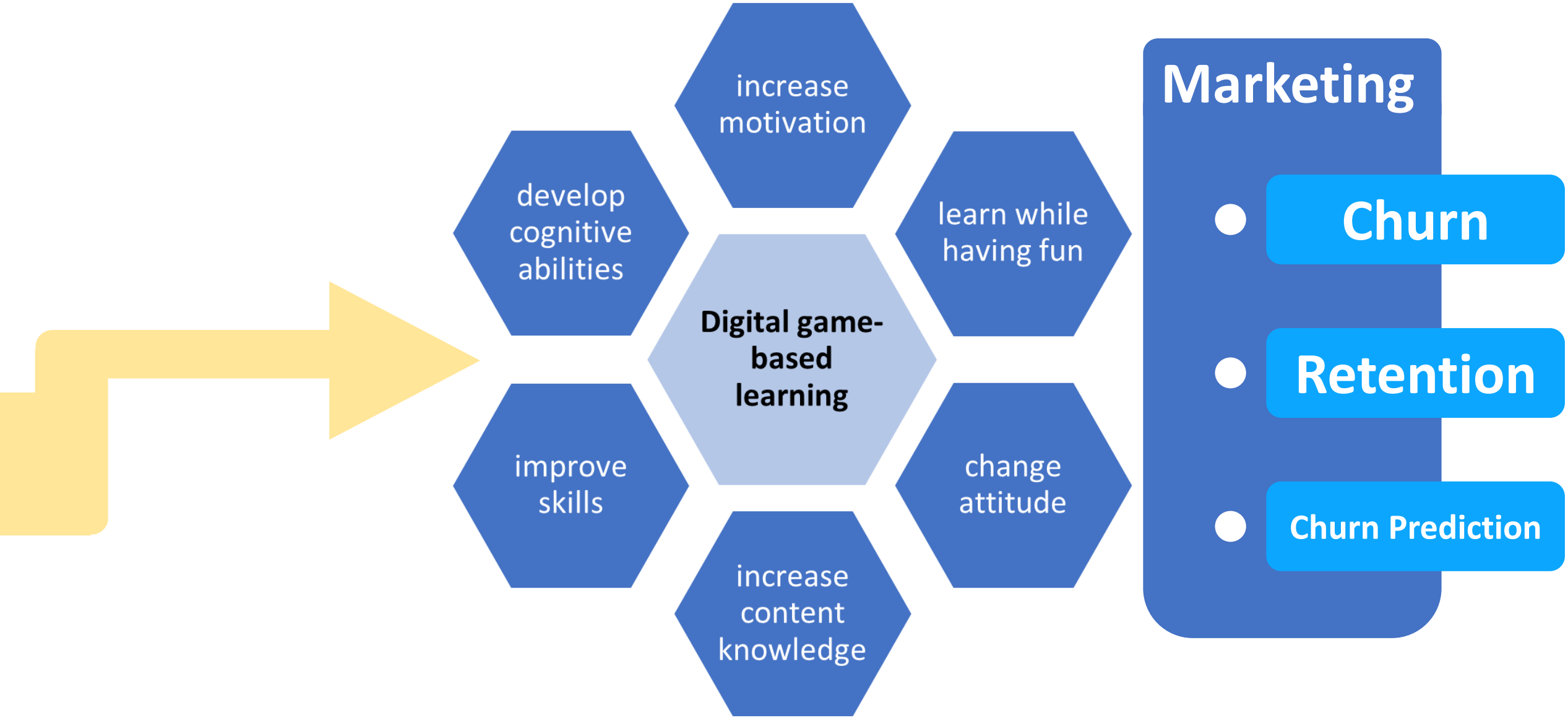
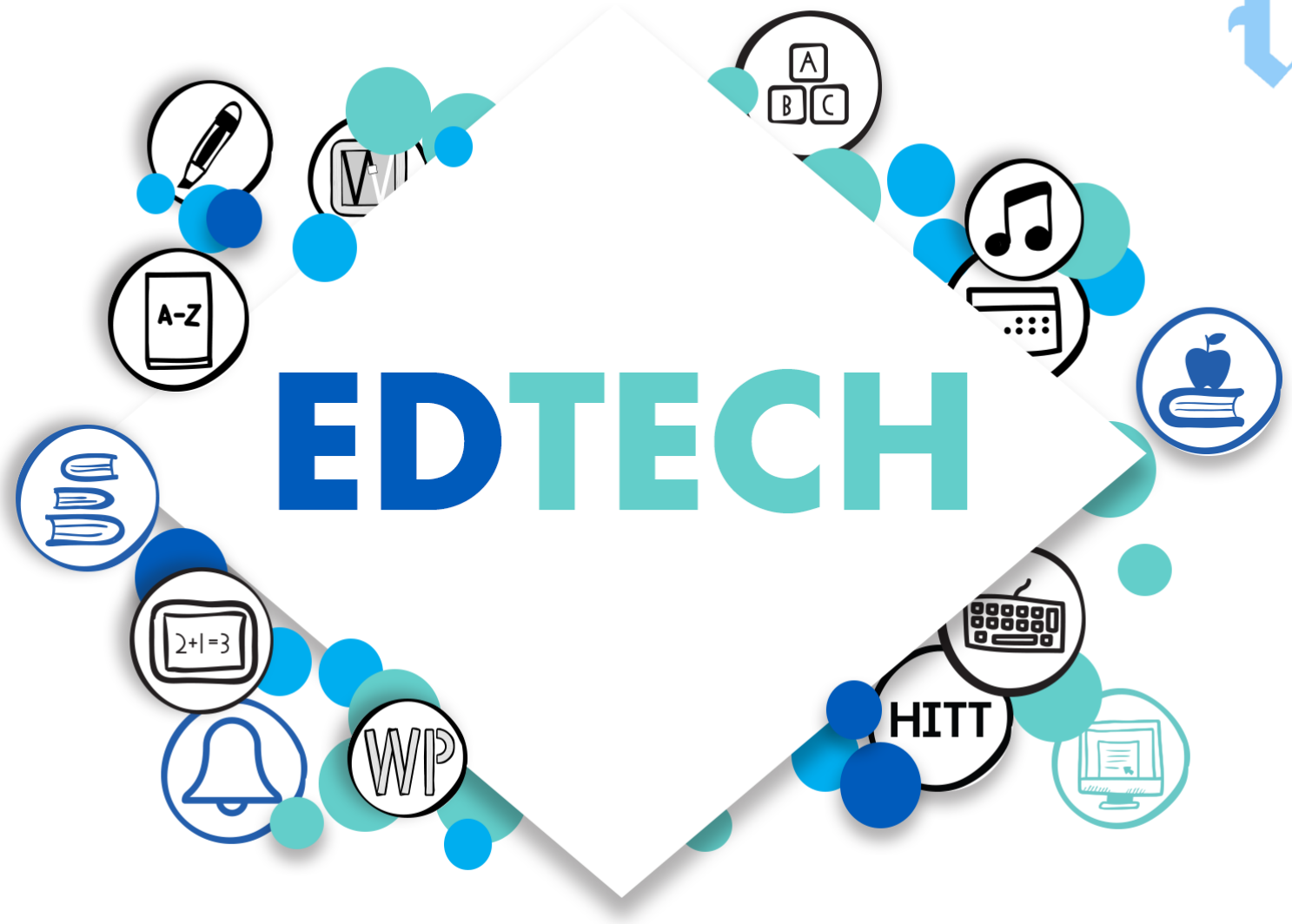
Parham Pishro

Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest

Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest

Educational Technology


increase motivation
develop cognitive abilities
learn while having fun
Digital game-based learning
improve skills
change attitude
increase content knowledge
Marketing
Churn
Retention
Churn Prediction



Graduation Rate:

Virtual School) 50.1%

Overall) 84%

Median Churn Rate in EdTech = 10.29%

- EdTech

Table 2
Summary of churn prediction literature in DGBL and gaming industries.

Reference	Category	Purpose	Variables used for prediction	Model	AUC/F1 results
[33]	DGBL	Prediction of abandonment in tutorial	Cumulative features (e.g. idle time, execution button clicks), learner features (e.g. age, gender, registration status)	LR, RF, DGBT	AUC = 0 .68
[21]	Online /mobile game	Prediction of churn	Number of sessions, number of days, average playtime per session, average playtime between sessions, etc.	LR, DT, NB, NN	F1 = 0 .916
[22]	Online /mobile game	Prediction of disengagement	Event frequency	SVM, DT, LR	AUC = 0 .7
[23]	Online game	Prediction of churn	Activities, performance, achievements	Hidden Markov model	AUC = 0 .77
[19]	Online game	Prediction of churn	Recency, frequency, monetary value, length of relationships, inter-play, bonuses/rewards, demographic	E-CHAID DT	AUC = 0 .88
[5]	Online game	Prediction of player	Lifetime engagement features (e.g. login frequency, average playtime), performance features (e.g. level, coins), social interaction features (e.g. number of in-game friends)	Stickiness based FCM for clustering NB, RBF	–
[20]	Online game	Prediction of churn	In-game activities (e.g. number of days user play, playtime)	RF, XG boost, generalized boosting regression	AUC = 0 .9358
[25]	Mobile games	Prediction of early churn	Activity features (e.g. playtime, session count), monetization features (e.g. in-game money spend), gameplay style features (e.g. auction usage)	LR, DT, RF, NB, gradient boosting	AUC = 0 .83
[34]	Mobile game	Prediction of churn	Play history (e.g. timestamp of play), game profiles (e.g. game genre, developer, rating), user information (e.g. device model, OS version)	semi-supervised deep NN, LR, RS, DT, RF, SVM	AUC = 0 .82
[24]	Mobile game	Prediction of churn	Activity indicators, Activity time, Engagement indicators	Proposed joint model	–
[35]	Gamification/Online game	Prediction of churn	In-game activities (e.g. playtime, frequency of usage, game actions and participation)	ANN, RF (Classification and Regression)	AUC = 0 .77

Notes: DGBT, Gradient Boosting Decision Tree; DT, Decision Tree; LR, Logistic Regression; NB, Naive Bayes; NN, Neural Networks; RBF, Radial Basis Function; RF, Random Forest; SVM, Support vector machines.

