پیش‌بینی ریزش در یادگیری مبتنی بر بازی دیجیتال با استفاده از تکنیک‌های داده کاوی: رگرسیون لجستیک، درخت تصمیم و جنگل تصادفی

دانشگاه فردوسی مشهد

دانشکده: علوم ریاضی

رشته: ریاضی کاربردی

گرایش: علوم داده

درس: نظریه یادگیری

استاد: دکتر جلال الدین نصیری

نویسنده: پرهام پیشرو

شماره دانشجویی: ۴۰۱۱۳۰۹۰۱۲

چکیده

در این فایل مستند که برای ارائه پاورپوینت آماده شده است، اطلاعات و توضیحات لازم برای فردی که قصد توضیح این ارائه را برای دیگران دارد ولی اطلاع کافی از مقاله و ارائه موجود ندارد، آورده شده است. مطالب این مستند در ۸ فصل به همراه یک مقدمه و یک پیوست تنظیم شده است. فصل‌های مورد نظر به ترتیب عبارتند از: آماده‌سازی داده، تعریف ریزش، تعیین ریزش، تحلیل بقا، بهینه‌سازی هایپرپارامتر، پیش‌بینی ریزش و ارزیابی، بحث و نتیجه‌گیری و مراجع.

تعداد اسلایدهای فایل ارائه پاورپوینت برابر با ۵۰ اسلاید، تعداد اسلایدهای فایل پی دی اف منطبق بر پاورپوینت برابر با ۵۳ اسلاید و تعداد صفحات اصلی خود مقاله مورد بررسی برابر با ۲۰ صفحه می‌باشد. هم‌چنین تعداد صفحات این فایل مستند نیز برابر با ۳۶ صفحه است.

مقدمه

برای دانشجویان ارشد علوم داده، مفاهیمی مثل پیش‌بینی ریزش[۱] و تکنیک‌های داده کاوی[۲] آشناست. عبارت نا آشنایی که در عنوان این مقاله جای دارد، یادگیری مبتنی بر بازی دیجیتال[۳] می‌باشد. یادگیری مبتنی بر بازی دیجیتال یکی از مقوله‌های فناوری آموزشی[۴] است؛ پس بهتر آن است که قبل از پرداختن به این موضوع یک نگاه رو به گذشته داشت و با این حوزه فناوری آموزشی آشنا شد.

فناوری آموزشی صنعتی است که آموزش و پیشرفت‌های فناوری را با یک‌دیگر ادغام می‌کند. در واقع هرگاه آموزش با کمک تکنولوژی های روز دنیا صورت گیرد، از فناوری آموزشی استفاده شده است. به طور مثال می‌توان به تخته‌های هوشمند، قلم نوری، پرژکتورهای موجود در کلاس، دوربین‌های کنفرانس، نرم افزارهای آموزشی و درسی، وبینار، نرم افزارهای ارائه مطالب و ... اشاره کرد. بخش آموزش به سرعت در حال دیجیتالی شدن است. هدف این صنعت جذاب‌تر و راحت‌تر کردن روند آموزش می‌باشد. [۱]

آموزش مبتنی بر بازی دیجیتال یک روش است که در آن از ایده بازی کردن برای رسیدن به اهداف خاص آموزشی اعم از کسب دانش، مهارت و یا نگرش‌های خاص استفاده می‌کند. یا به طور خلاصه میتوان گفت که فرد بازی می‌کند تا یاد بگیرد. نتایج استفاده از این استراتژی ثابت کرده است که استفاده از بازی‌ها، اثربخشی بسیار بالایی دارد. [۲] این روش فواید زیادی دارد که به ۶ مورد از آن فواید در زیر اشاره شده است:

۱) **یادگیری در هنگام تفریح:** این مورد باعث می‌شود که فرایند آموزش خسته‌کننده نباشد.
۲) **افزایش انگیزه:** افرادی که از این روش استفاده می‌کنند انگیزه‌ی بیشتری نسبت به دیگر افراد دارند که صرفاً به صورت سنتی مشغول به یادگیری هستند. این روش باعث می‌شود که فرد بازه‌های زمانی بیشتری به یادگیری اختصاص دهد.
۳) **توسعه توانایی‌های شناختی:** این روش باعث رشد مهارت‌های مرتبط با تفکر و منطق فرد می‌شود.
۴) **تقویت مهارت:** بازی کردن باعث تقویت مهارت می‌شود. به طور مثال بازی‌های استراتژیک باعث تقویت قدرت تفکر، برنامه‌ریزی و آینده‌نگری می‌شود. یا بازی‌های اول شخص اکشن باعث تقویت سرعت عمل افراد می‌شود.
۵) **افزایش دانش:** با توجه به این که فرد مدت زمان بیشتری به یادگیری مشغول است، دانش بیشتری در مقایسه با دیگر افراد خواهد داشت.
۶) **تغییر نگرش:** به روز شدن روند یادگیری باعث تغییر دیدگاه فرد می‌شود.

۶ مورد فوق تنها بخشی از فواید حوزه یادگیری مبتنی بر بازی دیجیتال می‌باشد. [۳]

نکته حائز اهمیت این است که هر صنعتی برای رشد و پویا بودن خود نیاز به کسب و حفظ کاربر دارد؛ همچنین با توجه گسترش روز افزون بازار، اهمیت رویکرد بازاریابی در این حوزه افزایش یافته است. برخی از شاخص‌های کلیدی و مهم رویکرد بازاریابی عبارتند از ریزش[۵]، حفظ[۶] و پیش‌بینی ریزش.

- **ریزش:** ریزش مشتری درصد مشتریانی است که استفاده از خدمات یا محصول را در یک دوره معین متوقف کرده‌اند.

---

۱ Churn Prediction
۲ Data Mining
۳ Digital Game-based Learning (DGBL)
۴ Educational Technology (Ed Tech)
۵ Churn
۶ Retention

- **حفظ:** اگر مشتری ریزش نکند، در واقع حفظ شده است.
- **پیش‌بینی ریزش:** به روند پیش‌بینی کردن ریزش‌کننده ها می‌گویند.

متاسفانه نرخ ریزش در حوزه فناوری آموزشی به طور معمول بالاتر بوده است. طبق گزارش واشنگتن پست در مورد آموزش مجازی در سال ۲۰۱۹ که به تفاوت نرخ فارغ‌التحصیلی بین مدارس عادی و مجازی پرداخته است؛ نرخ فارغ التحصیلی مدارس مجازی تنها ٪ ۵۰/۱ بوده در حالی که نرخ کلی فارغ‌التحصیلی در ایالات متحده ٪ ۸۴ می‌باشد. [۴] علاوه بر این، میانه نرخ ریزش در صنعت آموزش ٪ ۱۰/۲۹ بوده است که رتبه سوم را از بین ۹ دسته به لحاظ زیاد بودن نرخ ریزش به خود اختصاص می‌دهد. [۵]

در حوزه یادگیری مبتنی بر بازی دیجیتال هیچ تحقیقی در مورد پیش‌بینی ریزش و حتی تعیین نرخ ریزش وجود ندارد. ولی می‌توان از تحقیقات در موردی بازی آنلاین[۷] و صنعت آموزش تا حدی استفاده کرد؛ زیرا DGBL ترکیبی از بازی آنلاین و آموزش می‌باشد. کارهای گذشته[۸] که در این رابطه صورت گرفته در فایل پاورپوینت به نمایش گذاشته شده است. همان‌طور که قابل مشاهده است فقط یک پژوهش با موضوع DGBL در کارهای گذشته وجود دارد که به دقت نزدیک ٪۶۰ رسیده است و بقیه پژوهش‌ها در زمینه‌های مشابه یعنی بازی و آموزش می‌باشد. اگر چه DGBL ترکیبی از آموزش و بازی است؛ ولی ذکر این نکته خالی از اهمیت نیست که بین این حوزه‌ها تفاوت‌های بسیاری وجود دارد. در نتیجه نمی‌توان از مفهوم ریزش موجود در این حوزه‌ها بهره گرفت و باید یک مفهوم ریزش جدید برای DGBL پیدا کرد.

پس از ذکر همه نکات لازم، به معرفی محصول ساخته شده و روش جمع‌آوری داده باید پرداخت. داده‌ی کاربران توسط یک شرکت ژاپنی ارائه شده است که داده‌های اولیه هستند (بدین معنی که داده‌ها کثیف[۹] هستند). این شرکت یک سرویس برنامه‌نویسی آنلاین ارائه می‌دهد که مانند یک بازی نقش آفرینی[۱۰](RPG) است. بازی نقش آفرینی اینگونه است که یک بازیکن قابلیت و کنترل یک شخصیت درون بازی را در دست می‌گیرد و با کمک صحبت با شخصیت‌های دیگر، جابه‌جایی در مناطق مختلف، حل پازل و یادگیری کدنویسی برای ایجاد جادو در داستان حرکت می‌کند. یکی از معروف‌ترین بازی‌های RPG بازی سوپر ماریو[۱۱] می‌باشد که بازیکن کنترل ماریو را در بازی به دست می‌گیرد.

پس از گفتن همه مقدمات لازم، باید یک دید کلی بر روی مباحث پیش رو داشت تا روند کلی ارائه را متوجه شد. فهرست زیر نشان‌دهنده‌ی مباحث می‌باشد:

- **فصل اول:** آماده‌سازی داده[۱۲]
- **فصل دوم:** تعریف ریزش[۱۳]
- **فصل سوم:** تعیین ریزش[۱۴]
- **فصل چهارم:** تحلیل بقا[۱۵]
- **فصل پنجم:** بهینه‌سازی هایپرپارامتر[۱۶]

---

[۷] Online Gaming
[۸] Related Works
[۹] Untidy, Messy
[۱۰] Role Play Game
[۱۱] Super Mario
[۱۲] Data Preparation
[۱۳] Churn Definition
[۱۴] Churn Determination
[۱۵] Retention Analysis
[۱۶] Hyperparameter Optimization

- **فصل ششم:** پیش‌بینی ریزش و ارزیابی[۱۷]
- **فصل هفتم:** بحث[۱۸] و نتیجه‌گیری[۱۹]

۱ آماده‌سازی داده

در این فصل، روش جمع‌آوری داده[۲۰]، انتخاب داده[۲۱]، تجمیع داده[۲۲]، تبدیل داده[۲۳]، ادغام داده[۲۴]، تحلیل اکتشافی داده[۲۵]، پاکسازی داده[۲۶]، معرفی ویژگی‌ها و هدف[۲۷] و انتخاب ویژگی[۲۸] مورد بررسی قرار خواهد گرفت.

۱,۱ جمع آوری داده

همان‌طور که گفته شد داده‌ها از یک شرکت ژاپنی گرفته شده که یک سرویس برنامه نویسی آنلاین همانند بازی ارائه می‌دهد. دو نوع داده از این سرویس در دسترس بود: داده‌های ساختار یافته از پایگاه داده مدیریت ارتباط با مشتری[۲۹] و داده‌های لاگ نیمه ساختار یافته‌ی کاربر[۳۰]. محتوای این داده‌ها را می‌توان در جدول ۴ مشاهده کرد:

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.

[۱۷] Evaluation
[۱۸] Discussion
[۱۹] Conclusion
[۲۰] Data Collection
[۲۱] Data Selection
[۲۲] Data Aggregation
[۲۳] Data Transformation
[۲۴] Data Integration
[۲۵] Exploratory Data Analysis (EDA)
[۲۶] Data Cleaning
[۲۷] Features and Target
[۲۸] Feature Selection
[۲۹] Customer Relationship Management (CRM)
[۳۰] Semi-Structured User Log Data

تعداد شرکت‌کنندگان ۳۵۵۷ نفر بودند و توزیع جنسیتی آن‌ها به این صورت بود که حدود ۵۶ درصد از افراد زن و ۴۴ درصد مرد بودند. همچنین از لحاظ سنی نیز حدود ۵۰ درصد از بازیکنان بین ۱۷ تا ۳۳ سال سن دارند؛ ۲۵ درصد بازیکنان زیر ۱۷ و ۲۵ درصد دیگر بالای ۳۳ سال هستند.

۲,۱ انتخاب داده

طبق مطالعات گذشته که از داده‌های کاربر استفاده کردند، دسته‌بندی‌های رایجی از متغیرها وجود دارد. و از آنجایی که داده‌های کاربر در DGBL مشابه با صنعت بازی است، بیشتر انتخاب داده‌ها تحت تاثیر همین صنعت بازی می‌باشد. با توجه به حجم زیاد داده‌ها و جداول متعددی که وجود دارد، انتخاب همه آن‌ها و ادغام کردنشان کاری اشتباه بوده و هزینه بسیاری دارد. در نتیجه با توجه اهداف مورد نظر فقط باید بخشی از داده‌ها را برای تجمیع، تبدیل و ادغام داده‌ها انتخاب کرد. داده‌های مختلفی برای این هدف انتخاب شدند که در زیر به چند مورد از آن‌ها اشاره شده است.

اکثر بازی‌های موجود از تعداد ورود کاربران به بازی، مدت زمان بازی، تعداد کلیک‌های بازیکنان و مقادیر مشابه استفاده می‌کنند. سه متغیر شناسه کاربر[۳۱]، مهر زمانی[۳۲] و محتوا[۳۳] از میان داده‌های لاگ برای محاسبه میانگین تعداد دفعات ورود، مدت زمان بازی و میانگین تعداد کلیک‌ها انتخاب شده است.

