacamp.com/community/tutorials/pca-analysis-r.
4. Therneau, T. M. (2021). A package for survival analysis in r. Retrieved from https://CRAN.Rproject.org/package=survival
5. Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77 (1), 1–17. doi: 10.18637/jss.v077.i01

Thank You !