Table 3
Summary of churn prediction literatures in education.

Reference	Category	Purpose	Data source	Variables used for prediction	Model	AUC/F1 results
[36]	Online/On-campus	Early prediction of failure	Online course and school database	On-campus data (e.g. age, gender, civil status, exam performance), distance learning data (e.g. access frequency, participation in the forum)	SVM, DT via J48, NN, NB	F1 = 0 .82
[37]	Online	Prediction of dropout	Survey	Demographic variables, self-efficacy, readiness, prior knowledge, and locus of control	KNN, DT, NB, and NN	AUC = 0 .866
[28]	On-campus	Prediction of retention rate	School database	Gender, residency status, ACT composite score, high school class rank, race/ethnicity, student, etc.	LR	–
[6]	On-campus	Prediction of dropout and failure	Survey, School database	Scores of each subject, level of motivation, GPA, smoking habits, physical disability, etc.	JRip, NNge, OneR, Prism, Ridor, ADTree, J48, RandomTree, REPTree, SimpleCart	–
[38]	On-campus	Prediction of failure	Survey	Personal and family-related (e.g. age, parents occupation), previous education (e.g. scores of multiple subjects of previous education), academic results (e.g. scores)	NNge, OneR, SimpleCart, Random Tree, NB	–
[31]	MOOCs	Temporal prediction of dropout	MOOCs database	Clickstream (which pages students visited and when or how many times students clicked on certain sources (e.g., syllabus, modules, quizzes, etc.)), quiz scores and discussion forum data	general Bayesian network, decision tree (C4.5), Stacking	AUC = 90.7
[7]	MOOCs	Prediction of decrease of engagement	MOOCs database	Video engagement, exercise engagement, assignment engagement	LR, stochastic gradient descent, RF and SVM	AUC = 0 .914
[39]	MOOCs	Prediction of dropout	MOOCs database	Clickstream data divided into 7 categories (e.g. video, access, wiki, problem, navigate, discussion and page close)	Proposed deep learning model named CLSA, LR, SVM, CNN, LSTM, CNN-LSTM, and DP-CNN	F1 = 0 .869

Role Play Game

CONTENTS

Data Collection

Two Types of Data

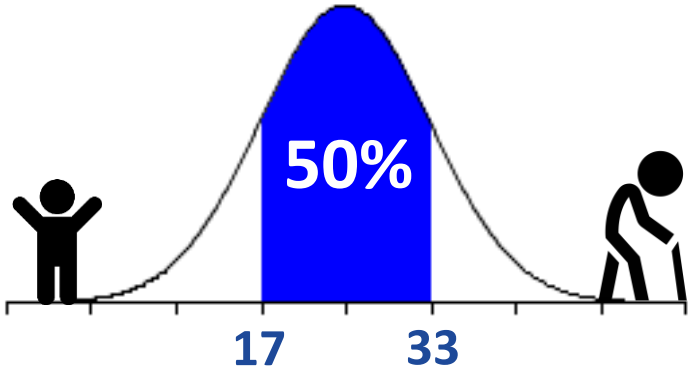
- Customer Relationship Management (CRM)
- Semi-Structured User Log Data

3557

56% Female
44% Male

Table 4
The content of the data.

ID	Number of observations	Description
1	6,973	Chapter related dates such as release, start and finish dates.
2	514	Chapter 7 finish date.
3	478,700	Historical experience points and coins acquirement data.
4	1,496	Historical user replay data.
5	3,982	CRM data. Contains demographic information such as user name, email, address, birth date, and gender.
6	8,587,940	User log of all activities except for learning contents related.
7	562,581	User log in the learning contents such as start and finish date-time of a lesson.



Data Selection

- User id
- Timestamp
- Contents

(from log data for computing the average or duration of logins, plays, and clicks number)

- Level
- Number of coins

(called performance features)

- Gender
- Date of birth
- Resident area

(from CRM data)

Data Aggregation, Transformation, and Integration

- ❑User activity and lesson log integration
- ❑Experience points, coins, and replay data
- ❑CRM data transformation
- ❑Chapter progress data
- ❑Player data

$$if\ difference\ between\ to\ timestamps \leq 60:$$
$$playtime = difference$$

$$else:$$
$$inactive = difference$$

Exploratory Data Analysis (EDA)

- Data Preparation
- Churn Definition
- Churn Determination
- Retention Analysis
- Hyperparameter
- Prediction Evaluation
- Discussion Conclusion

Table 5
The descriptive statistics of all variables.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
total_playtime (min)	3701	1640.15	2207.54	0	339.55	885.82	2213.1	38390.98
total_login	3701	36.01	50.27	1	700	2000	4700	742
total_inactive (min)	3701	307397.91	255321.92	0	66484.59	251926.17	496751.22	928384.54
playtime_average (min)	3701	52.09	28.87	0	33.93	46.74	63.15	298
inactive_average (min)	3701	16507.71	25549.83	0	4375.9	9074.17	18022.08	379752.63
first_login	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last_login	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
entire_period (days)	3701	211.51	178.95	0	42	173	344	645
churn_status	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
chapter1_playtime (min)	3701	404.72	373.23	0	243.65	344.1	479.33	7442.73
chapter2_playtime (min)	3701	479.59	535.10	0	0	386.45	804.52	5842.65
chapter3_playtime (min)	3701	210.84	399.96	0	0	0	383.18	8615.63
chapter4_playtime (min)	3701	234.51	642.30	0	0	0	274.37	22511.83
chapter5_playtime (min)	3701	117.44	360.91	0	0	0	0	7105.2
chapter6_playtime (min)	3701	99.97	331.14	0	0	0	0	4787.72
chapter7_playtime (min)	3701	93.08	492.92	0	0	0	0	14452.2
exp	3521	13753.37	14314.31	150	4570	8670	17700	135010
coins	3521	15485.90	15841.23	200	4700	9920	21760	155850
replay	3521	0.37	4.1	0	0	0	0	177
wait_average (days)	3701	11.43	37.92	0	0	1.48	6.46	644.54
gender	3557	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	3557	29.63	70.44	0	17	26	33	2010
prefecture	3533	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Data Cleaning

Missing Value:

Gender, Age) Trial accounts $\xrightarrow{Handling}$ *Remove*

Prefecture) … $\xrightarrow{Handling}$ *Replace with "Tokyo"*

Exp, Coins, Replay) New players $\xrightarrow{Handling}$ *Fill with* 0

chapter5_playtime (min)	3701
chapter6_playtime (min)	3701
chapter7_playtime (min)	3701
exp	3521
coins	3521
replay	3521
wait_average (days)	3701
gender	3557
age	3557
prefecture	3533

Impossible Value:

Age) 0 to 2010 $\begin{cases} -4 \xrightarrow{Handling} Replace\ with\ median\ (26) \\ +117(119{,}825{,}2002, \dots) \xrightarrow{Handling} Replace\ with\ median\ (26) \end{cases}$

Outlier:

All of them) … $\xrightarrow{Handling}$ *Standardization*

Features and Target

Table 6
The variables in *features* dataset.

Variable name	Category	Details
total_login	Engagement	Total number of logins of the user.
entire_period (days)		The engagement period. Subtract *first_login* from *last_login*.
avr_ch_wait (days)		The average period between open and start. Total wait divide by the number of chapters played.
replay		Total number of replay per user.
total_playtime (min)		Total playtime of the user in minutes.
total_inactive (min)		Total inactive time between logins.
average_playtime (min)		Average playtime per login. Calculated by *total_playtime* divided by *total_login*.
average_inactive (min)		The average inactive time between logins. Calculated by *total_inactive* divided by *total_login*.
ch1_playtime (min)	Performance	Playtime of chapter 1
ch2_playtime (min)		Playtime of chapter 2
ch3_playtime (min)		Playtime of chapter 3
ch4_playtime (min)		Playtime of chapter 4
ch5_playtime (min)		Playtime of chapter 5
ch6_playtime (min)		Playtime of chapter 6
ch7_playtime (min)		Playtime of chapter 7
exp		Total exp points per user
coins		Total coins per user
gender	Demographic	Gender in binary 0 or 1 (0 = Female and 1 = Male)
age		Age of the player
prefecture		Prefecture from 0 to 47
churn_status	Target	Churn status in binary. 0 or 1 (0 = False and 1 = True)

Feature Selection

- Data Preparation
- Churn Definition
- Churn Determination
- Retention Analysis
- Hyperparameter
- Prediction Evaluation
- Discussion Conclusion

Table A.1
Explained variance of principal components.

	total_login	*total_playtime* (min)	*total_inactive* (min)	*average_playtime* (min)	*average_inactive* (min)	*entire_period* (days)	*ch1_playtime* (min)	*ch2_playtime* (min)	*ch3_playtime* (min)
PC-1	2.00E−06	1.00E−06	1.00E+00	−6.69E−07	1.00E−06	0.0007	4.46E−07	0.000002	1.00E−06
PC-2	−5.00E−04	0.00023	−7.88E−06	1.23E−03	−0.002	0.0099	2.32E−03	0.002199	7.00E−05
PC-3	0.0102	0.01323	−2.06E−06	5.76E−03	−0.001	0.0061	5.67E−03	0.012157	0.0116
PC-4	0.1663	0.18799	−6.05E−04	1.75E−02	−0.096	0.8606	8.42E−02	0.116752	0.1384
PC-5	0.2719	0.30871	3.52E−04	3.62E−02	−0.129	−0.506	7.84E−02	0.161545	0.2256
PC-6	−0.029	0.03238	−2.08E−05	1.11E−01	0.4419	0.0295	1.32E−01	−0.382636	−0.242
PC-7	0.1508	−0.03262	−7.21E−06	−7.75E−01	−0.159	0.011	−5.39E−02	−0.169805	−0.134
PC-8	0.0694	0.0559	1.86E−05	7.53E−02	−0.246	−0.026	7.83E−01	0.17535	0.0091
PC-9	0.1422	0.02436	7.13E−06	−4.70E−01	−0.128	−0.012	−9.13E−02	0.15487	0.1345
PC-10	−0.069	−0.02935	−2.39E−05	1.63E−01	0.0029	0.0344	−4.78E−01	0.11993	0.2582
PC-11	0.1205	0.08686	8.32E−07	−2.45E−01	0.7706	−0.004	2.10E−01	0.093908	0.4173
PC-12	0.056	0.05772	7.58E−07	2.64E−02	0.098	−0.001	−1.33E−01	0.331262	0.1194
PC-13	−0.14	−0.0956	−3.09E−06	−4.76E−02	−0.119	0.005	1.27E−01	−0.655647	0.2889
PC-14	−0.081	0.01833	2.38E−06	6.26E−02	−0.208	−0.002	−2.28E−02	−0.229848	0.6743
PC-15	−0.047	−0.01299	2.87E−06	5.21E−03	−0.034	−0.004	1.01E−02	0.106602	0.0718
PC-16	0.1724	0.04301	−1.46E−06	1.73E−02	−0.034	0.0022	2.07E−02	−0.059972	−0.038
PC-17	0.4638	0.08592	8.15E−06	2.14E−01	−0.102	−0.011	−2.69E−02	−0.170329	0.0664
PC-18	0.7444	−0.25832	−5.03E−06	1.41E−01	0.0578	0.0068	−1.21E−01	−0.16183	−0.076
PC-19	−0.007	0.00207	2.78E−07	−3.74E−04	0.0051	−5.00E−04	−5.72E−03	0.008591	0.003