هم‌چنین مورد مهم بعدی، عملکرد هر بازیکن می‌باشد. سطح[۳۴] هر بازیکن و تعداد سکه‌های آن فرد برای به دست آوردن ویژگی‌های عملکرد[۳۵] بازیکن نیز انتخاب شده اند. علاوه بر مواردی که در بالا ذکر شد، جنسیت[۳۶]، تاریخ تولد[۳۷]، محل سکونت[۳۸] از میان پایگاه داده مدیریت ارتباط با مشتری به‌عنوان ویژگی‌های جمعیت شناختی[۳۹] نیز انتخاب شده اند.

۳,۱ تجمیع، تبدیل و ادغام داده

پس از انتخاب داده‌های مورد نیاز، باید این داده‌ها را با توجه به صورت مسئله تجمیع، تبدیل و ادغام کرد. در این مرحله نیز چندین کار انجام شد که فقط به برخی از آن‌ها اشاره شده و دو مثال از موارد زیر نیز آورده شده است:

- ادغام گزارش درس با فعالیت‌های کاربر
- امتیازات تجربه، سکه‌ها و داده‌های بازپخش
- تبدیل داده‌های CRM: به طور مثال تبدیل کلمه مرد که ژاپنی نوشته شده به ”male“
- داده‌های پیشرفت هر فصل
- داده‌های بازیکن: به طور مثال اگر تفاوت بین دو مهر زمانی کمتر از ۶۰ دقیقه باشد؛ بدین معنی است که بازیکن همچنان در حال بازی کردن می‌باشد و اگر بیش‌تر باشد، یعنی بازیکن فعالیتی نداشته است.

---

[۳۱] User Id
[۳۲] Timestamp
[۳۳] Contents
[۳۴] Level
[۳۵] Performance Features
[۳۶] Gender
[۳۷] Date of Birth
[۳۸] Resident Area
[۳۹] Demographic

۴,۱ تحلیل اکتشافی داده

حال مجموعه داده برای EDA و پاکسازی داده آماده است. آمار توصیفی همه متغیرهای محاسبه شده در جدول ۵ نشان داده شده است:

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

۵,۱ پاکسازی داده

در روند پاکسازی داده‌ها سه گام مختلف را باید انجام داد. گام اول پر کردن داده‌های گمشده است. گام دوم تصحیح داده‌های غیر ممکن و گام سوم نیز رسیدگی به داده‌های پرت می‌باشد.

در رابطه با اولین گام، باید به جدول ۵ استناد کرد. اگر به این جدول دقت شود، می‌توان فهمید که ۶ سطر از ۷ سطر آخر جدول با بقیه سطرها از لحاظ تعداد تفاوت دارند. در واقع تعداد داده‌های موجود در متغیرهای تجربه (exp)، سکه‌ها (coins)، بازپخش (replay)، جنسیت (gender)، سن (age) و محل سکونت (prefecture) از بقیه متغیرها کمتر است. طبق گفته شرکت ژاپنی، مشاهدات با مقادیر گمشده به صورت دستی بررسی شدند؛ زیرا نباید داده‌هایی با مقادیر گمشده وجود داشته باشد. این مشاهدات با مقادیر گمشده دارای الگوهای یکسانی بودند. در واقع همه آن‌ها بدون ویژگی‌های جمعیت شناسی هستند.

با بررسی های صورت گرفته مشخص شد که تعداد ۱۴۴ تا از مشاهدات، در واقع حساب‌های آزمایشی بودند که به منظور بازاریابی برای مدارس و کمپین‌ها ایجاد شده است. این حساب‌های آزمایشی فاقد اطلاعات معتبر بودند. بنابراین، این ۱۴۴ حساب که فاقد متغیر جنسیت و سن بودند از مجموعه داده حذف شدند.

پس از حذف سطرهای گفته شده، متغیر محل سکونت ۲۴ عدد از سطرهای باقی‌مانده نیز همچنان خالی باقی ماند. برای رفع این مشکل، همه محل‌های سکونت موجود بررسی شدند و "توکیو" بر اساس نتایج آماری رایج‌ترین محل سکونت بوده است. پس مقادیر خالی با "توکیو" جایگزین شد.

پس از رسیدگی به ۳ متغیر فوق، همچنان تعداد ۱۲ مشاهده دیگر وجود دارد که ستون‌های تجربه، سکه‌ها و بازپخش آنان مقادیر خالی می‌باشد. پس از بررسی به عمل آمده، مشخص شد که این متغیرها فقط برای بازیکنانی وجود دارد که تا یک منطقه

خاص در فصل ۱ بازی کرده‌اند؛ بنابراین بازیکنان جدیدتر این داده‌ها را ندارند. از این رو، همه‌ی این مقادیر از دسته رفته با مقدار صفر پر شدند.

گام دوم تصحیح مقادیر غیر ممکن می‌باشد. پس از بررسی همه داده‌ها با کمک نمودار جعبه‌ای مشخص شد که متغیر سن دارای مقادیری از صفر تا ۲۰۱۰ می‌باشد. برای سن کمتر، تحقیقات در مورد سن توانایی خواندن برای تعیین کم‌ترین سن ممکن مورد بررسی قرار گرفت. طبق تحقیقات ژاپنی، ۶۰ درصد از کودکان چهار ساله می‌توانند ۸۰ درصد کلمات هیراگانا را بخوانند، حتی اگر سطح یادگیری بسته به خانواده و یا محیط مهد کودک متفاوت باشد. این بدان معناست که سنین زیر ۴ سال باید توانایی خواندن کم‌تری داشته باشند و اجرای این سرویس را دشوارتر می‌کند. بنابراین، سنین زیر ۴ سال با میانه یعنی ۲۶ جایگزین شد.

از سوی دیگر، سنین بالای ۱۰۰ سال نیز وجود داشتند (یعنی داده‌های ۲۰۱۰، ۲۰۰۹، ۲۰۰۴، ۲۰۰۲، ۹۷۲، ۸۲۵ و ۱۱۹). مسن‌ترین سن در ژانویه ۲۰۲۰ (که آخرین تاریخ موجود برای ثبت داده‌ها بود) در ژاپن ۱۱۷ سال بوده است. بنابراین سنین بالای ۱۱۷ سال نیز با میانه ۲۶ جایگزین شدند.

در نهایت، موارد پرت باید رسیدگی شود. مقادیر پرت یافت شده نباید حذف یا جایگزین شوند، زیرا این مقادیر واقعی هستند. با این وجود، نقاط پرت تأثیر زیادی بر مدل‌سازی دارند، بنابراین استانداردسازی باید تأثیر را کاهش دهد. استانداردسازی متغیرها با مقادیر پرت انجام شد.

۶,۱ ویژگی‌ها و هدف

پس از انجام عمل پاکسازی داده‌ها، باید بردارهای ویژگی و ستون هدف معرفی شوند. همان طور که در جدول ۶ قابل مشاهده است؛ بردارهای ویژگی به ۳ دسته‌ی تعهد[۴۰]؛ کارایی[۴۱] و جمعیت شناختی تقسیم شده اند. این مجموعه داده به عنوان مجموعه داده ویژگی‌ها شناخته می‌شود. متغیر هدف نیز وضعیت ریزش می‌باشد. این وضعیت ریزش در این مرحله خالی می‌باشد و در فصل آینده که مفهوم ریزش برای DGBL تعریف شد، این متغیر نیز پر خواهد شد. در ادامه توضیحات هر کدام از متغیر ها آورده شده است:

- **مجموع ورود (total_login)**: تعداد کل ورودهای کاربر.
- **کل دوره (entire_period)**: دوران تعهد (دوران اشتغال). فاصله بین اولین ورود و آخرین ورود به بازی.
- **میانگین انتظار فصل (avr_ch_wait)**: میانگین مدت زمان بین باز کردن بازی و شروع شدن آن. مجموع مدت زمان انتظار تقسیم بر تعداد فصل‌های بازی شده.
- **بازپخش (replay)**: تعداد کل بازپخش‌های هر کاربر.
- **مجموع مدت زمان بازی (total_playtime)**: کل زمان بازی کاربر در دقیقه.
- **مجموع مدت عدم فعالیت (total_inactive)**: کل زمان غیر فعال بودن بین هر ورود به بازی.
- **متوسط مدت زمان بازی (average_playtime)**: میانگین زمان بازی کردن کاربر در هر ورود. حاصل تقسیم مجموع مدت زمان بازی بر کل تعداد دفعات ورود کاربر.

[۴۰] Engagement
[۴۱] Performance

- **متوسط مدت عدم فعالیت (average_inactive)**: میانگین زمان غیر فعال بودن کاربر بین هر ورود. حاصل تقسیم مجموع مدت عدم فعالیت بر کل تعداد دفعات ورود کاربر.
- **مدت زمان بازی فصل ۱ (ch1_playtime)**: مدت زمان بازی فصل ۱
- **مدت زمان بازی فصل ۲ (ch2_playtime)**: مدت زمان بازی فصل ۲
- **مدت زمان بازی فصل ۳ (ch3_playtime)**: مدت زمان بازی فصل ۳
- **مدت زمان بازی فصل ۴ (ch4_playtime)**: مدت زمان بازی فصل ۴
- **مدت زمان بازی فصل ۵ (ch5_playtime)**: مدت زمان بازی فصل ۵
- **مدت زمان بازی فصل ۶ (ch6_playtime)**: مدت زمان بازی فصل ۶
- **مدت زمان بازی فصل ۷ (ch7_playtime)**: مدت زمان بازی فصل ۷
- **تجربه (exp)**: مجموع امتیازهای تجربه‌ی هر کاربر
- **سکه‌ها (coins)**: مجموع سکه‌های هر کاربر
- **جنسیت (gender)**: جنسیت هر کاربر به صورت دودویی (صفر برای زن و یک برای مرد)
- **سن (age)**: سن هر کاربر
- **محل سکونت (prefecture)**: استان محل سکونت هر کاربر از ۰ تا ۴۷
- **وضعیت ریزش (churn_status)**: متغیر هدف. وضعیت ریزش هر کاربر به صورت دودویی (صفر = غلط و یک = درست)

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)

۷,۱ انتخاب ویژگی

پس از معرفی ویژگی‌های موجود که ۱۹ بردار می‌باشند، حال باید عمل تحلیل مولفه‌های اساسی را بر روی آن پیاده کرد. به این علت که در مطالعات گذشته از این عمل استفاده شده است. با کمک PCA، ۱۹ مؤلفه اساسی تولید می‌شود که از "PC-1" تا "PC-19" نامگذاری شدند. واریانس توضیح داده شده این مؤلفه‌های اساسی در جدول A.1 که در پیوست می‌باشد، توضیح داده شده است. این مؤلفه‌های اساسی تولید شده در یک مجموعه داده جدید به نام features_pc ذخیره شدند. دو مجموعه داده موجود یعنی features و features_pc برای مدل سازی و ارزیابی مورد استفاده قرار می‌گیرند.

۲ تعریف ریزش

با توجه به نبود یک مفهوم ریزش برای حوزه DGBL، باید از حوزه‌های مشابه یعنی آموزش و بازی استفاده کرد. در این فصل مفهوم ریزش در حوزه‌های مذکور مورد بررسی قرار گرفته تا در فصل آینده از این مفاهیم استفاده کرده و یک نرخ ریزش برای حوزه DGBL تعریف کنیم.

زمانی که صحبت از ریزش به میان باشد، یک مثال معروف در جوامع و سایت‌های مختلف وجود دارد که مکان صنعتی مورد نظر را به یک سطل تشبیه می‌کند که آب در آن ریخته می‌شود. این سطل دارای سوراخ می‌باشد و به همین علت مقداری آب از سطل خارج خواهد شد. مقادیر زیادی آب که درون سطل باقی مانده است را به افرادی که حفظ شده اند تشبیه می‌کنند و آب‌های از دست رفته را به افراد ریزش‌کننده تشبیه می‌کنند. علت قرار گرفتن عکس در فایل ارائه‌ی پاورپوینت نیز همین است.

۱,۲ آموزش

بسیاری از دوره‌های آنلاین دانشگاهی هفته‌های ثابتی دارند و در فواصل زمانی مشخصی، تکالیف یا امتحاناتی را ارائه می‌کنند. در صنعت آموزش، معمولاً اگر درسی ناتمام باقی بماند و یا فرد در امتحان پایانی آن درس رد شود، به معنای این است که کاربر در آن درس قبول نشده است. پس می‌توان گفت که در صنعت آموزش، ریزش‌کننده‌ها کسانی هستند که کار را در هفته یا پایان دوره تکمیل نکرده باشند.

نکته قابل توجه این است که در این حوزه آموزش، با مسئله‌ای به نام زمان باید مواجه شد؛ زیرا محصول ارائه شده در قالب بازی است و مدت زمان بازی کردن و شروع بازی در اختیار خود کاربر می‌باشد. پس نمی‌توان کاربر را محدود به یک بازه ۱۶ هفته‌ای برای اتمام بازی کرد و بازیکن آزاد است که هر طور میل دارد بازی را انجام دهد تا یادگیری اتفاق بیفتد. در نتیجه، از این مفهوم ریزش در صنعت آموزش نمی‌توان استفاده کرد. برای پرداختن به این مشکل می‌توان به حوزه‌های دیگر رجوع کرد که کاربرد بیش‌تری می‌تواند داشته باشد.