	ch4_playtime (min)	*ch5_playtime* (min)	*ch6_playtime* (min)
PC-1	0.000001	1.00E−06	8.98E−07
PC-2	−0.001188	−0.001	4.60E−04
PC-3	0.007957	0.0091	1.01E−02
PC-4	0.147524	0.1502	1.50E−01
PC-5	0.281479	0.2694	2.48E−01
PC-6	0.168011	0.1533	6.46E−02
PC-7	0.072236	0.0241	−1.11E−02
PC-8	0.024768	−0.171	−3.29E−01
PC-9	0.062961	−0.065	−7.69E−02
PC-10	0.440438	−0.146	−3.06E−01
PC-11	0.02843	−0.056	−8.19E−03
PC-12	−0.211524	−0.392	−3.56E−01
PC-13	0.161819	−0.402	−1.43E−01
PC-14	−0.39824	0.3272	4.34E−02
PC-15	−0.140912	−0.587	7.19E−01
PC-16	0.202304	−0.021	−1.09E−02
PC-17	0.377125	−0.183	9.63E−02
PC-18	−0.409904	−0.036	−1.08E−01
PC-19	−0.001526	0.0035	−9.50E−04

	ch7_playtime (min)	*avr_ch_wait* (days)	*exp*	*coins*	*replay*
	5.97E−07	7.69E−07	2.00E−06	2.00E−06	3.49E−07
	−1.01E−03	−9.65E−04	0.001	0.0012	−1.27E−03
	8.35E−03	1.47E−03	0.0128	0.012	−1.08E−03
	1.33E−01	−3.16E−02	0.1635	0.1639	8.41E−02
	2.29E−01	−7.15E−02	0.2839	0.2834	2.03E−01
	2.74E−01	2.91E−01	−0.157	−0.149	5.52E−01
	8.13E−02	−4.48E−01	−0.139	−0.129	2.24E−01
	−2.21E−01	−2.78E−02	−0.111	−0.1	2.66E−01
	−8.02E−02	8.19E−01	−0.007	−0.01	2.76E−03
	−3.69E−01	−1.23E−01	−0.041	−0.04	4.40E−01
	−2.00E−01	−1.33E−01	0.0109	0.0045	−1.87E−01
	7.06E−01	−5.77E−02	−0.041	−0.047	9.85E−02
	1.26E−01	3.37E−02	0.3078	0.3135	−8.56E−02
	4.93E−02	7.15E−03	−0.269	−0.268	1.36E−01
	−1.07E−01	−4.31E−03	−0.063	−0.058	2.69E−01
	2.98E−02	−1.19E−02	−0.113	−0.12	−1.88E−01
	1.09E−01	7.23E−03	−0.358	−0.347	−3.43E−01
	−1.91E−01	8.90E−04	0.1577	0.1674	1.94E−01
	1.98E−03	2.27E−03	−0.704	0.7096	−7.52E−03

	gender	*age*	*prefecture*
PC-1	−3.25E−08	−2.00E−06	−6.73E−07
PC-2	1.20E−03	−0.104	−9.94E−01
PC-3	−2.74E−03	0.9939	−1.04E−01
PC-4	7.86E−03	−0.021	1.14E−02
PC-5	−2.73E−04	−0.025	−1.88E−03
PC-6	−1.61E−02	0.0043	−3.57E−03
PC-7	4.50E−02	0.0098	−2.40E−03
PC-8	1.07E−02	0.001	2.21E−03
PC-9	1.06E−02	−0.002	−9.19E−04
PC-10	−2.61E−02	0.0037	−1.34E−03
PC-11	1.46E−02	−0.005	−2.80E−04
PC-12	−3.76E−02	−0.002	2.31E−05
PC-13	−4.20E−03	0.0016	−1.18E−04
PC-14	−7.22E−02	0.0009	−8.36E−04

(*continued on next page*)

Table A.1 (*continued*).

	gender	*age*	*prefecture*
PC-15	−8.61E−02	0.0002	1.04E−03
PC-16	−9.26E−01	−0.003	−1.11E−03
PC-17	3.53E−01	0.0003	−5.30E−04
PC-18	−3.58E−02	0.0009	−3.28E−04
PC-19	−7.05E−03	0.0004	1.07E−04

Education Industry)

work in the week
or
the end of the course

Game Industry)

Inactivity Duration→ Cutoff (C)

Gamble Industry)

Gambling ⇒ Online Gaming ⇒ Recency

Recency

Game (I)

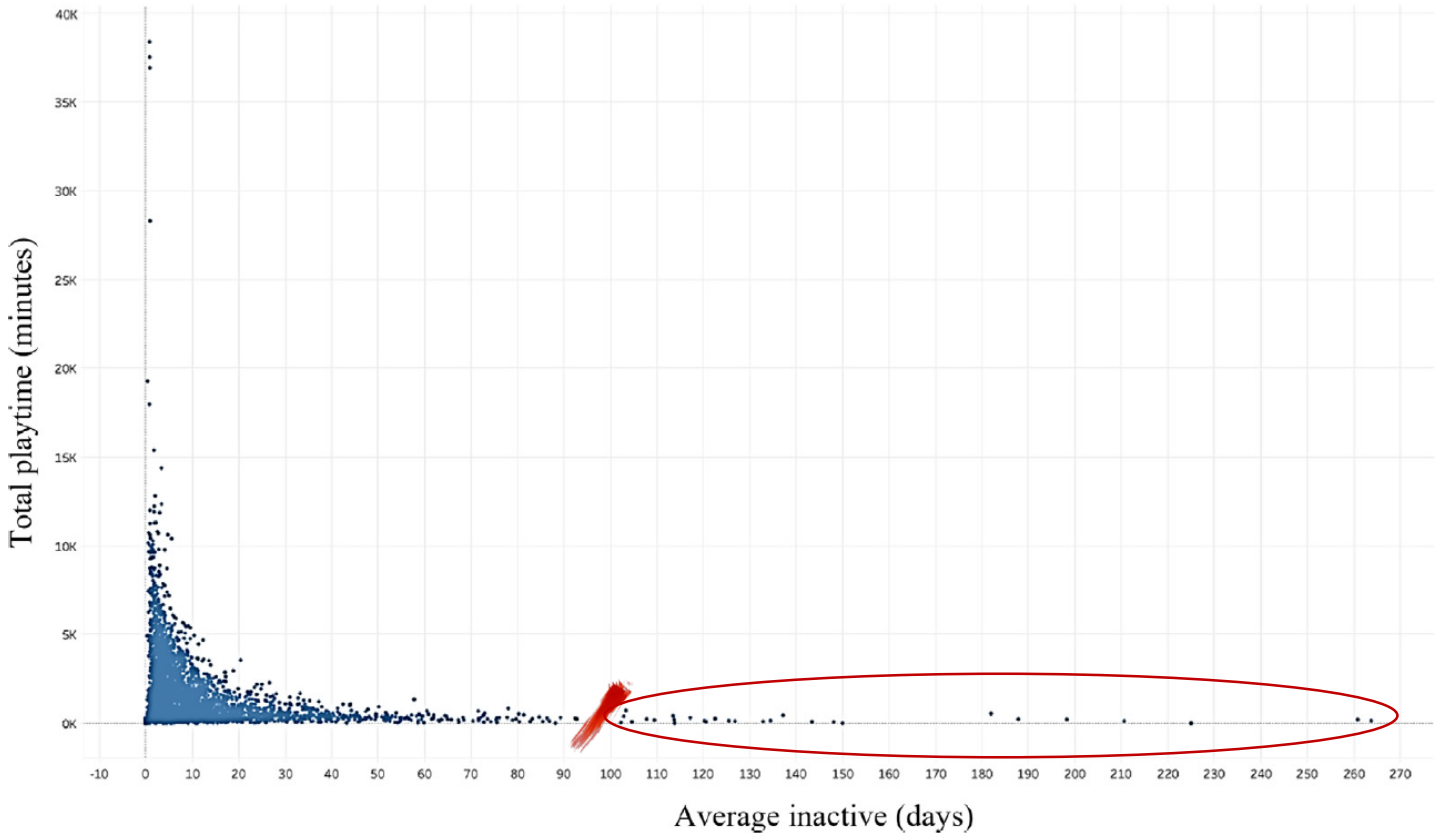


Fig. 1. Scatter plot 1 with average inactive (days) and total playtime.

$$average_{inactive}(Hours) = \left(\frac{1440}{60 \times 24} - \frac{total_{playtime}}{total_{login}} \right) \div 60 = \frac{600}{20} = (1440 - 30) \div 60 = 23.5$$

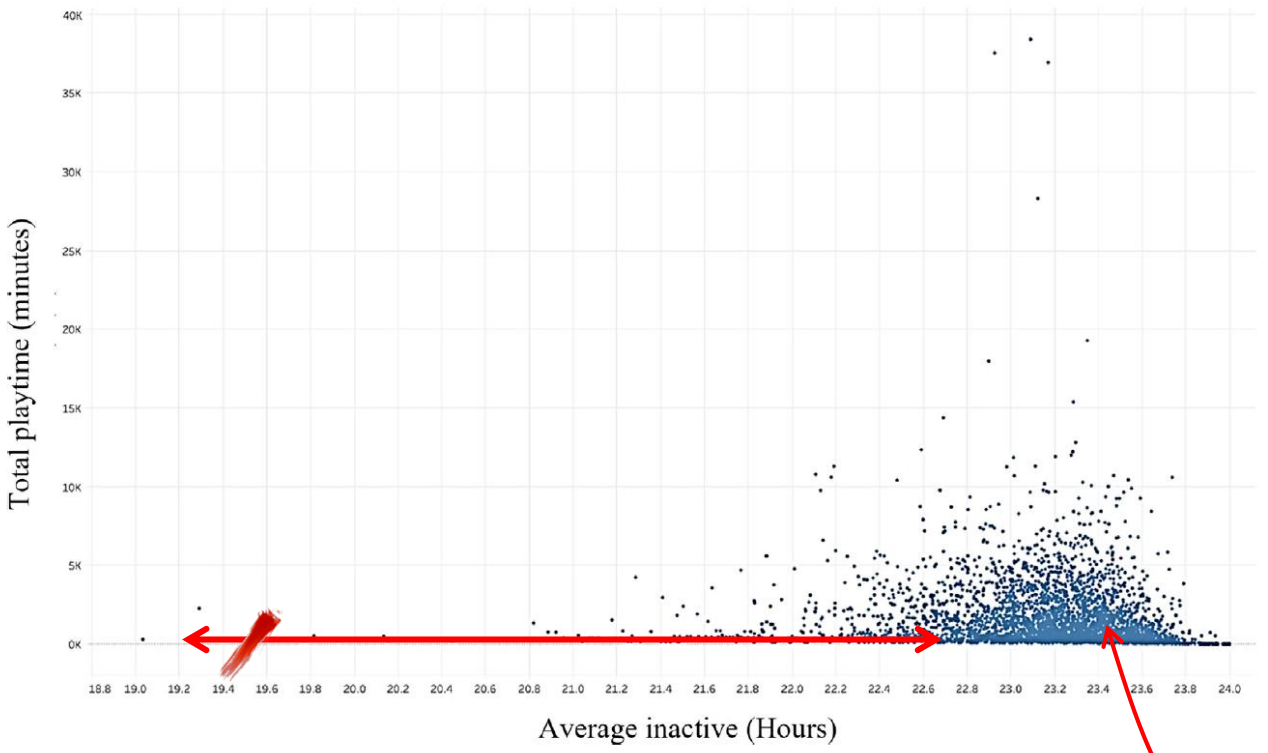


Fig. 2. Scatter plot 2 with average inactive in a day (hours) and total playtime.

Determination of Churn

- Recency
- Avr_inactive
- Std_inactive

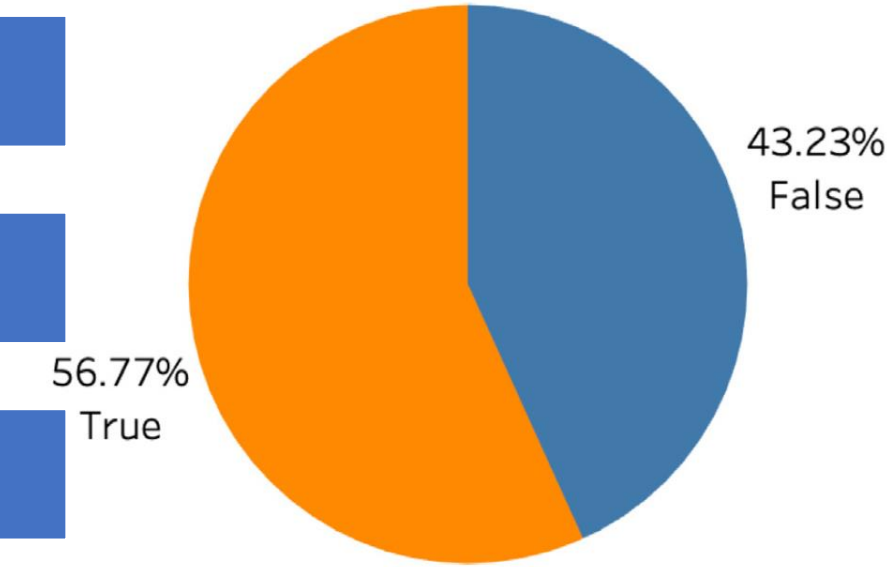


Fig. 3. The proportion of churners (True) and non-churners (False).

Recency = Jan 28th 2020 – Last Login

Cutoff = avr + 2 std

$$\Rightarrow \begin{cases} if\ recency > cutoff: \\ \quad churn\_status = True \\ else: \\ \quad churn\_status = False \end{cases}$$

Descriptive Analysis of the User Logs

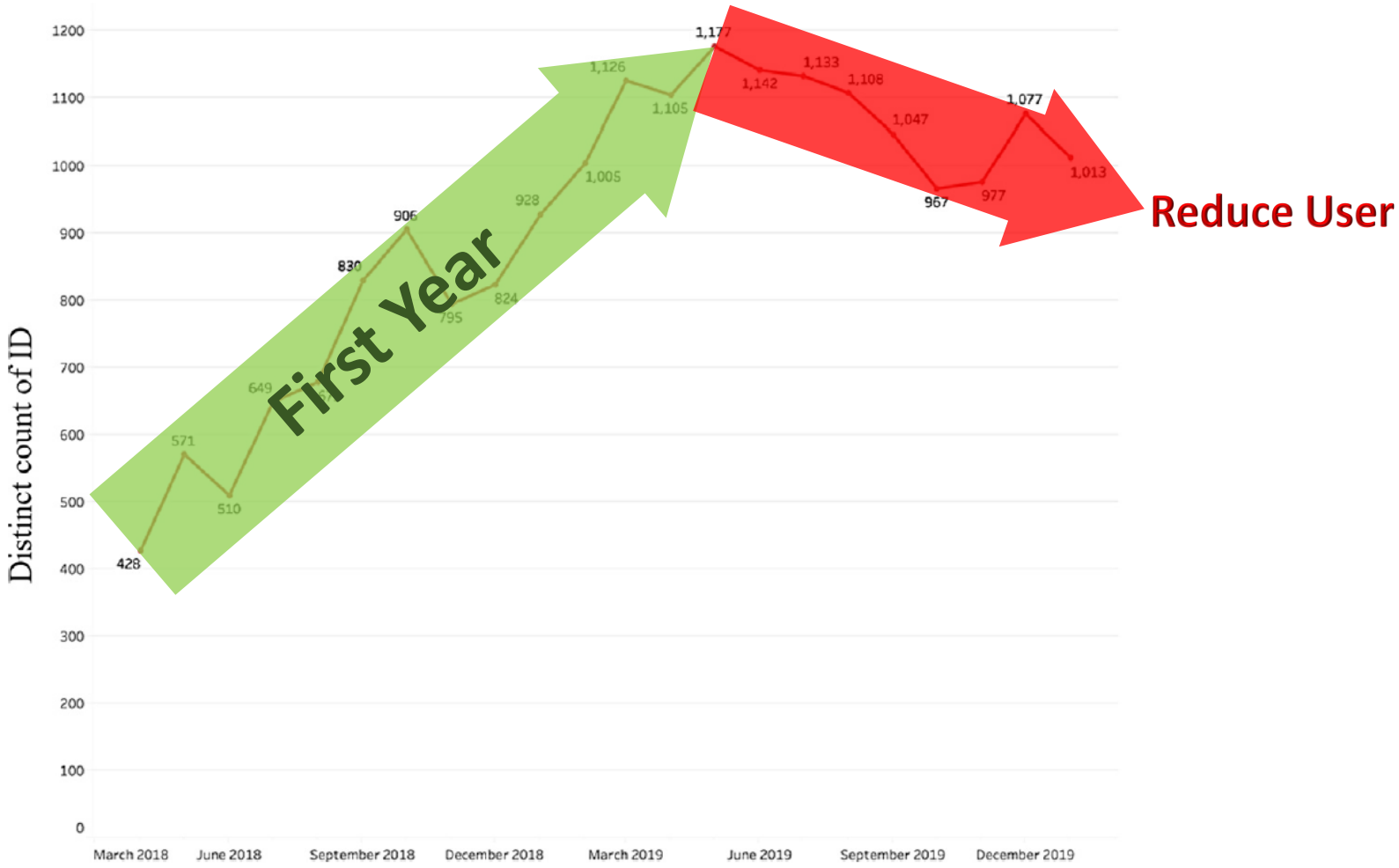


Fig. 4. Monthly active users (MAU).