۲,۲ بازی

در صنعت بازی از یک مفهوم به نام «مدت زمان عدم فعالیت کاربر پس از آخرین ورود» استفاده می‌شود. می‌توان از این دوره عدم فعالیت به عنوان یک برش استفاده کرد و نرخ ریزش را مشخص کرد. به این صورت که یک C باید پیدا کرد که نشان‌دهنده

مدت زمان فعال نبودن بازیکن است و هر بازیکنی که بیش‌تر از این C فعال نباشد، به عنوان ریزش‌کننده شناسایی خواهد شد. ولی اگر رویکرد فوق با مجموعه داده ارائه شده به خوبی کار نکند، باید به دنبال یک راه جایگزین بود. یا به عبارت دیگر باید یک چارچوب زمانی دلخواه برای هر کاربر و شناسایی وضعیت ریزش آن محاسبه شود.

۳,۲ قمار

در صنعت قمار مفهومی وجود دارد که در آینده وارد صنعت بازی‌های آنلاین نیز شده است. این مفهوم تازگی[۴۲]نامیده می‌شود. تازگی بدین معناست که کاربر برای آخرین بار، کِی از محصول یا خدمت مورد نظر استفاده کرده است؟ این مفهوم هم‌اکنون در بخش‌های مختلفی از جمله در تحلیل سبد خرید، بورس، سری‌های زمانی و ... استفاده می‌شود. در صورتی که نتوان از مفاهیم موجود در صنایع قبلی استفاده کرد، می‌توان از این مفهوم بهره گرفت.

۳ تعیین ریزش

در فصل گذشته مفهوم ریزش در صنایع مختلف بررسی شد. در این فصل هدف این است که با توجه به مفهوم موجود که در صنایع مختلف نیز متفاوت است، نرخ ریزش را برای حوزه DGBL محاسبه و تعیین کرد.

مفهوم ریزش در صنعت آموزش با توجه به ماهیتی که دارد اصلاً برای حوزه DGBL مناسب نیست. یعنی از مفهوم ریزش در آموزش نمی‌توان در DGBL هم استفاده کرد. پس در این فصل فقط به بررسی دو صنعت بازی و قمار (یا حتی می‌توان گفت بازی‌های آنلاین) پرداخته می‌شود.

۱,۳ بازی

طرح اولیه در این حوزه، بررسی نمودار پراکندگی کل زمان پخش و میانگین زمان عدم فعالیت است. اگر این طرح نقطه برش خاصی برای نرخ ریزش را نشان ندهد، باید از یک طرح دیگر استفاده کرد. همان‌طور که در شکل ۱ می‌توان مشاهده کرد، محور عمودی به مجموع زمان پخش (به دقیقه) و محور افقی به میانگین مدت زمان عدم فعالیت (به روز) اختصاص داده شده است. با توجه به این نمودار می‌توان گفت که پراکندگی داده‌ها از روز ۱۰۰ به بعد کم‌تر خواهد شد. یا یک نقطه آشکارتر را می‌توان ذکر کرد که بین ۱۵۰ تا ۱۸۰ می‌باشد. با این حال، با توجه به حجم ریزش‌کننده‌های موجود در حوزه فناوری آموزشی، این مقدار قاطعی نمی‌تواند باشد. به طور مثال حتی اگر نقطه برش روز ۱۰۰ تعیین شود، تعداد بازیکنان ریزش‌کننده فقط ۲۸ نفر است که تنها ۰/۷۵٪ از جمعیت را پوشش می‌دهد. علاوه بر این، هیچ دلیل محکمی برای تعیین محدودیت در روزهای دیگر وجود ندارد.

پس می‌توان گفت که مسئله اصلی در این طرح عدم انتخاب یک نقطه برش مناسب برای افراد ریزش‌کننده می‌باشد. در نتیجه باید به دنبال یک طرح دیگر بود که این مسئله را رفع کند. طرح دیگر اینگونه است که باید با کمک یک محاسبه از میانگین عدم فعالیت (به ساعت) و مجموع مدت زمان بازی (به دقیقه) نمودار را تغییر داد. در واقع یک فرمول باید ساخته شود که شکل نمودار را به‌گونه‌ای عوض کند که پراکندگی میان داده‌ها زیاد شود.

---

[۴۲] Recency

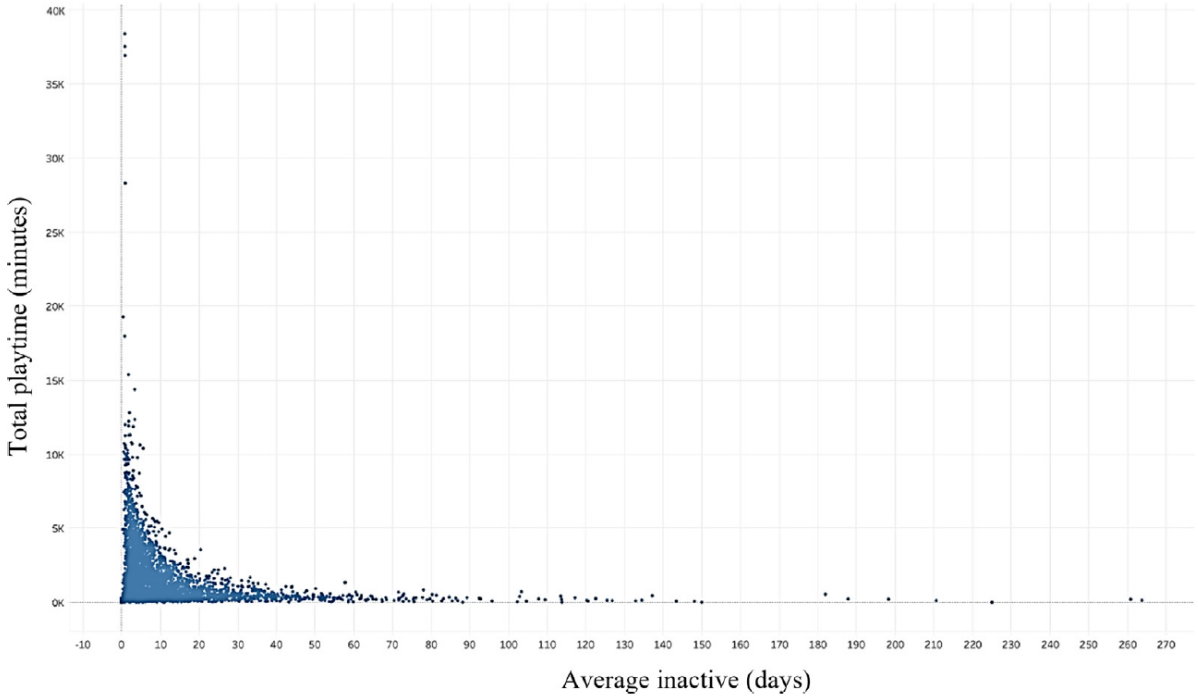


Fig. 1. Scatter plot 1 with average inactive (days) and total playtime.

محاسبه لازم برای انجام این تغییرات در نمودار، بدین صورت است که باید میانگین عدم فعالیت کاربر در روز به دست آید. فرمول لازم برای انجام این کار به صورت زیر می‌باشد:

$$average_{inactive}(Hours) = \left(۶۰ \times ۲۴ - \frac{total_{playtime}}{total_{login}}\right) \div ۶۰$$

به طور مثال فرض کنید که مجموع مدت زمان بازی یک کاربر برابر ۶۰۰ دقیقه باشد و این فرد به تعداد ۲۰ بار وارد بازی شده باشد. در نتیجه میانگین عدم فعالیت این کاربر در روز برابر خواهد بود با:

$$average_{inactive}(Hours) = \left(۱۴۴۰ - \frac{۶۰۰}{۲۰}\right) \div ۶۰ \Longrightarrow$$

$$average_{inactive}(Hours) = (۱۴۴۰ - ۳۰) \div ۶۰ = ۲۳/۵$$

$$average_{inactive}(Hours) = ۱۴۱۰ \div ۶۰ = ۲۳/۵$$

با کمک فرمول فوق که بر روی همه داده‌ها اعمال می‌شود، نمودار پراکندگی تغییر خواهد کرد. شکل ۲ نشان دهنده‌ی نمودار پراکندگی جدید می‌باشد که در زیر قابل مشاهده است:

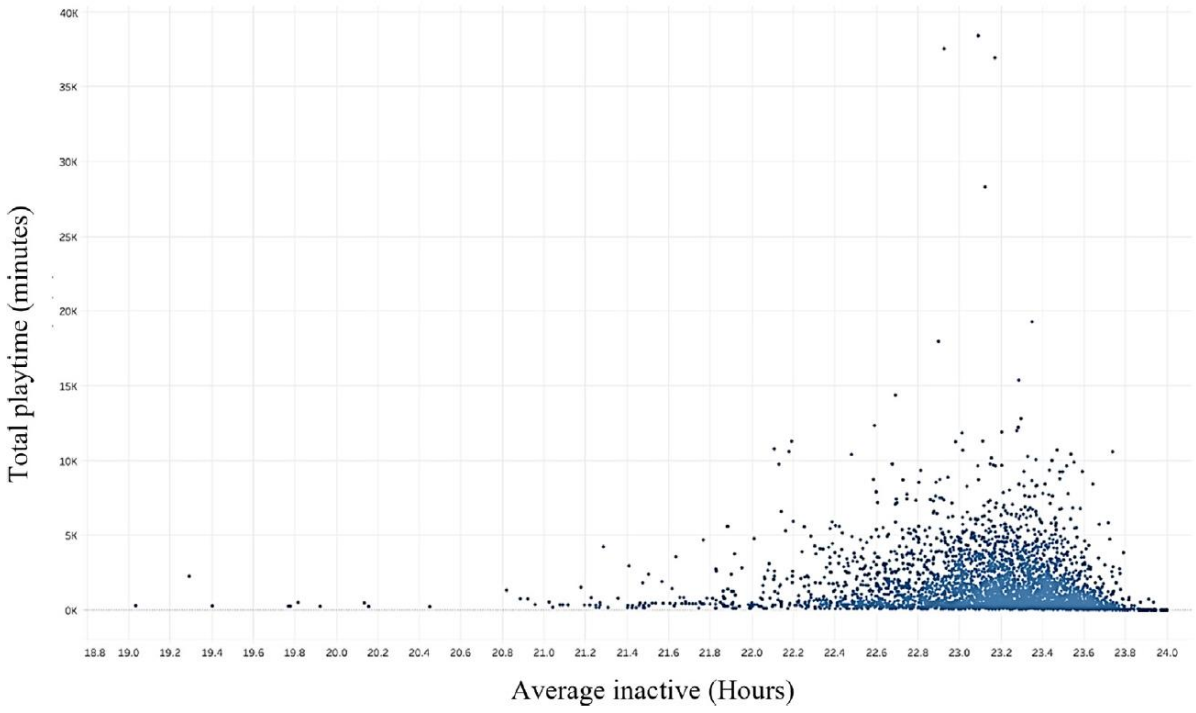


Fig. 2. Scatter plot 2 with average inactive in a day (hours) and total playtime.

به طور واضح می‌توان گفت که نمودار تغییر پیدا کرده است و مقداری پراکندگی میان داده‌ها بیش‌تر شده است ولی همچنان مسئله اصلی یعنی انتخاب یک مقدار به عنوان نقطه برش برای مشخص کردن افراد ریزش‌کننده بر جای خود باقی است. این نمودار پس از ۲۰ ساعت عدم فعالیت تراکم بسیار بالایی بین داده‌ها را نشان می‌دهد. دلیل بالقوه این موضوع این است که DGBL یک دوره غیرفعال طولانی دارد و نسبت به داده‌های صنعت بازی کم‌تر قابل تشخیص است (یعنی داده‌ها تراکم بیش‌تری نسبت به داده‌های متداول در صنعت بازی دارند). پس می‌توان گفت که این رویکرد برای DGBL مناسب نیست.

۲,۳ تعیین نهایی نرخ ریزش

رویکرد نهایی که می‌تواند جوابگوی مسئله ما باشد به ۳ مقدار مختلف نیاز دارد: تازگی، میانگین عدم فعالیت و انحراف معیار عدم فعالیت کاربران. با کمک این ۳ مقدار می‌توان یک نقطه برش مناسب به دست آورد. دو مقدار میانگین و انحراف معیار کاربران که به سادگی به دست می‌آید. مقدار تازگی نیز با کمک فرمول زیر به دست خواهد آمد:

$$Recency = Jan\ ۲۸^{th}\ ۲۰۲۰ - Last\ Login$$

علت استفاده از تاریخ ۲۸ام ژانویه ۲۰۲۰ این است که داده‌ها در تاریخ ۲۷ام ژانویه ۲۰۲۰ به دست پژوهشگر رسیده؛ در نتیجه برای محاسبه معیار تازگی باید تاریخ آخرین ورود هر کاربر را از ۲۸ام ژانویه کسر کرد. حال با کمک میانگین و انحراف معیار موجود باید نقطه برش را تعیین کرد. این نقطه برش به صورت زیر محاسبه می‌شود:

$$Cutoff = avr + ۲std$$

پس از محاسبه معیار تازگی و همچنین نقطه برش مورد نظر می‌توان گفت که اگر تازگی بیش‌تر از این نقطه برش بود، فرد را باید ریزش‌کننده دانست (یعنی وضعیت ریزش این فرد را باید مساوی با ۱ قرار داد)؛ در غیر این‌صورت وضعیت ریزش فرد مساوی با صفر قرار خواهد گرفت. درصد ریزش‌کننده‌ها و افرادی که ریزش نکرده اند در شکل ۳ به نمایش درآمده است. از میان ۳۵۵۷ مشاهده‌ی موجود ۵۶/۷۷ ٪ آن‌ها ریزش‌کننده و ۴۳/۲۳ ٪ آن‌ها افرادی هستند که ریزش نکرده‌اند. این نسبت موجود بین افراد ریزش‌کننده و غیر ریزش‌کننده‌ها با در نظر گرفتن نرخ ریزش بالای EdTech همان‌طور که قبلاً توضیح داده شد، قابل قبول است. از این رو می‌توان گفت که نرخ ریزش مورد نظر برای داده‌ها به دست آمده است.