Descriptive Analysis of the User Logs

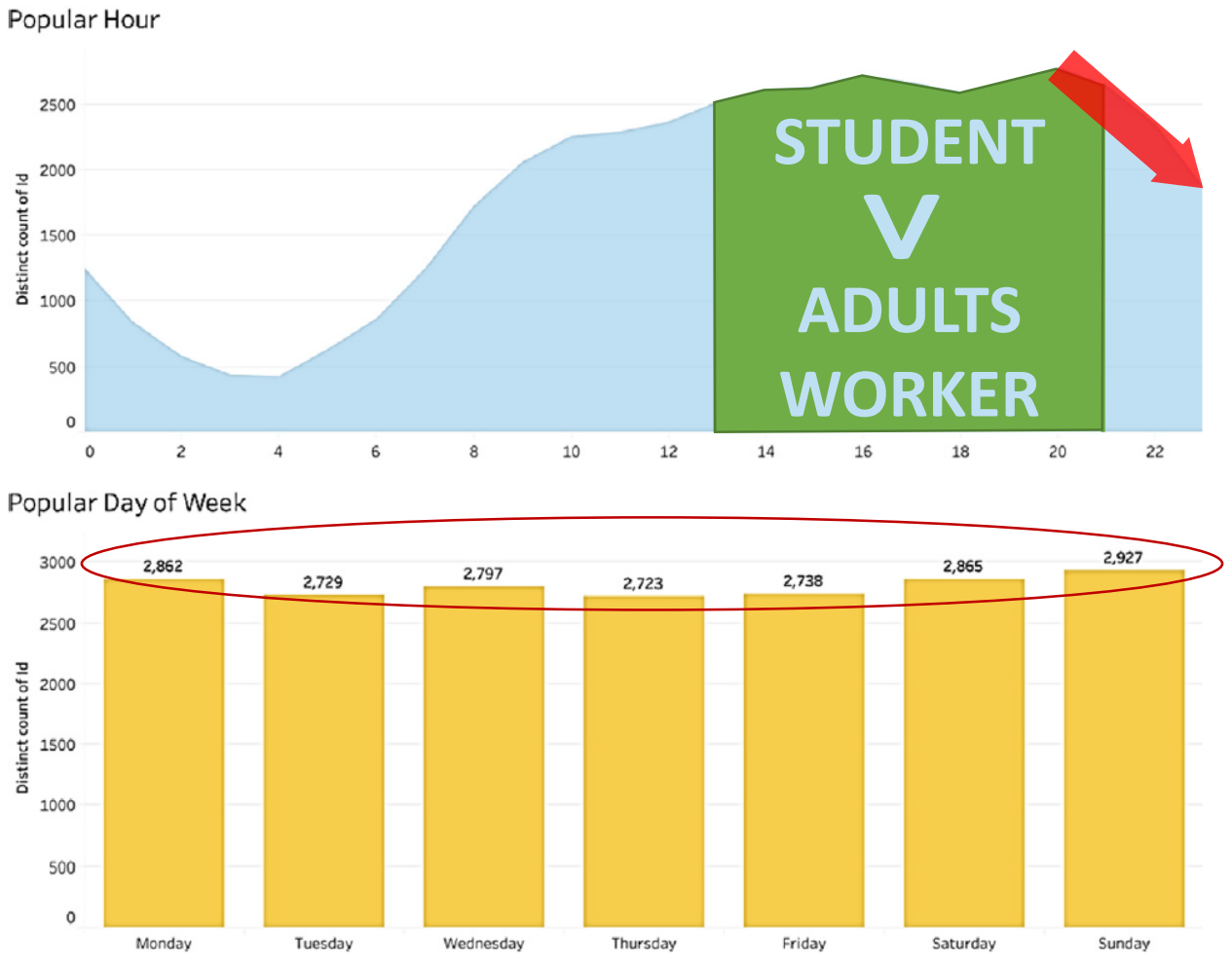


Fig. 5. Popular hour and day of the week.

Retention Analysis by Chapter

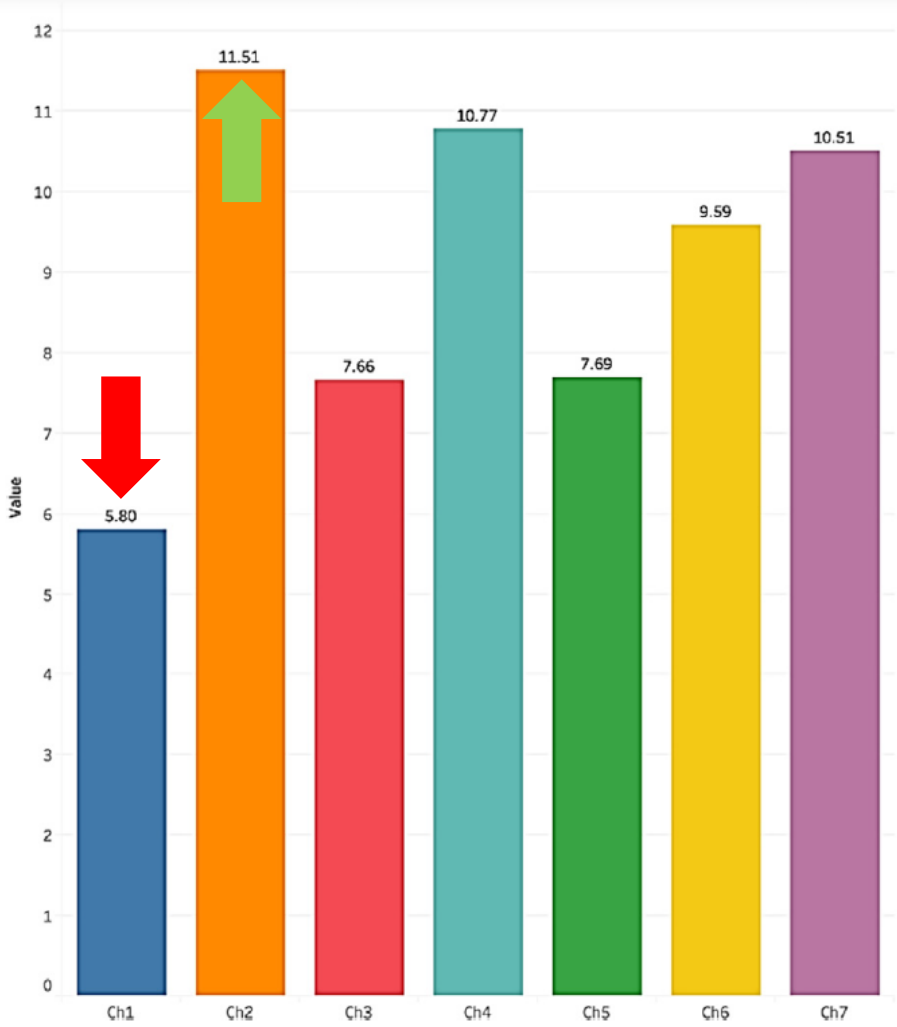


Fig. 6. Median playtime per chapter (hour).

Outlier

Descriptive Analysis of the User Logs

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

Descriptive Analysis of the User Logs

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

25%

Churn

Shortening or simplify its lesson content

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

Churn (in the end)

Examining last content of this chapter

Descriptive Analysis of the User Logs

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

Churn (like chapter 2)

Shortening or simplify its lesson content

Descriptive Analysis of the User Logs

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

Unknown!!!

Adding data

Table 7
AUC results to find out the best dataset and splitting percentage for modeling.

Model	Dataset	AUC		
		80%–20%	85%–15%	90-10%
Decision tree	*features*	0.7106	0.7213	0.7081
	features_pc	0.8345	0.8410	0.8492
Logistic regression	*features*	0.7150	0.7207	0.7222
	features_pc	0.7832	0.7708	0.7808
Random forest	*features*	0.8617	0.8571	**0.9605**
	features_pc	0.9582	0.9584	**0.9605**

Notes: The best result for each model is underlined.
The best result in the table is in boldface.

Decision Tree

Decision Tree

Random Forest

Logistic Regression

Decision Tree

Table 10
Decision tree hyperparameters settings.

Hyperparameter	Description	Values
criterion	It measures the quality of a split.	Gini and entropy
splitter	A strategy that is used to select the split at each node.	best and random
max_depth	The maximum depth of the tree.	[1 - 32]
minimum_samples_split	The minimum number of samples required to split.	The 0.01 to 0.5 are set for the comparison. 30 evenly spaced values between 0.01 and 0.5 are created and evaluated.
minimum_samples_leaf	The minimum number of samples required to be a leaf node.	1 and 30 values between 0.01 and 0.5 are set.

Decision Tree

```
Decision Tree ==========================================
Gini Criteria AUC with Validation:  0.8492491456309462
Entropy Criteria AUC with Validation:  0.853239233455893
Best Splitter AUC with Validation:  0.8492491456309462
Random Splitter AUC with Validation:  0.8183504809576065
```

Fig. 7. The result of AUC with different criteria and splitter values.