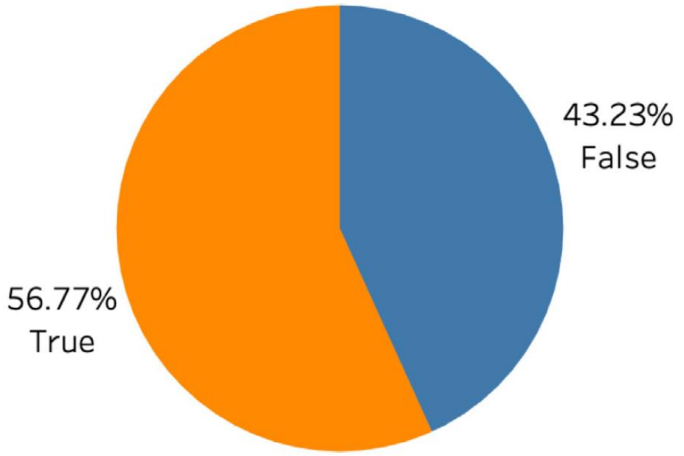


Fig. 3. The proportion of churners (True) and non-churners (False).

۴ تحلیل بقا

پس از انجام همه موارد لازم برای آماده‌سازی داده‌ها باید تجزیه و تحلیل‌های لازم را برای رسیدن به اطلاعات لازم و تحلیل آن‌ها انجام داد.

۱,۴ تجزیه و تحلیل توصیفی از داده‌های کاربر

برای درک وضعیت فعلی محصول، کاربران فعال ماهانه در شکل ۴ نمایش داده شده است. تعداد کاربران فعال ماهانه عموماً در طول سال اول در حال افزایش بود و سپس در حدود ۱۰۰۰ نفر، این تعداد رو به کاهش رفته است. با توجه به کاهش تعداد کاربران می‌توان فهمید که این محصول در دراز مدت نمی‌تواند کاربران خود را حفظ کند. علت این امر را با توجه به جداول، نمودارها و تحلیل‌های بعدی می‌توان متوجه شد.

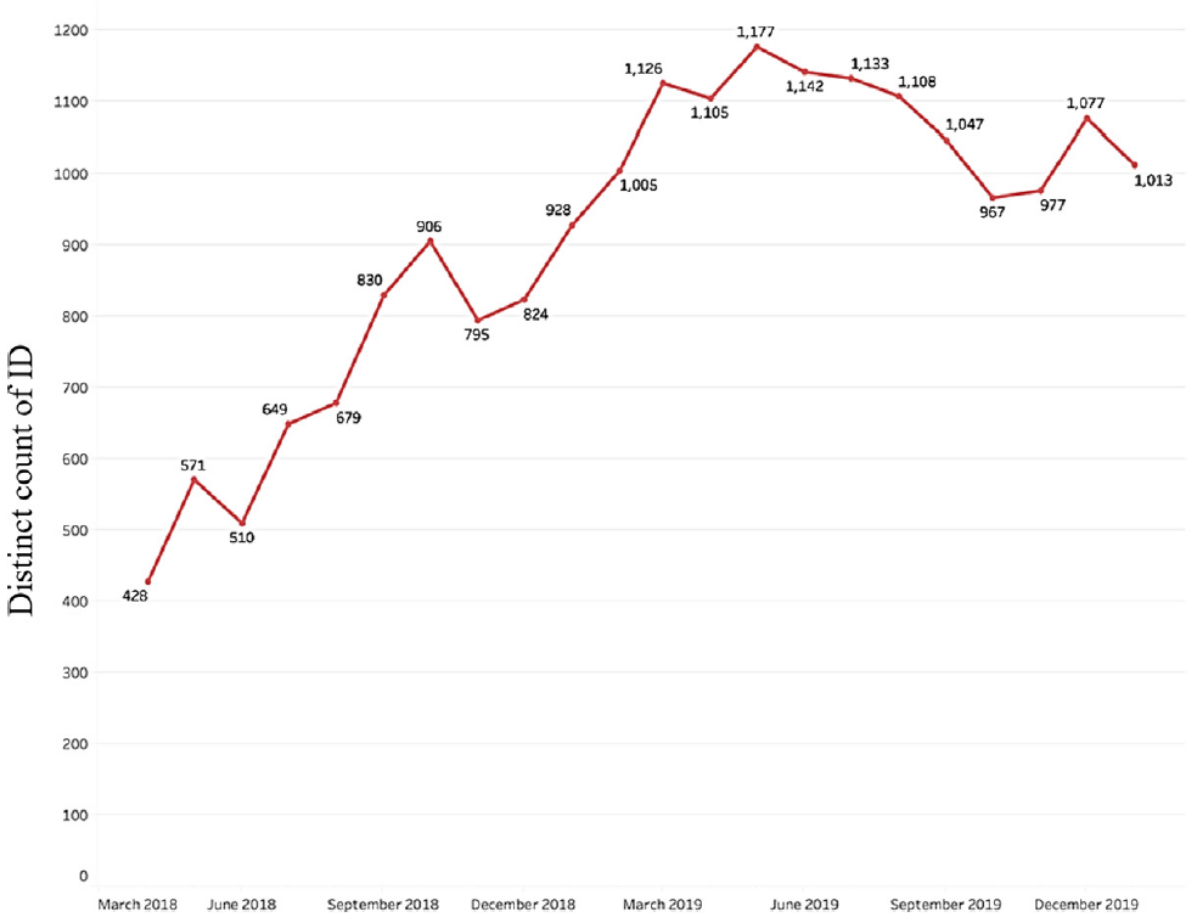
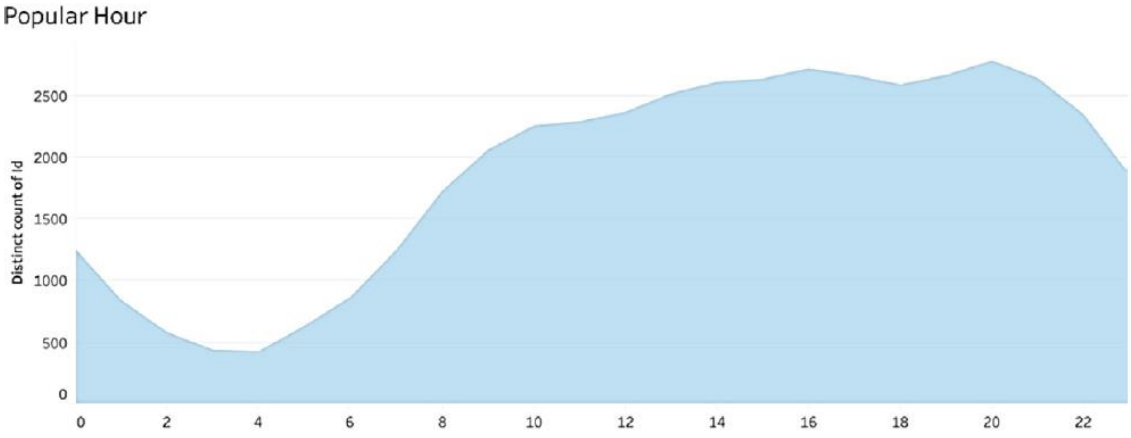


Fig. 4. Monthly active users (MAU).

در ادامه ساعت و روز محبوب هفته در شکل زیر نمایش داده شده است. با توجه به ساعت و روز پرطرفدار هفته، ساعت پرطرفدار مورد انتظار بعد از ساعات اداری و تا نیمه شب است. پرطرفدارترین ساعات، از ساعت ۱۳ تا ۲۱ می‌باشد. با توجه به ساعت تعطیلی مدارس و ادارات می‌توان گفت که تعداد دانش‌آموزان بیش‌تر از بزرگسالانی است که مشغول به کار هستند. پس از ساعت ۲۰ نیز می‌توان یک افت در بین افراد مشاهده کرد که نشان‌دهنده ساعت خواب آنان می‌باشد.



همچنین در ادامه می‌توان گفت که با توجه به شکل ۵ که محبوب‌ترین روز در هفته را نشان می‌دهد، می‌توان گفت که تفاوت فاحشی در بین روزهای هفته وجود ندارد و در همه روز به طور تقریباً یکسانی افراد در بازی حضور داشتند. ولی در روزهای آخر هفته (یعنی شنبه و یکشنبه) به طور معمول افراد بیشتری در بازی حضور داشتند.

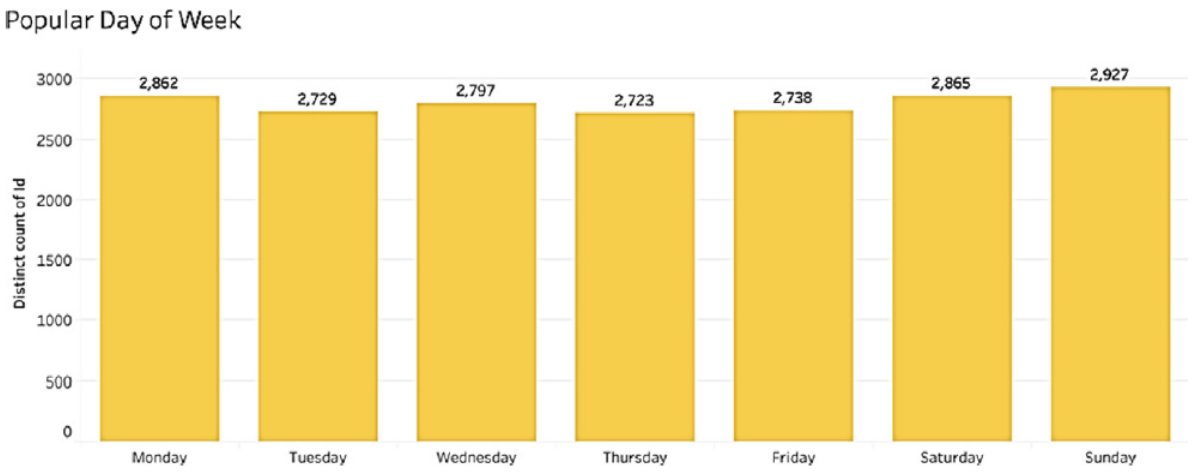


Fig. 5. Popular hour and day of the week.

۲,۴ تحلیل بقا بر اساس فصل

پس از درک کلی رفتار کاربر، حال باید این تجزیه و تحلیل را بر روی فصل‌های مختلف با کمک تعریف ریزش جدید انجام داد. اول از همه، میانه مدت زمان بازی هر فصل در شکل ۶ نمایش داده شده است. علت استفاده از میانه، به جای میانگین طبق گفته پژوهشگران از بین بردن تاثیر داده‌های پرت می‌باشد. همان‌طور که در شکل ۶ قابل مشاهده است، کم‌ترین مدت زمان بازی مربوط به فصل اول می‌باشد و هم‌چنین بیش‌ترین مدت زمان بازی مربوط به فصل دوم است.

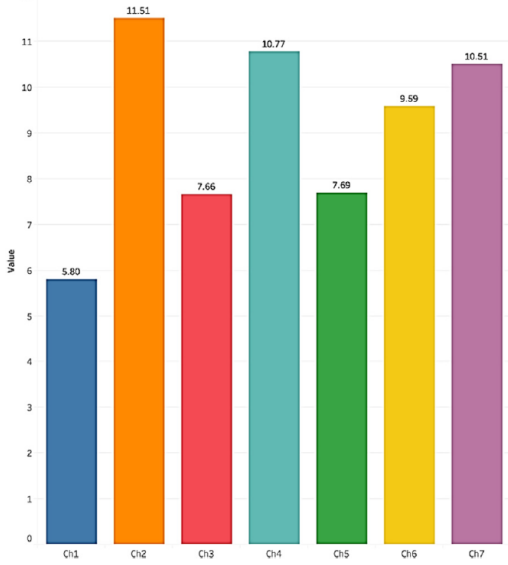


Fig. 6. Median playtime per chapter (hour).

با در نظر گرفتن ۵۶/۷۷ درصد ریزش‌کننده‌ها و ۴۳/۲۳ درصد غیر ریزش‌کننده‌ها، می‌توان گفت که تفاوت قابل توجهی بین ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها وجود دارد. به طور کلی افرادی که ریزش می‌کنند کم‌تر از افرادی که باقی می‌مانند، بازی می‌کنند؛ اما این فاصله زمانی به هر فصل بستگی دارد. جدول ۸ نشان‌دهنده‌ی تفاوت میانگین مدت زمان بازی کردن را برای هر فصل نشان می‌دهد.

در مرحله بعد، برای مشاهده میزان تکمیل هر فصل، تعداد کاربرانی که توانستند هر فصل را به پایان برسانند محاسبه شد. سپس درصد تکمیل فصل جاری از تکمیل فصل قبل محاسبه شد. درصد تکمیل برای افرادی که ریزش نکرده اند معمولاً بالا است، هر چند که ۱۰۰ ٪ نیست. این به این معنی است که درصد مشخصی از کاربرانی هستند که در حال حاضر فصل را بازی می‌کنند. این درصد می‌تواند به عنوان یک معیار برای ریزش‌کننده‌ها استفاده شود. تفاوت در میزان تکمیل هر فصل در جدول ۹ خلاصه شده است.

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

با مقایسه نتایج جداول ۸ و ۹ به پنج نکته جالب می‌توان دست یافت:

۱) **فصل اول:** تفاوت زیادی بین ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها وجود ندارد. بنابراین ممکن است ریزش‌کننده‌ها تمام محتوای فصل ۱ را بازی کرده باشند و دیگر به آن برنگردند.

۲) **فصل دوم:** بیش‌ترین زمان پخش برای ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها برای فصل دوم می‌باشد. با این حال تفاوت زمان پخش بین آن‌ها تقریباً ۲ ساعت است. همچنین نرخ تکمیل برای ریزش‌کننده‌ها از فصل ۱ که برابر با ۷۵/۲۲ ٪ بود به میزان ۵۰/۳۵ ٪ در فصل دوم رسیده است. **پس** می‌توان گفت که بسیاری از کاربران در فصل دوم ریزش داشته اند. با کوتاه کردن و یا ساده‌تر کردن محتوا می‌توان این مشکل را تا حد زیادی حل نمود.

۳) **فصل سوم:** فقط یک ساعت تفاوت بین ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها وجود دارد؛ ولی فاصله نرخ تکمیل بسیار زیاد بوده و برابر ۱۷/۶۱ درصد است. **پس** می‌توان گفت که ریزش‌کننده‌ها احتمالاً در اواخر این فصل بازی و آموزش را ترک کردند. بنابراین محتواهای آخر این فصل باید بازنگری شود.

۴) **فصل چهارم:** این فصل نیز به طور مشابه با فصل دوم دارای مدت زمان بازی زیادی است. همچنین نرخ تکمیل این فصل به نسبت فصل گذشته افت داشته است. **پس** می‌توان گفت که همانند فصل دوم ریزش وجود داشته و محتواها باید کوتاه و یا ساده‌تر شوند.

۵) **فصل پنجم، ششم و هفتم:** تفاوت بین مدت زمان بازی ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها در فصل‌های ۶ و ۷ بیش‌تر از همه فصول دیگر است. همچنین تفاوت بین نرخ تکمیل ریزش‌کننده‌ها و غیر ریزش‌کننده‌ها در فصل‌های ۵ و ۶ بیش‌تر از همه فصل‌های دیگر می‌باشد. علاوه بر این، میزان نرخ تکمیل فصل هفتم برای افراد غیر ریزش‌کننده برابر صفر است! علت این امر ناشناخته است و حدس زده می‌شود که علت آن کمبود داده‌های لازم باشد. بنابراین این مسئله نیز با افزودن داده حل خواهد شد.

۵ بهینه‌سازی هایپرپارامتر

برای انتخاب بهترین مدل ابتدا باید حالت مطلوب را پیدا کرد. برا این کار داده‌ها را به دو مجموعه داده آموزشی[۴۳] و آزمایشی[۴۴] در سه الگوی متفاوت تقسیم شدند. الگوی اول ۹۰٪ برای تمرین و ۱۰٪ برای مجموعه تست است. مورد دوم ۸۵٪ برای تمرین و ۱۵٪ برای تست و آخرین مورد نیز ۸۰٪ برای تمرین و ۲۰٪ برای تست است. هم‌چنین روش k-fold cross-validation با مقدار $k=10$ برای انجام این تقسیم‌ها مورد نظر قرار گرفته است.

در مورد الگوریتم‌ها سه الگوریتم رایج برای پیش‌بینی ریزش بر اساس آموزش و صنعت بازی انتخاب شدند. زیرا DGBL ترکیبی از این صنایع است و هیچ تحقیقی در مورد پیش‌بینی ریزش DGBL وجود ندارد. الگوریتم‌های رایج این دو صنعت برای اعمال مجموعه داده‌های موجود انتخاب شدند. این الگوریتم‌ها عبارتند از درخت تصمیم[۴۵]، جنگل تصادفی[۴۶] و رگرسیون لجستیک.[۴۷]

درخت تصمیم یکی از قابل تفسیرترین مدل‌ها توسط انسان است. درخت تصمیم یک الگوریتم طبقه‌بندی است که برای مسائل یادگیری بانظارت استفاده می‌شود. درخت نشان می‌دهد که در هر مرحله از کدام ویژگی استفاده شده و با چه مقداری این تفکیک صورت می‌گیرد. بنابراین تفسیر ساده‌تر است.

جنگل تصادفی یکی دیگر از الگوریتم‌های طبقه بندی مبتنی بر درخت است که از درخت‌های تصمیم‌گیری زیادی تشکیل شده است. در واقع جنگل تصادفی، چندین درخت تصمیم را با نمونه‌گیری تصادفی می‌سازد و هر کدام از آن درخت‌ها عمل طبقه‌بندی را انجام داده و نتیجه را برمی‌گردانند. سپس جنگل تصادفی نتایج را از درختان جمع‌آوری کرده و بهترین نتیجه را به عنوان نتیجه نهایی با رأی گیری انتخاب می‌کند.

رگرسیون لجستیک یک الگوریتم احتمالی برای طبقه‌بندی باینتری است. خروجی واضح از نظر ریاضی یکی از دلایل محبوبیت آن است؛ در حالی که در تفسیر توسط انسان ضعف دارد.

---

[۴۳] Train
[۴۴] Test
[۴۵] Decision Tree (DT)
[۴۶] Random Forest (RF)
[۴۷] Logistic Regression (LR)

به عنوان معیار ارزیابی، معیار سطح زیر منحنی[۴۸]انتخاب شده است. هر چه مساحت این ناحیه زیر منحنی به یک نزدیک‌تر باشد یعنی مساحت بزرگتری دارد و این نشان‌دهنده طبقه‌بندی بهتر است. بسیاری از مطالعات از AUC به عنوان معیاری برای مقایسه مدل‌های پیش‌بینی در صنعت بازی و آموزش استفاده کردند.

نتایج AUC برای یافتن بهترین مجموعه داده و درصد تقسیم برای مدل‌های مختلف در جدول ۷ خلاصه شده است. از این جدول می‌توان دریافت که با استفاده از features_pc عملکرد بهتری در همه مدل‌ها به نسبت مجموعه داده معمولی دارد. علاوه بر آن، بهترین تقسیم برای دو تا از مدل‌ها با مجموعه داده PCA مربوط به تقسیم ۹۰-۱۰ می‌باشد.

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.

۱,۵ درخت تصمیم

الگوریتم درخت تصمیم دارای ۵ هایپرپارامتر به نام‌های معیار[۴۹]، تقسیم‌کننده[۵۰]، حداکثر عمق[۵۱]، حداقل نمونه‌های تقسیم[۵۲]و حداقل نمونه‌های برگ[۵۳]می‌باشد. هایپرپارامترها همراه با توضیحات و مقادیر قابل دریافت آن‌ها در جدول ۱۰ گردآوری شده است؛ هر چند که هر کدام از هایپرپارامترهای مذکور در ادامه به تفکیک توضیح داده شده اند و پس از بررسی آن‌ها بهترین هایپرپارامتر انتخاب شده است.

[۴۸] Area Under Curve (AUC)
[۴۹] Criterion
[۵۰] Splitter
[۵۱] Max Depth
[۵۲] Minimum Samples Split
[۵۳] Minimum Samples Leaf

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.

۱،۱،۵ معیار و تقسیم‌کننده

معیار برای سنجش کیفیت تقسیم‌ها است. دو معیار جینی[۵۴] و آنتروپی[۵۵] برای درخت تصمیم وجود دارد. هم‌چنین تقسیم‌کننده نیز نشان‌دهنده روش‌های مختلف برای انتخاب تقسیم در هر گره درخت تصمیم است. تقسیم‌کننده نیز دارای دو حالت تصادفی و بهترین حالت ممکن است. نتایج استفاده از معیارها و تقسیم‌کننده‌های گفته شده در شکل ۷ که در زیر آمده است، خلاصه‌سازی شده است. همان‌طور که قابل مشاهده است بهترین معیار، معیار آنتروپی بوده و بهترین تقسیم‌کننده نیز بهترین حالت است. پس این دو معیار برای الگوریتم درخت تصمیم انتخاب می‌شوند.

```
Decision Tree ==============================================
Gini Criteria AUC with Validation:  0.8492491456309462
Entropy Criteria AUC with Validation:  0.8533239233455893
Best Splitter AUC with Validation:  0.8492491456309462
Random Splitter AUC with Validation:  0.8183504809576065
```

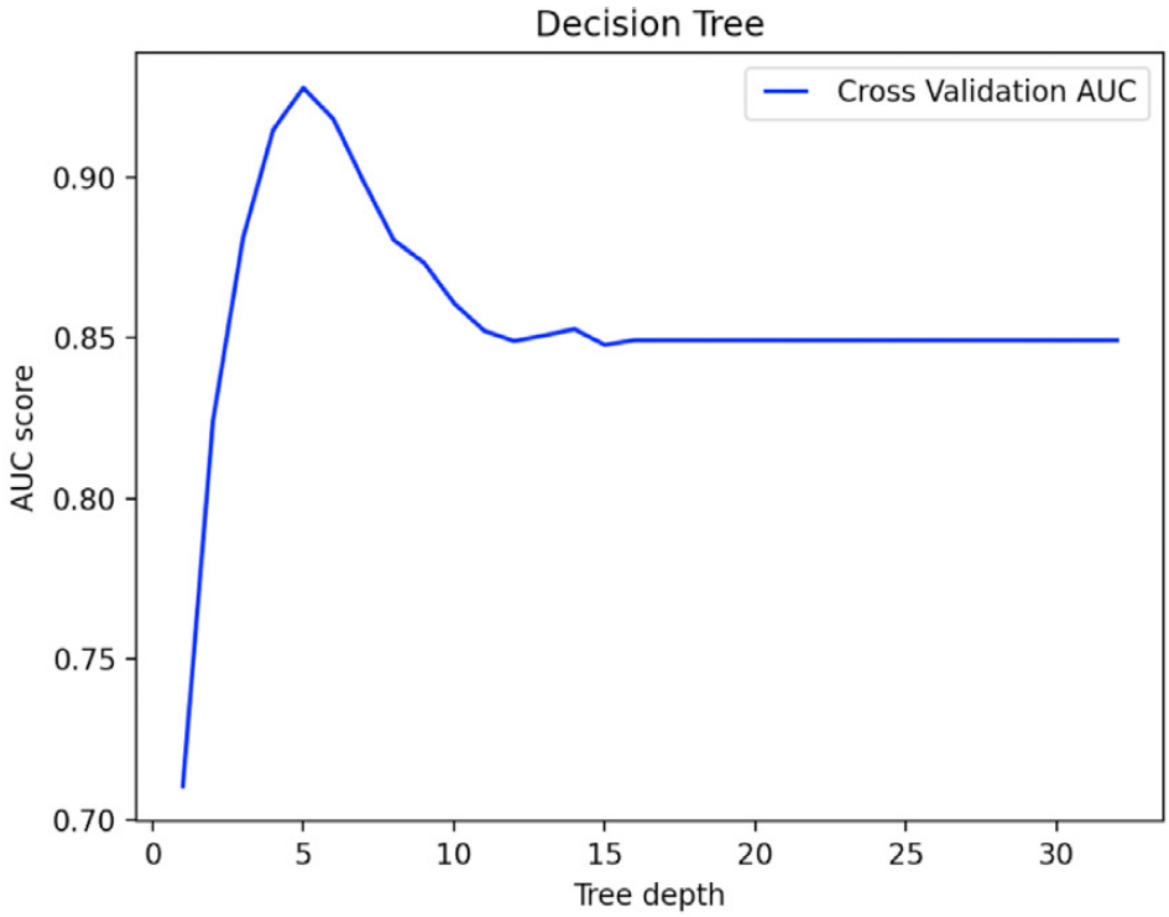
Fig. 7. The result of AUC with different criteria and splitter values.

۲،۱،۵ حداکثر عمق

حداکثر عمق درخت، اعداد صحیح بین ۱ تا ۳۲ در نظر گرفته شده است. نتایج استفاده از حداکثر عمق‌های متفاوت برای الگوریتم مورد نظر در شکل ۸(الف) نشان داده شده است. همان‌طور که قابل مشاهده است، پس از حدود ۶ عمق درخت، عملکرد

[۵۴] Gini [۵۵] Entropy

مدل به وضوح کاهش می‌یابد و پس از ۱۰ عمق درخت، عملکرد تقریباً یکسان است. به این معنی که عمق درخت بیش‌تر باعث عملکرد پایین‌تر می‌شود. هم‌چنین هزینه محاسبات نیز بیش‌تر می‌شود.



(a) With different *max_depth* values.