Decision Tree

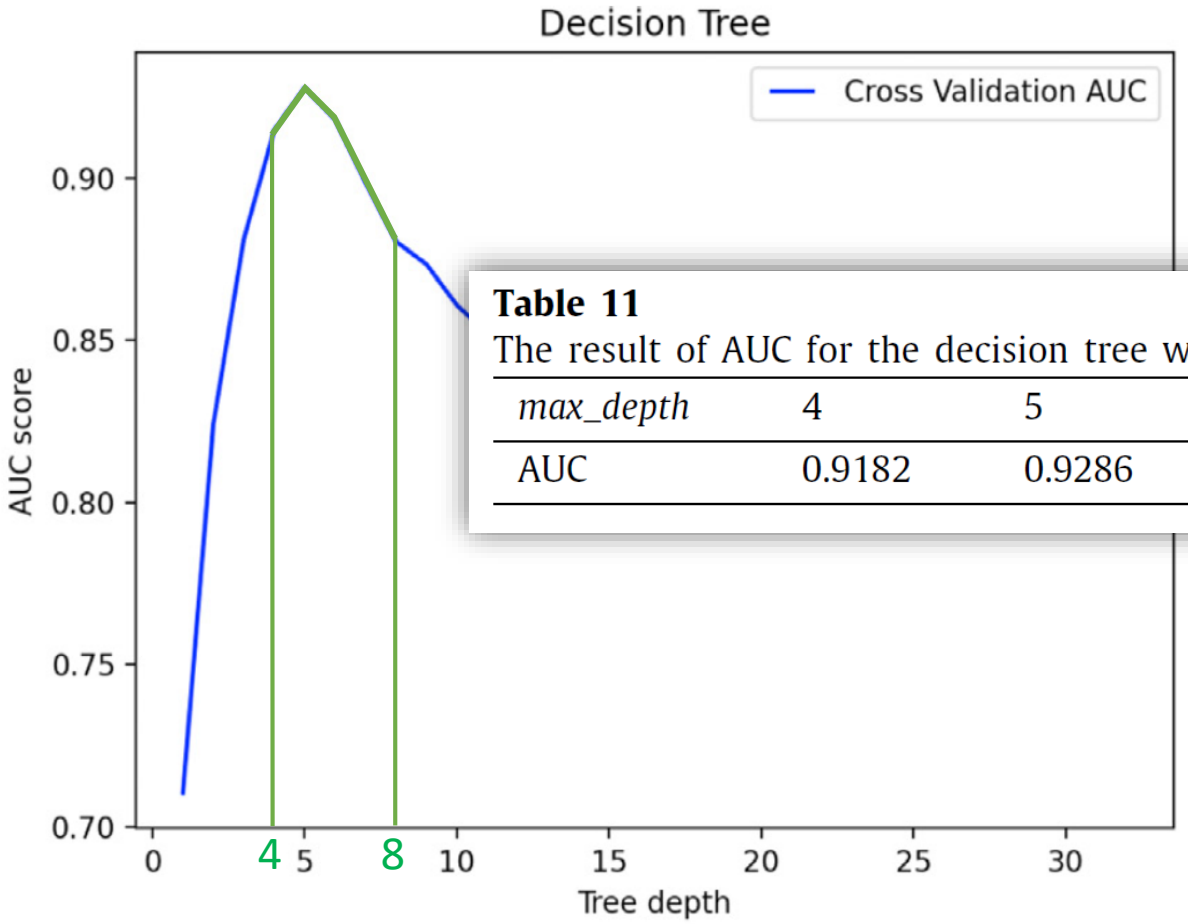
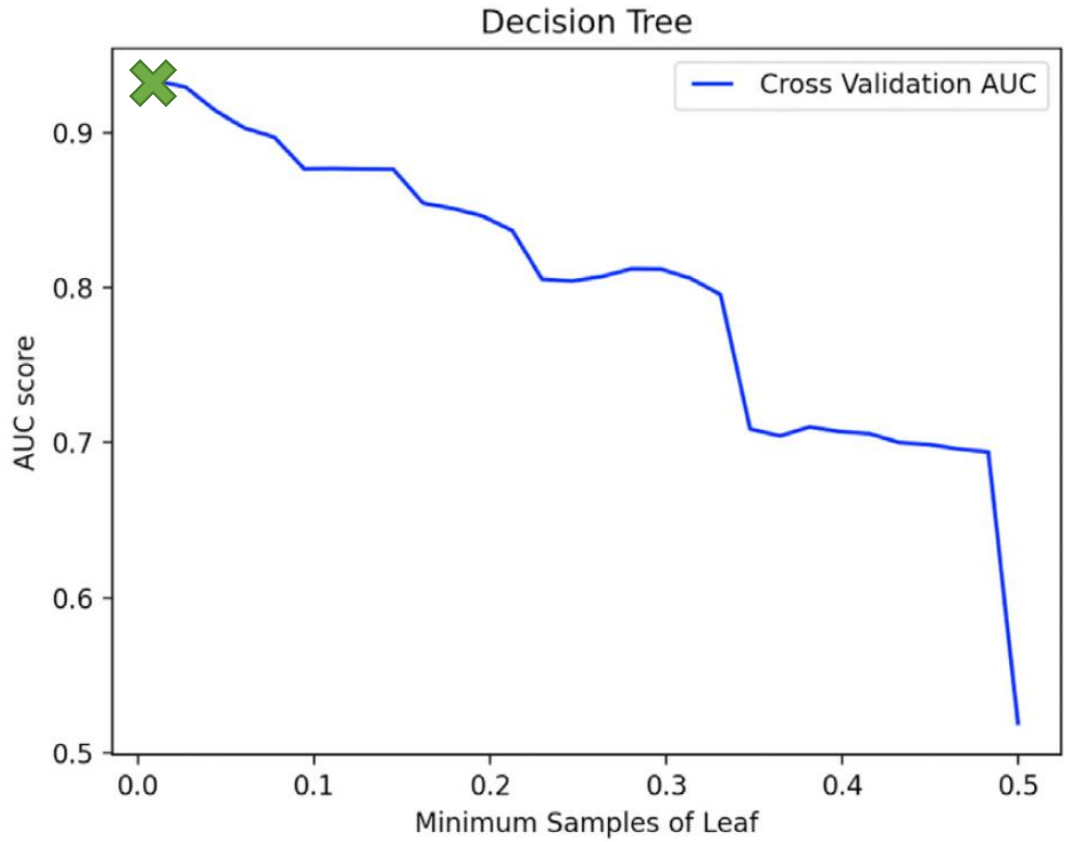


Table 11
The result of AUC for the decision tree with different *max_depth* values.

max_depth	4	5	6	7	8
AUC	0.9182	0.9286	**0.9290**	0.9286	0.9289

(a) With different *max_depth* values.

Decision Tree



(c) With different *minimum_samples_leaf* values.

$$best\ performance = smallest\ value$$

Decision Tree

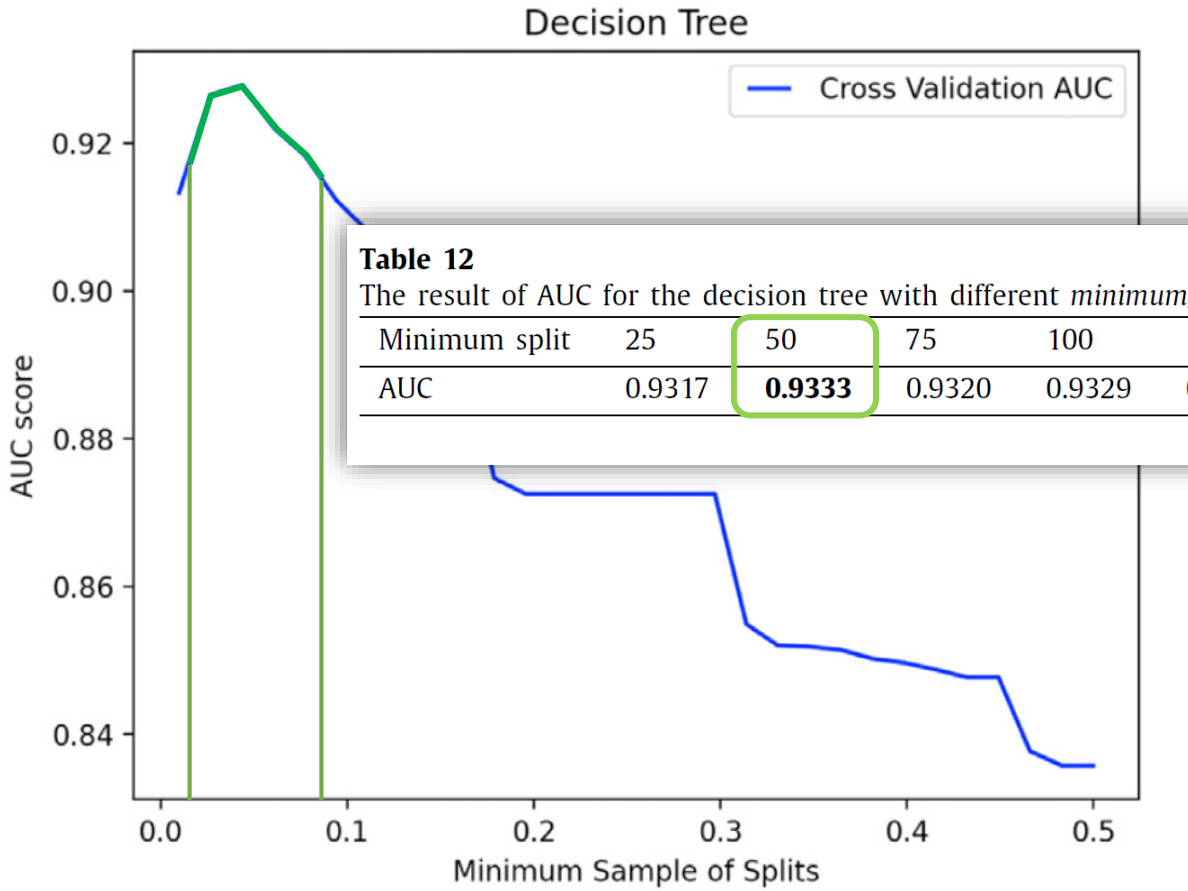
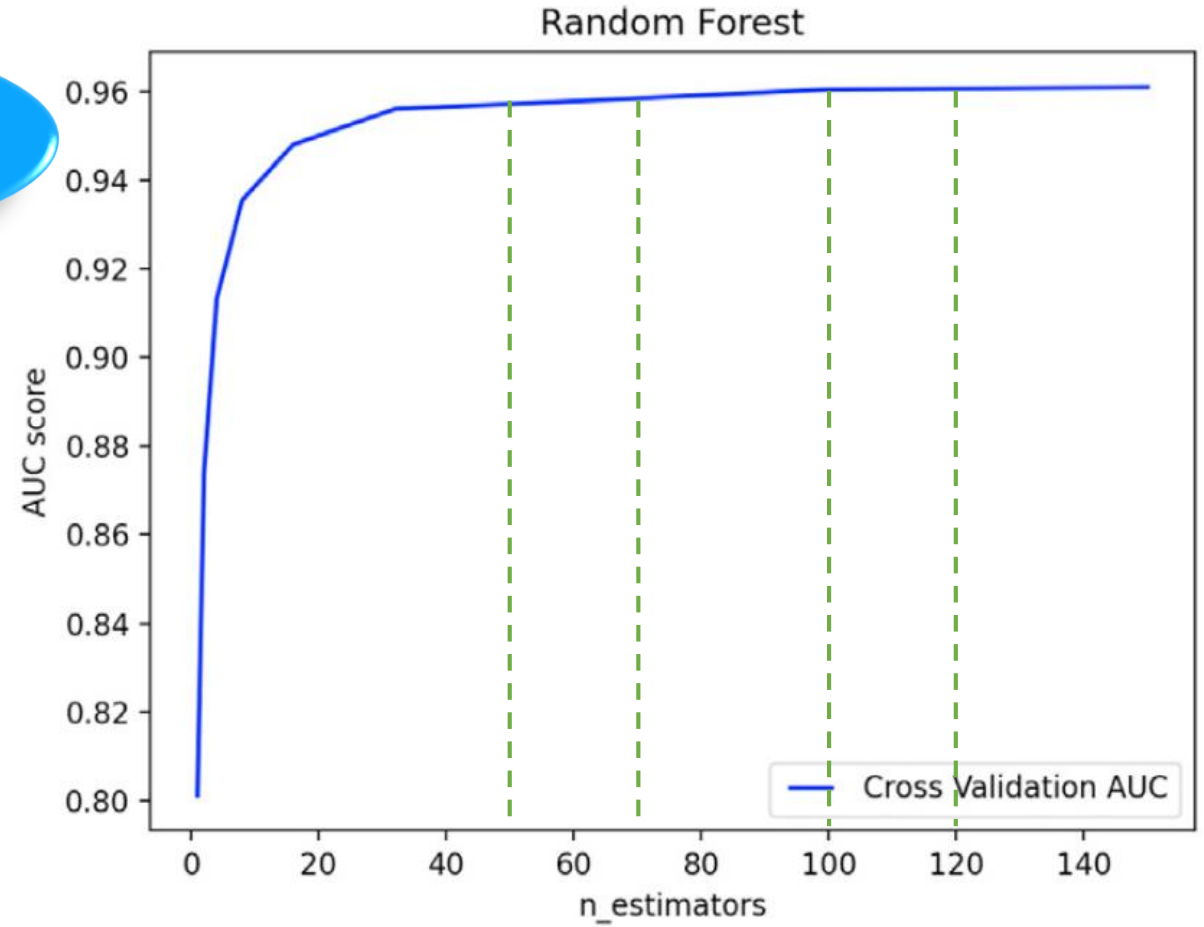


Table 12
The result of AUC for the decision tree with different *minimum_samples_split* values.

Minimum split	25	50	75	100	125	150	175	200
AUC	0.9317	**0.9333**	0.9320	0.9329	0.9290	0.9284	0.9258	0.9262

(b) With different *minimum_samples_split* values.

Random Forest



(a) With different estimators.

Random Forest

Table 13
The result of AUC for the random forest with different estimators.

n_estimators	50	75	100	125
AUC	0.9570	0.9590	**0.9605**	0.9603

Random Forest

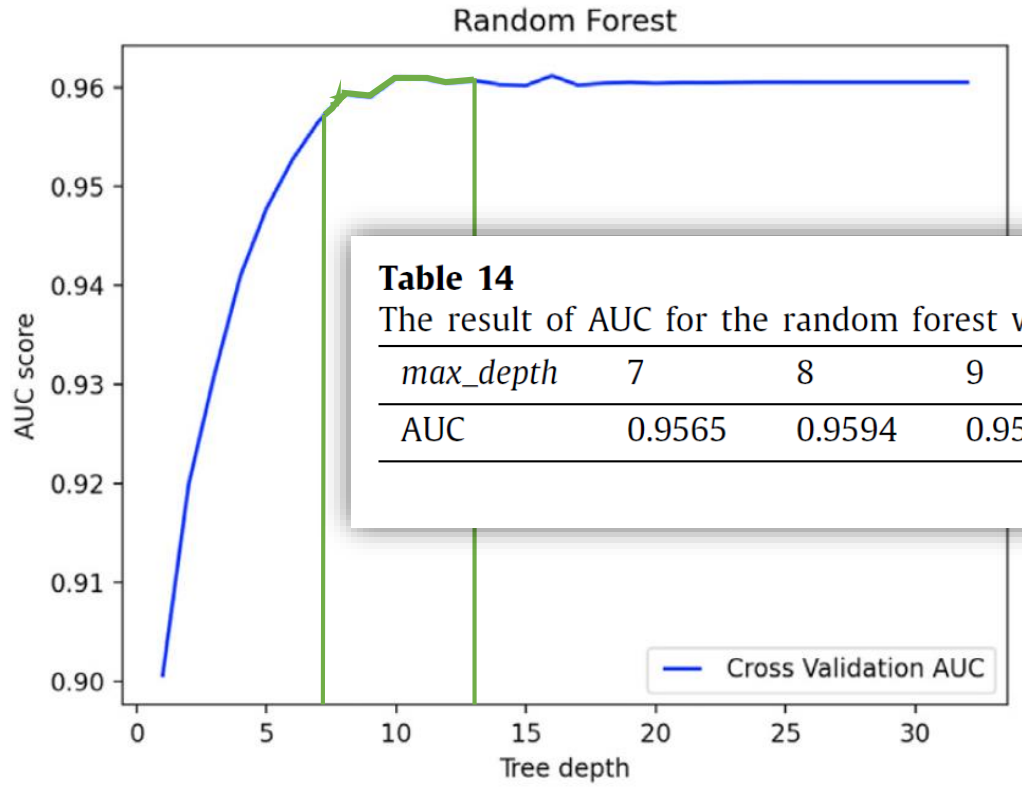
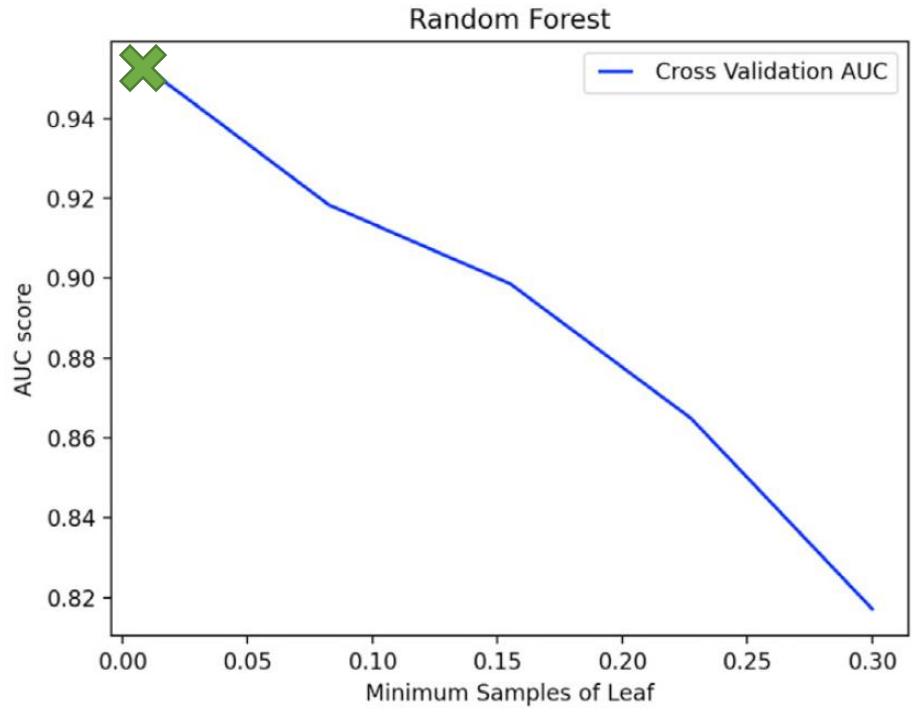


Table 14
The result of AUC for the random forest with different *max_depth* values.

max_depth	7	8	9	10	11	12	13
AUC	0.9565	0.9594	0.9591	**0.9610**	0.9609	0.9605	0.9607

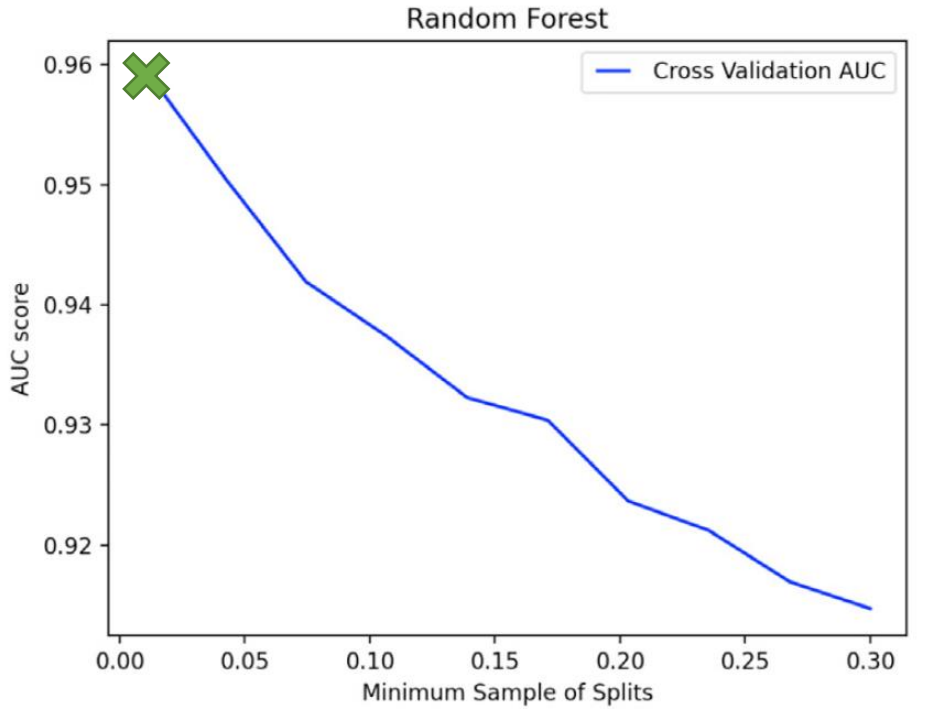
(b) With different *max_depth* values.

Random Forest



(d) With different $minimum\_samples\_leaf$ values.

$$best\ performance = smallest\ value$$



(c) With different *minimum_samples_split* values.

$$best\ performance = smallest\ value$$

Logistic Regression

Penalty (L1 , L2)

Inverse of regularization strength (C)

```
Logistic Regression ========================
Penalty L1 AUC :  0.9745084802343669
Penalty L2 AUC :  0.7807881974415437
Penalty None AUC :  0.7807922014455476
```

Fig. 10. The result of AUC with different penalty values.