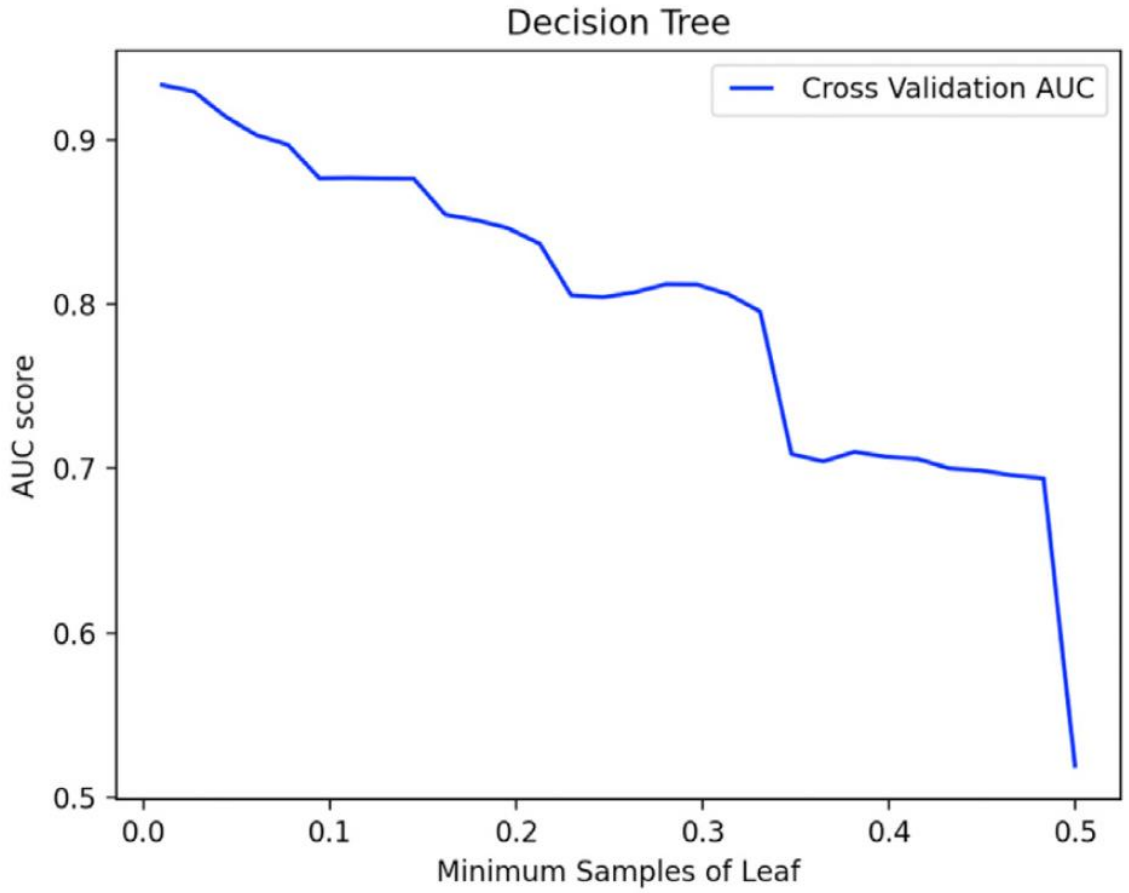
با توجه به نمودار فوق، حداکثر عمق درخت بین ۴ و ۸ تنظیم می‌شود و دوباره برای مشاهده AUC آزمایش صورت می‌گیرد. نتیجه حاصل از آزمایش دوباره در جدول ۱۱ خلاصه‌سازی شده است. بالاترین AUC در حداکثر عمق ۶ است. بنابراین بهترین مقدار برای حداکثر عمق بر روی ۶ تنظیم خواهد شد.

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

۳,۱,۵ حداقل نمونه‌های برگ

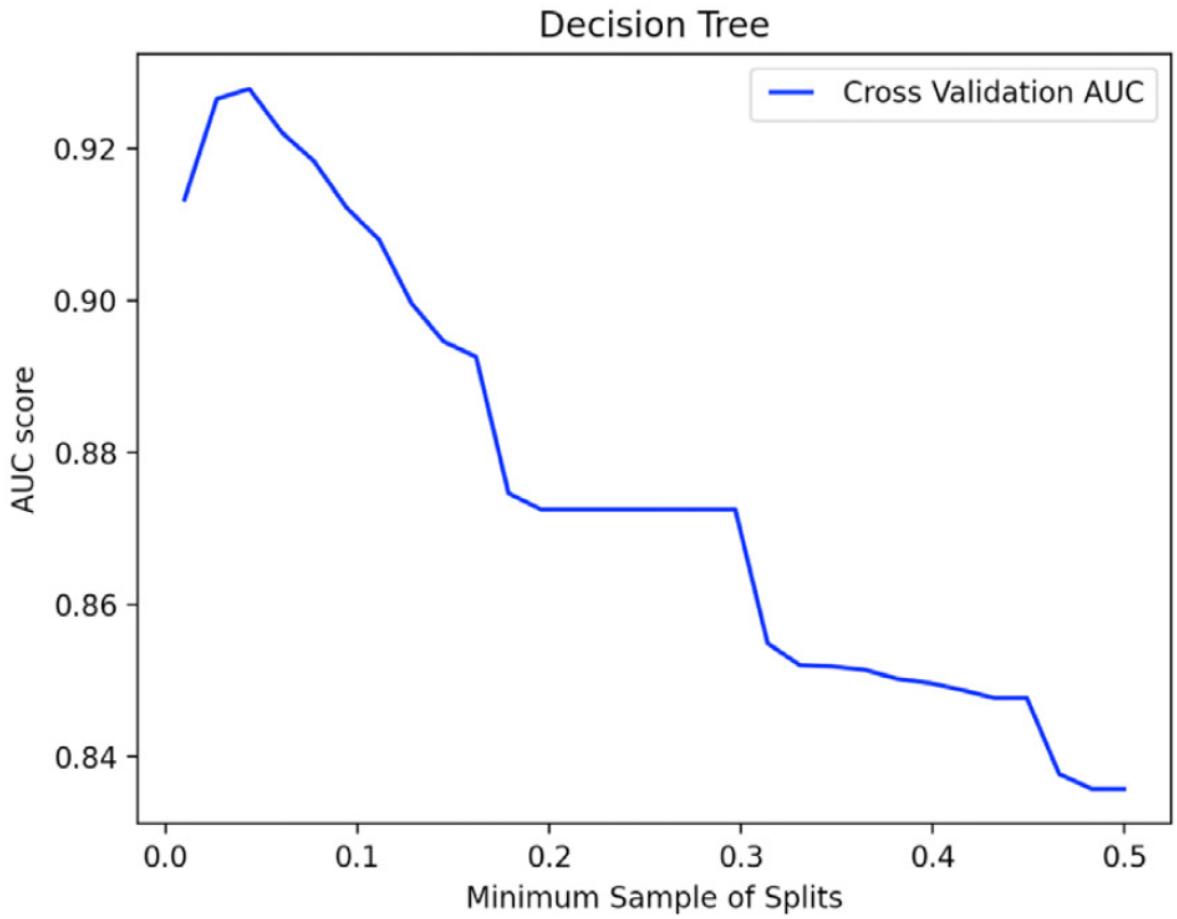
حداقل نمونه‌های برگ در واقع همان حداقل تعداد نمونه مورد نیاز برای تشکیل یک برگ می‌باشد. مقدار آن بین ۰/۰۱ و ۰/۵ تنظیم شده است. نتایج حاصل از این آزمایش در شکل ۸(ج) نشان داده شده است. همان‌طور که می‌توان مشاهده کرد، بهترین عملکرد مربوط به کوچک‌ترین مقدار است.



(c) With different *minimum_samples_leaf* values.

۴,۱,۵ حداقل نمونه‌های تقسیم

حداقل نمونه‌های تقسیم در واقع همان حداقل تعداد نمونه مورد نیاز برای انجام عمل تقسیم می‌باشد. مقدار آن بین ۰/۰۱ و ۰/۵ تنظیم شده است. نتایج حاصل از این آزمایش در شکل ۸(ب) که در زیر وجود دارد، آمده است. همان‌طور که قابل مشاهده است بهترین حالت مدل در حدود بین ۵٪ داده ها قرار گرفته است. ۵٪ داده‌ها برابر است با ۱۴۲ مشاهده. بنابراین این آزمایش دوباره برای حدود ۵٪ داده‌ها انجام خواهد شد.



(b) With different *minimum_samples_split* values.

نتایج حاصل از انجام دوباره آزمایش بهینه‌سازی برای حداقل نمونه‌های تقسیم ۲۵، ۵۰، ۷۵، ۱۰۰، ۱۲۵، ۱۵۰، ۱۷۵ و ۲۰۰ در جدول ۱۲ خلاصه سازی شده است. همان‌طور که قابل مشاهده است، بهترین حداقل نمونه مورد نیاز برای تقسیم برابر ۵۰ داده است.

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

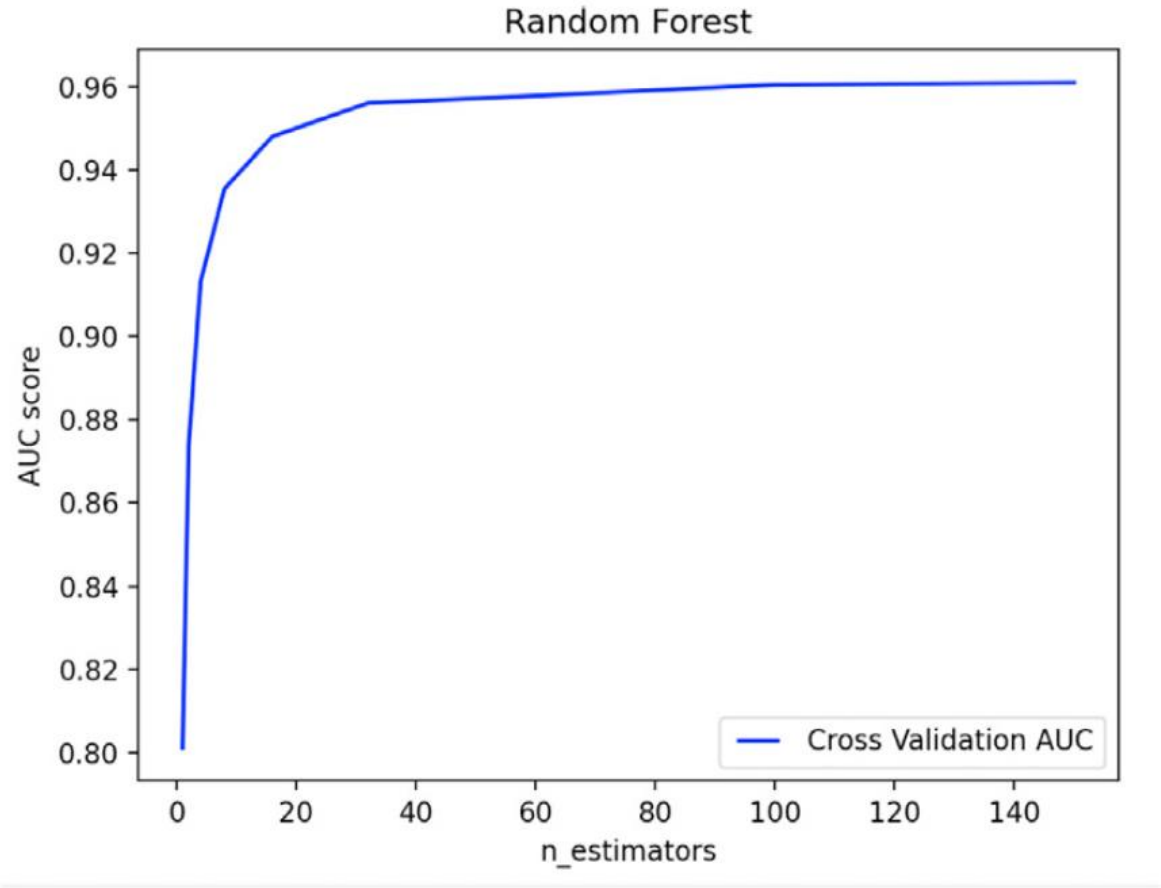
Minimum split	25	50	75	100	125	150	175	200
AUC	0.9317	**0.9333**	0.9320	0.9329	0.9290	0.9284	0.9258	0.9262

۲,۵ جنگل تصادفی

الگوریتم جنگل تصادفی مشابه با درخت تصمیم می‌باشد. یعنی دقیقاً همان هایپرپارامترهای درخت تصمیم، در جنگل تصادفی نیز وجود دارند و فقط یک هایپرپارامتر دیگر به آن اضافه می‌شود که در واقع تعداد شبیه‌سازها[۴]یا همان تعداد درخت‌ها) می‌باشد. با توجه به تشابه میان درخت تصمیم و جنگل تصادفی، دیگر توضیحات مربوطه تکرار نشده و فقط به بررسی هایپرپارامترها پرداخته می‌شود.

۱,۲,۵ تعداد درخت‌ها

هر جنگل تصادفی متشکل از چندین درخت تصمیم است. در این بخش تعداد مناسب درخت‌ها بررسی می‌شود. همان‌طور که در شکل ۹(الف) یعنی شکل زیر قابل مشاهده است، طبقه‌بند خیلی سریع به مقدار ۱ که مقدار مطلوب AUC است، می‌رسد.



(a) With different estimators.