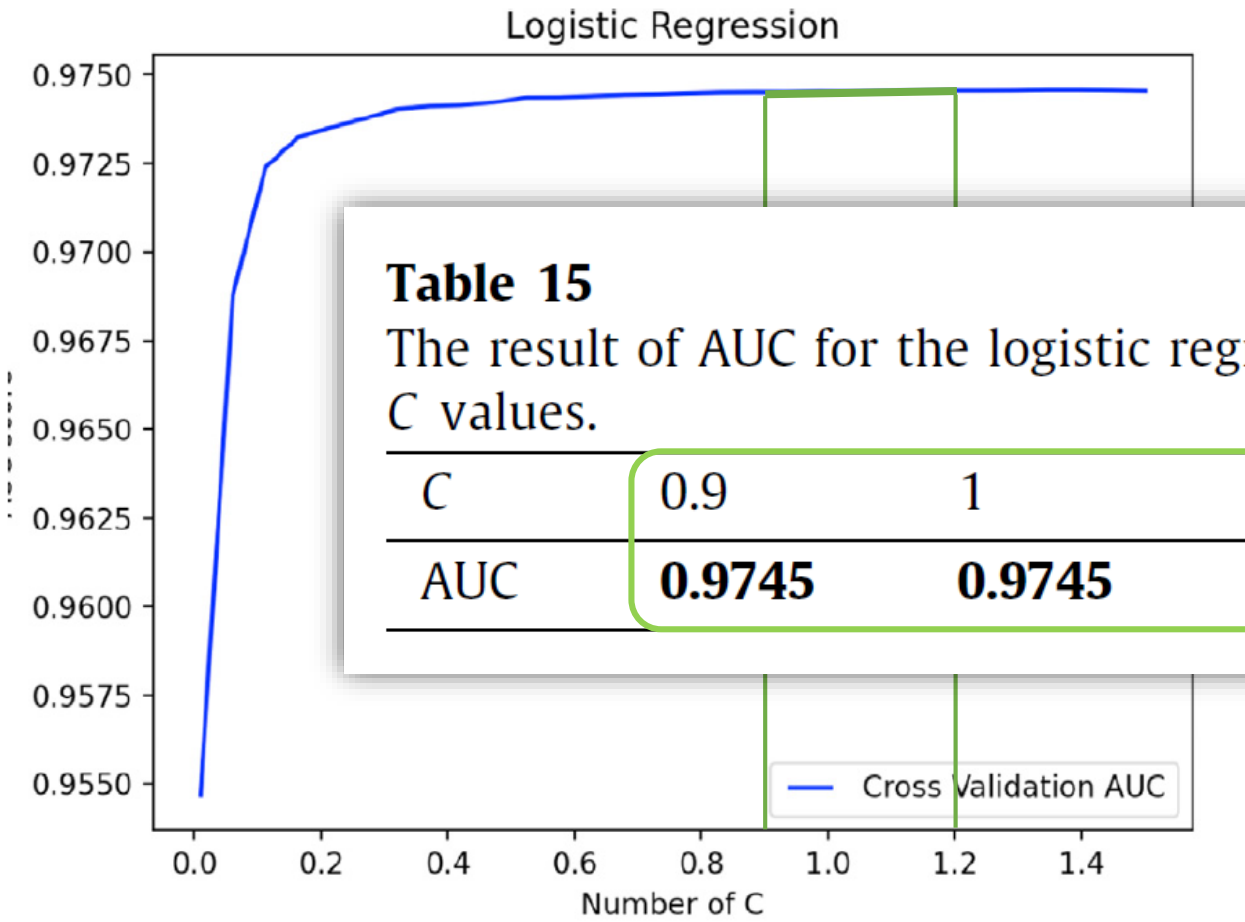


Table 15
The result of AUC for the logistic regression with different C values.

C	0.9	1	1.1	1.2
AUC	**0.9745**	**0.9745**	**0.9745**	**0.9745**

Fig. 11. The result of AUC with different C values.

Hyperparameter Optimization

Decision Tree

$criterion$ = "entropy"
$splitter$ = best
max_depth = 6
$min\_samples\_split$ = 50
$min\_samples\_leaf$ = 1

Random Forest

$criterion$ = "entropy"
$splitter$ = best
max_depth = 10
$min\_samples\_split$ = 2
$min\_samples\_leaf$ = 1
$n\_estimators$ = 100

Logistic Regression

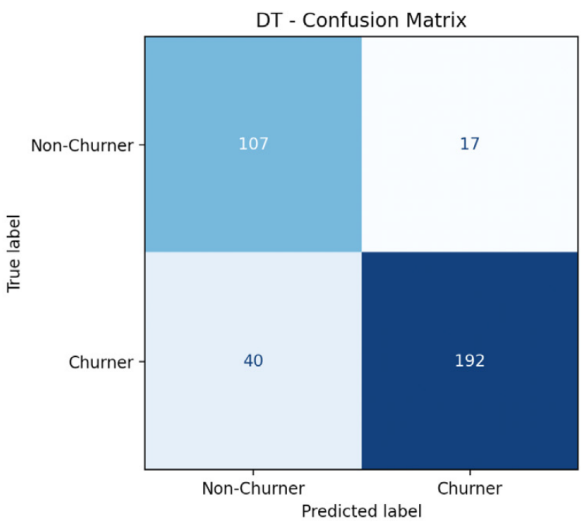
$penalty$ = "L1"
C = 0.9

Table 7
AUC results to find out the best dataset and splitting percentage for modeling.

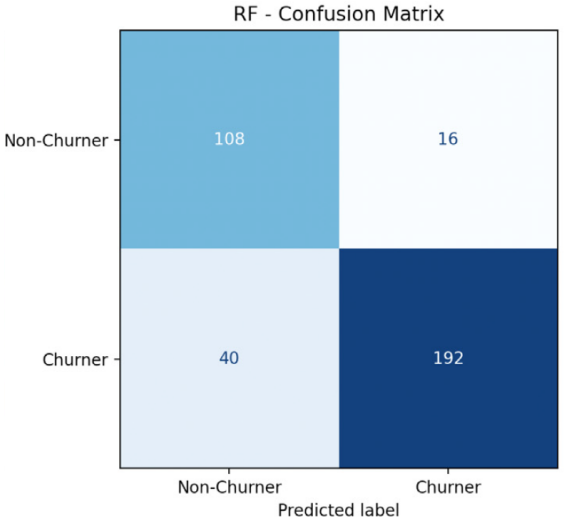
Model	Dataset	AUC		
		80%–20%	85%–15%	90-10%
Decision tree	*features*	0.7106	0.7213	0.7081
	features_pc	0.8345	0.8410	<u>0.8492</u>
Logistic regression	*features*	0.7150	0.7207	0.7222
	features_pc	<u>0.7832</u>	0.7708	0.7808
Random forest	*features*	0.8617	0.8571	**<u>0.9605</u>**
	features_pc	0.9582	0.9584	**<u>0.9605</u>**

Notes: The best result for each model is underlined.
The best result in the table is in boldface.

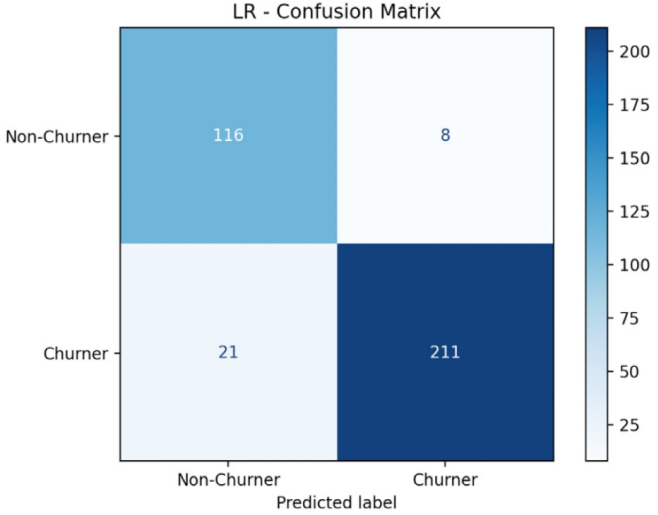
Model Evaluation



(a) Decision tree.



(b) Random forest.



(c) Logistic regression.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+TN}$$

$$\text{F1-score} = 2 \times \frac{precision\,.\,recall}{precision+recall}$$

Model Evaluation

Decision Tree ==========
[Test] ROC AUC of DT: 0.8452447163515017
▼▼▼ [Test] Classification Report of DT ▼▼▼

	precision	recall	f1-score	support
0	0.7279	0.8629	0.7897	124
1	0.9187	0.8276	0.8707	232
accuracy			0.8399	356
macro avg	0.8233	0.8452	0.8302	356
weighted avg	0.8522	0.8399	0.8425	356

(a) Decision tree.

Random Forest ==========
[Test] ROC AUC of RF: 0.8492769744160178
▼▼▼ [Test] Classification Report of RF ▼▼▼

	precision	recall	f1-score	support
0	0.7297	0.8710	0.7941	124
1	0.9231	0.8276	0.8727	232
accuracy			0.8427	356
macro avg	0.8264	0.8493	0.8334	356
weighted avg	0.8557	0.8427	0.8453	356

(b) Random forest.

Logistic Regression ==========
[Test] ROC AUC of LR: 0.9224833147942157
▼▼▼ [Test] Classification Report of LR ▼▼▼

	precision	recall	f1-score	support
0	0.8467	0.9355	0.8889	124
1	0.9635	0.9095	0.9357	232
accuracy			0.9185	356
macro avg	0.9051	0.9225	0.9123	356
weighted avg	0.9228	0.9185	0.9194	356

(c) Logistic regression.

Fig. 13. Prediction results. Non-churner class is represented by (0) and the churner class by (1). The values 124 and 232 in the support column represents the number of samples for non-churner and churner, respectively.

$$\frac{232}{356} = 65.16\% \quad > \quad 56.7\%$$

More bias ⇒ Weighted avg

Table 16
Results summary.

Metrics/Algorithms	Decision tree	Random forest	Logistic regression
Precision	0.8522	0.8557	**0.9228**
Recall	0.8399	0.8427	**0.9185**
F1-Score	0.8425	0.8453	**0.9194**
AUC	0.8452	0.8493	**0.9225**

Logistic regression

LR is Better than DT and RF

- LR tends to overfit with high dimensional data, but a low dimensional dataset was used.
- LR works well with linearly separatable datasets, whereas DT has less efficiency with continuous numerical data.
- Many of the used dataset variables are continuous numerical variables with a wide-range and more significant gap between mean and standard deviation, so this could affect the tree-based models, and worked better with LR model than DT and RF.

Discussion

- The defined churn can be helpful to many DGBL services.
- Three categories (demographic, engagement, and performance) can be implemented with any DGBL company.
- This approach allows us to have the flexibility to define churners even if the users are playing in mid-course.
- The applicability and flexibility of the proposed churn determination can handle the challenges.

Discussion

About half the students we surveyed said educational games were an effective way to learn new material.

Source: Paper, 2023.

Thanks for your attention

Churn Prediction in digital game-based learning using data mining techniques: Linear regression, decision tree, and random forest