[۴] Estimators

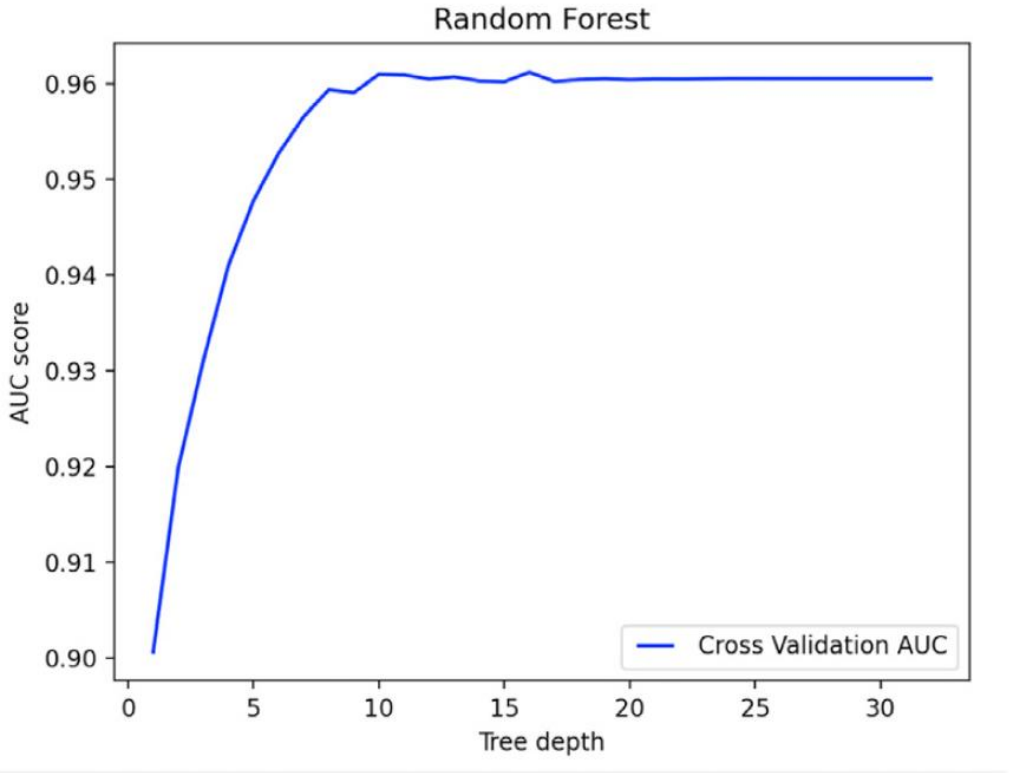
برای مشاهده دقیق مقادیر AUC، آزمایش دوباره برای مقادیر ۵۰، ۷۵، ۱۰۰ و ۱۲۵ تکرار خواهد شد. نتیجه‌ی این آزمایش در جدول ۱۳ نشان داده شده است. بهترین مقدار تعداد درخت برابر ۱۰۰ می‌باشد و پس از آن مطابق نمودار بهبود نمی‌یابد. داشتن درختان بیش‌تر مستلزم هزینه‌های محاسباتی بیش‌تری است.

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

۲,۲,۵ حداکثر عمق

به طور مشابه آزمایش حداکثر عمق بر روی جنگل تصادفی به ازای مقادیر ۱ تا ۳۲ صورت می‌گیرد. نتایج حاصل از این آزمایش در نمودار ۹(ب) نمایش داده شده است.



(b) With different *max_depth* values.

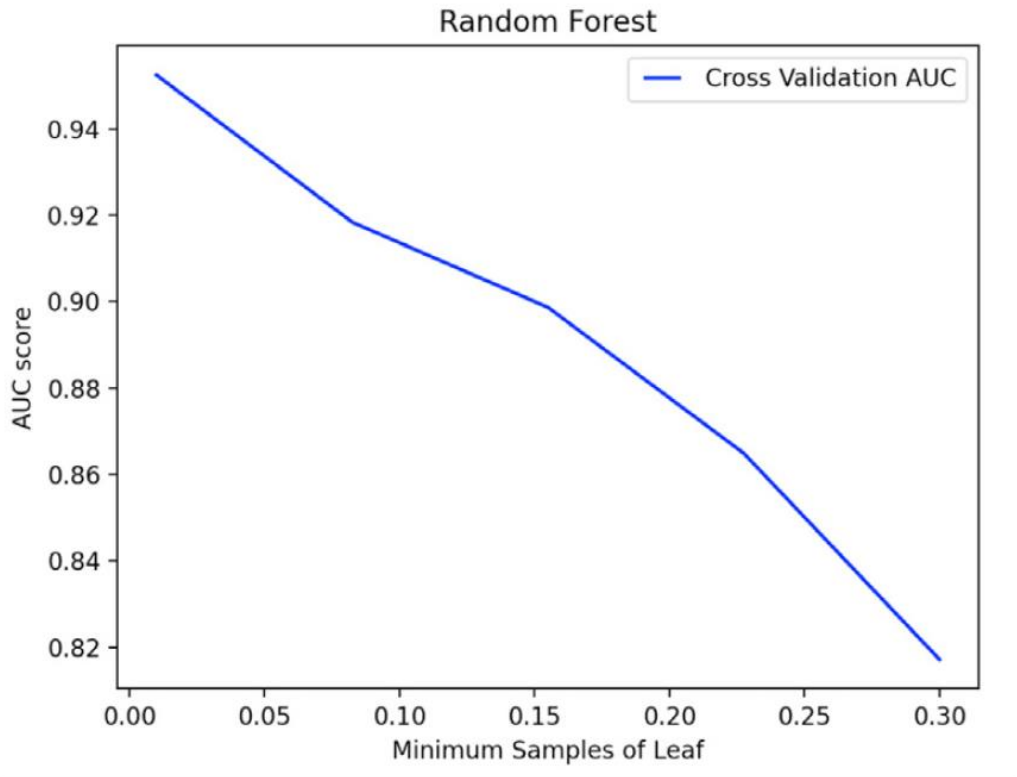
همان‌طور که مشاهده می‌شود به ازای حداکثر عمق ۷ به بعد میزان دقت‌های مطلوبی به دست می‌آید؛ در نتیجه آزمایش به ازای مقادیر از ۷ تا ۱۳ تکرار خواهد شد تا بهترین حالت ممکن انتخاب شود. نتایج حاصل از تکرار آزمایش در جدول ۱۴ خلاصه‌سازی شده است.

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

۳,۲,۵ حداقل نمونه‌های برگ

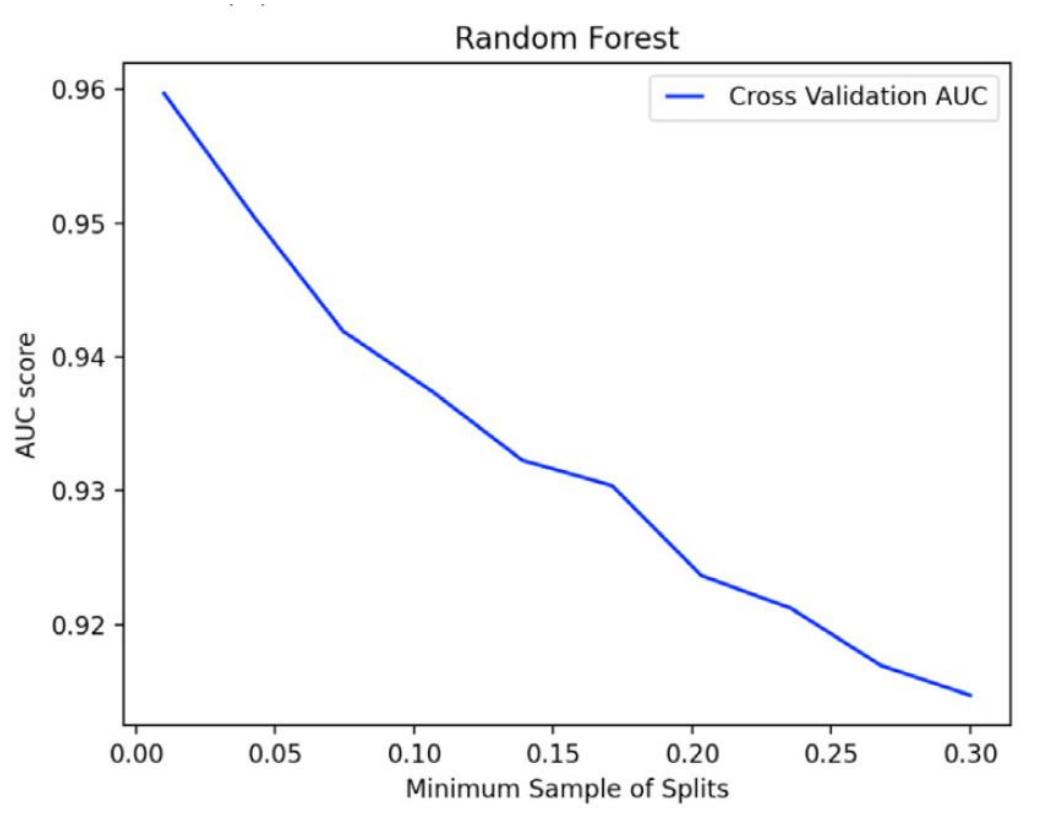
به طور مشابه با درخت تصمیم، آزمایش لازم بر روی هایپرپارامتر حداقل تعداد نمونه‌های مورد نیاز برای برگ صورت می‌گیرد. نتایج حاصل از این آزمایش در شکل ۹(د) نمایش داده شده است. همان‌طور که قابل مشاهده است، بهترین مقدار ممکن برابر است با کوچک‌ترین مقدار.



(d) With different *minimum_samples_leaf* values.

۴,۲,۵ حداقل نمونه‌های تقسیم

به طور مشابه با درخت تصمیم، آزمایش لازم بر روی هایپرپارامتر حداقل تعداد نمونه‌های مورد نیاز برای تقسیم صورت می‌گیرد. نتایج حاصل از این آزمایش در شکل ۹(ج) نمایش داده شده است. همان‌طور که قابل مشاهده است، بهترین مقدار ممکن برابر است با کوچک‌ترین مقدار.



(c) With different *minimum_samples_split* values.

۳,۵ رگرسیون لجستیک

الگوریتم رگرسیون لجستیک دارای ۲ هایپرپارامتر است. اولین مورد جریمه[۴۷]است که نوعی نرمال ساز است و در الگوریتم از آن استفاده می‌شود. هایپرپارامتر دوم معکوس قدرت منظم‌سازی[۴۸]است. در واقع مقدار کوچک‌تر این هایپرپارامتر به معنای سخت‌گیری بیش‌تر الگوریتم است. یا به عبارت دیگر می‌توان گفت که مقدار کوچک‌تر یعنی تنظیم قوی‌تر. در ادامه الگوریتم با هایپرپارامترهای مختلف بررسی و آزمایش خواهد شد.

[۴۷] Penalty

[۴۸] Inverse of Regularization Strength (C)

۱،۳،۵ جریمه

جریمه نوعی نرمال ساز است که در الگوریتم از آن استفاده می‌شود. می‌توان جریمه را برابر با L1 و یا L2 در نظر گرفت و یا این‌که اصلاً جریمه‌ای در نظر نگرفت. به ازای همه مقادیر گفته شده، دقت‌های متفاوتی به دست خواهد آمد. نتایج حاصل از آزمایش جریمه‌ها در شکل ۱۰ خلاصه‌سازی شده است.

```
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.

۲،۳،۵ معکوس قدرت منظم‌سازی

برای این هایپر پارامتر ۳۰ مقدار با فاصله یکسان بین ۰/۰۱ تا ۱/۵ بر روی C تنظیم شده است. همان‌طور که در شکل ۱۱ نشان داده شده است، AUC در حدود ۱ به بالاترین حد خود می‌رسد و AUC پس از آن تغییری نمی‌کند.

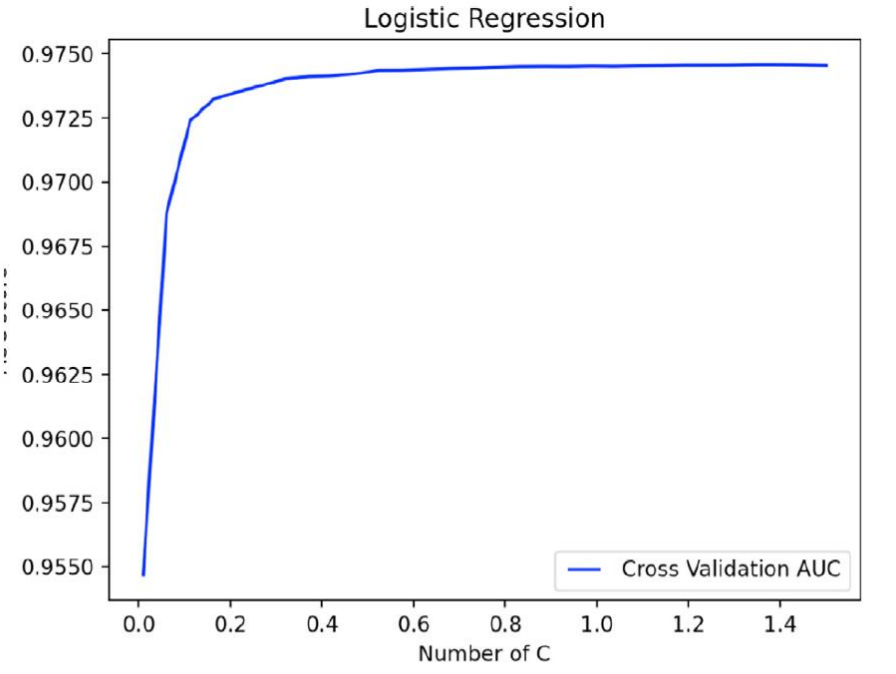


Fig. 11. The result of AUC with different *C* values.

برای اطمینان از این‌که کدام مقدار با مقدار دقیق بهترین است، مقدار AUC در ۰/۹، ۱/۰، ۱/۱ و ۱/۲ ارزیابی شد و نتایج در جدول ۱۵ خلاصه شده است. با توجه به یکسان بودن مقادیر، کوچک‌ترین مقدار C یعنی ۰/۹ انتخاب می‌شود.

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**

۴,۵ خلاصه هایپرپارامترها

در فصلی که گذشت، آزمایش‌های لازم برای به دست آوردن بهترین حالت هایپرپارامتر انجام شد. پس از انجام آزمایش روی هایپرپارامترهای متفاوت، نتایج به دست آمده به صورت زیر هستند:

- درخت تصمیم
  - معیار = آنتروپی
  - تقسیم‌کننده = بهترین حالت
  - حداکثر عمق = ۶
  - حداقل نمونه‌های تقسیم = ۵۰
  - حداقل نمونه‌های برگ = ۱
- جنگل تصادفی
  - معیار = آنتروپی
  - تقسیم‌کننده = بهترین حالت
  - حداکثر عمق = ۱۰
  - حداقل نمونه‌های تقسیم = ۲
  - حداقل نمونه‌های برگ = ۱
  - تعداد درخت‌ها = ۱۰۰
- رگرسیون لجستیک
  - جریمه = L1
  - معکوس قدرت منظم‌سازی = ۰/۹

۶ پیش‌بینی ریزش و ارزیابی

پس از پیدا کردن بهترین هایپرپارامترهای ممکن، باید هر ۳ مدل را بر روی داده‌ها پیاده کرد و پس از عمل مدل‌سازی، باید ارزیابی را انجام داد تا بهترین مدل پیدا شود.

۱,۶ مدل‌سازی و پیش‌بینی ریزش

مدل نهایی هر الگوریتم با بهترین هایپرپارامترهای گزارش شده در زیربخش‌های قبلی ایجاد شد. همچنین همان‌طور که در جدول ۷ نمایش داده شد، از مجموعه داده features_pc با تقسیم ۹۰-۱۰ باید بهره گرفته شود که مطلوب‌ترین حالت برای داده‌ها می‌باشد. جدول ۷ برای یادآوری مجدداً در زیر آورده شده است.

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.

پس از انجام عمل مدل‌سازی طبق هایپرپارامترهای پیدا شده، باید ارزیابی را بر روی مدل‌ها انجام داد تا از صحت و بهتر بودن هر کدام از مدل‌ها مطمئن شد.

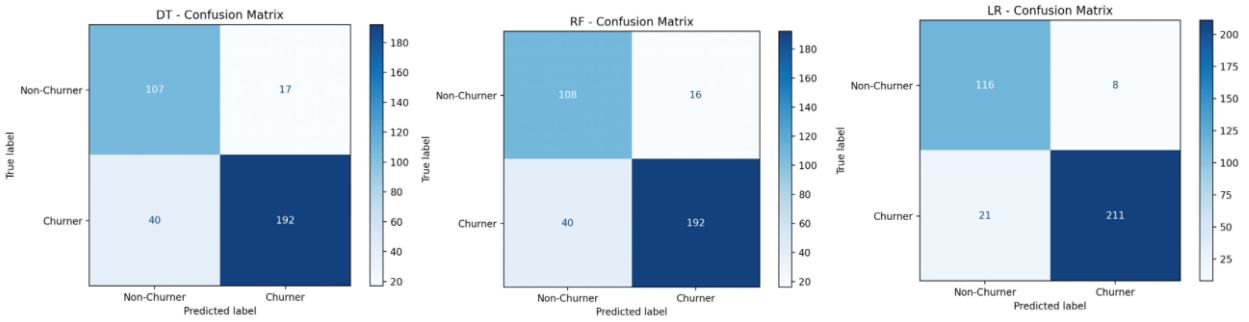
۲,۶ ارزیابی مدل

برای انجام عمل ارزیابی از ماتریس درهم‌ریختگی (ماتریس اغتشاش) کمک گرفته شده است. معیار AUC صرفاً برای پیدا کردن بهترین حالت ممکن (یعنی مجموعه داده و تقسیم داده‌ها به دو حالت آموزشی و آزمایشی) بود. دلیل بررسی سایر معیارها این است که AUC کل عوامل پشت صحنه مدل‌ها را توضیح نمی‌دهد و به معیارهای ارزیابی دیگر هم نیاز است. به طور مثال ۳ معیار ارزیابی دقت[۵۹] یادآوری[۶۰] امتیاز F1 نیز برای ارزیابی آورده شده اند.

دقت به معنای نسبت موارد مثبت پیش‌بینی شده صحیح به کل موارد مثبت پیش‌بینی شده است. یادآوری نسبت مشاهدات مثبت واقعی است که به درستی مثبت پیش‌بینی شده اند. به طور کلی، دقت و یادآوری در تنش هستند. بنابراین، بهبود یک معیار باعث کاهش دیگری می‌شود. امتیاز F1 معیاری برای مشاهده تعادل بین دقت و یادآوری است.

[۵۹] Precision
[۶۰] Recall

ماتریس درهم‌ریختگی برای هر ۳ مدل در شکل ۱۲ ترسیم شده است. در این ماتریس مثبت واقعی (TP) نشان‌دهنده یک کاربر ریزش‌کننده است که به درستی به صورت ریزش‌کننده شناسایی شده است. منفی واقعی (TN) نشان‌دهنده یک فرد غیر ریزش‌کننده است که به درستی یک فرد غیر ریزش‌کننده شناسایی شده است. از نتایج کلی پیش‌بینی ۳ مدل می‌توان گفت که پیش‌بینی‌ها عمدتاً درست هستند؛ زیرا تعداد TP و TN بدون بررسی سایر معیارها از اکثریت بیش‌تر است. هم‌چنین می‌توان متوجه شد که تعداد TP ها به طور قابل توجهی از TN ها بیش‌تر است که علت این امر، عدم تعادل مقدار هدف است. این در مورد رابطه منفی کاذب (FN) و مثبت کاذب (FP) نیز صدق می‌کند.



(a) Decision tree. (b) Random forest. (c) Logistic regression.

فرمول معیارهایی که در بالا قرار داشتند، یعنی معیارهای دقت، یادآوری و امتیاز F1 در ادامه آورده شده است. موارد مثبت واقعی، منفی واقعی، مثبت کاذب و منفی کاذب درون ماتریس‌های درهم‌ریختگی قابل مشاهده است.

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

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

$$\text{F1} - \text{score} = ۲ \times \frac{precision \cdot recall}{precision + recall}$$

سپس با استفاده از گزارش طبقه‌بندی مقادیر مشخصی بررسی شد. گزارش‌ها در شکل ۱۳ نشان دادهد شده است. با توجه به نتیجه، کل مشاهدات ۳۵۶ است و تعداد ریزش‌کننده‌ها (۱ در گزارش) برابر با ۲۳۲ می‌باشد و غیر ریزش‌کننده‌ها (۰ در گزارش) نیز برابر با ۱۲۴ می‌باشد. این بدین معنی است که ۶۵/۱۶ ٪ از داده‌های تست، ریزش‌کننده‌ها هستند که این درصد بیش‌تر از نسبت واقعی افراد ریزش‌کننده و غیر ریزش‌کننده در کل داده‌ها می‌باشد. در واقع فقط ۵۶/۷ ٪ داده‌ها را افراد ریزش‌کننده تشکیل می‌دادند که در این داده‌های تست اینگونه نمی‌باشد. دو میانگین کلان[۶۱] و میانگین وزن‌دار[۶۲] نیز در گزارش طبقه‌بندی وجود دارد که می‌توان از آن‌ها استفاده کرد. میانگین وزن‌دار برای داده‌های نامتوازن[۶۳] مناسب‌تر می‌باشد. می‌توان در محاسبه معیارها از میانگین وزن‌دار استفاده کرد؛ زیرا داده‌های تست نامتوازن هستند.

[۶۱] Macro Average
[۶۲] Weighted Average
[۶۳] Imbalanced Data

```
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.

خلاصه معیارهای هر ۳ مدل در جدول ۱۶ ارائه شده است. دقت، یادآوری و امتیاز F1 از میانگین وزن‌دار گرفته شده است. طبق نتایج موجود در جدول می‌توان نتیجه گرفت که مدل رگرسیون لجستیک بهترین مدل را دارد؛ زیرا بالاترین AUC و هم‌چنین امتیاز F1 مربوط به این مدل است. امتیاز F1 این مدل که برابر با ۰/۹۱۹۴ می‌باشد، بالا است و این بدین معنی است که دقت و یادآوری به خوبی متعادل هستند. اگرچه که درخت تصمیم و جنگل تصادفی عملکرد خوبی از AUC دارند؛ ولی عملکرد آن‌ها ضعیف‌تر از مدل رگرسیون لجستیک است.

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**

۳,۶ رگرسیون لجستیک بهتر از درخت تصمیم و جنگل تصادفی است

ویژگی‌های الگوریتم مبتنی بر درخت ممکن است باعث تشابه نتایج درخت تصمیم و جنگل تصادفی شده باشد. دلایلی که رگرسیون لجستیک را به عنوان مدل بهتر معرفی می‌کند، می‌تواند به شرح زیر باشد:

- رگرسیون لجستیک تمایل دارد که با داده‌ها با ابعاد بالا بیش از حد سازگار[۶۴]شود، اما در این مسئله یک مجموعه داده با ابعاد پایین وجود داشت.
- رگرسیون لجستیک با مجموعه داده‌هایی که به صورت خطی جداپذیر هستند، به خوبی کار می‌کند. در حالی که درخت تصمیم با داده‌های عددی پیوسته کارایی کم‌تری دارد.
- بسیاری از متغیرهای استفاده شده در دادگان به صورت عددی پیوسته با یک بازه‌ی بزرگ بودند و تفاوت میان انحراف معیار و میانگین قابل توجه بود. پس این می‌تواند بر مدل‌های مبتنی بر درخت تأثیر بگذارد. به همین علت مدل رگرسیون لجستیک بهتر از مدل‌های درخت تصمیم و جنگل تصادفی عمل می‌کند.

۷ بحث و نتیجه‌گیری

ترکیبی از نرخ ریزش تعریف شده و پیش‌بینی ارائه شده در بالا، اثربخیش رویکرد پیشنهادی در DGBL را نشان می‌دهد. این تعریف می‌تواند برای بسیاری از سرویس‌هایی که در حوزه یادگیری مبتنی بر بازی دیجیتال فعالیت می‌کنند، کمک‌کننده باشد؛ حتی اگر مدت دوره مشخص نباشد یا مدت‌های متفاوتی وجود داشته باشد. هم‌چنین سه دسته‌بندی استفاده شده (ویژگی‌های جمعیت شناختی، تعهد و کارایی) می‌تواند برای هر شرکت مرتبط با DGBL پیاده‌سازی شود.

[۶۴] Overfit

رویکرد فوق به ما اجازه می‌دهد که با انعطاف زیادی ریزش‌کننده‌ها را تعریف کنیم؛ حتی اگر کاربران در اواسط دوره وارد بازی شوند. این قابلیت اجرا و انعطاف تعریف ریزش پیشنهاد ششده می‌تواند چالش‌های متفاوتی که امکان دارد به وجود بیاید را برطرف کند.

در نهایت به یک نقل قول مهم در یکی از مقاله‌هایی که در سال ۲۰۲۳ چاپ شده اشاره می‌شود: «حدود نیمی از دانش‌آموزان که مورد بررسی قرار دادیم، بازی‌های آموزشی را راهی مؤثر برای یادگیری مطالب جدید دانستند.»

۸ مراجع

[۱] https://google.com/amp/s/resources.owllabs.com/blog/educational-technology%3fhs_amp=true

[۲] https://dayamooz.co/یادگیری-مبتنی-بر-بازی/

[۳] https://ilearn4health.eu./digital-game-based-learning-for-health-education/

[۴] V. Strauss, New report on virtual education: ‘it sure sounds good. As it turns out, it’s too good to be true, ۲۰۱۹, URL: https://www.washingtonpost.com/education/۲۰۱۹/۰۵/۲۹/new-report-virtual-education-it-sure-sounds-good-it-turns-out-its-too-good-be-true

[۵] Recurly, Benchmarks for subscription E-commerce, 2019: https://info.recurly.com/research/benchmarks-for-subscription-ecommerce

پیوست

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)	*ch*1_*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)	*ch7_playtime* (min)	*avr_ch_wait* (days)	*exp*	*coins*	*replay*
PC-1	0.000001	1.00E−06	8.98E−07	5.97E−07	7.69E−07	2.00E−06	2.00E−06	3.49E−07
PC-2	−0.001188	−0.001	4.60E−04	−1.01E−03	−9.65E−04	0.001	0.0012	−1.27E−03
PC-3	0.007957	0.0091	1.01E−02	8.35E−03	1.47E−03	0.0128	0.012	−1.08E−03
PC-4	0.147524	0.1502	1.50E−01	1.33E−01	−3.16E−02	0.1635	0.1639	8.41E−02
PC-5	0.281479	0.2694	2.48E−01	2.29E−01	−7.15E−02	0.2839	0.2834	2.03E−01
PC-6	0.168011	0.1533	6.46E−02	2.74E−01	2.91E−01	−0.157	−0.149	5.52E−01
PC-7	0.072236	0.0241	−1.11E−02	8.13E−02	−4.48E−01	−0.139	−0.129	2.24E−01
PC-8	0.024768	−0.171	−3.29E−01	−2.21E−01	−2.78E−02	−0.111	−0.1	2.66E−01
PC-9	0.062961	−0.065	−7.69E−02	−8.02E−02	8.19E−01	−0.007	−0.01	2.76E−03
PC-10	0.440438	−0.146	−3.06E−01	−3.69E−01	−1.23E−01	−0.041	−0.04	4.40E−01
PC-11	0.02843	−0.056	−8.19E−03	−2.00E−01	−1.33E−01	0.0109	0.0045	−1.87E−01
PC-12	−0.211524	−0.392	−3.56E−01	7.06E−01	−5.77E−02	−0.041	−0.047	9.85E−02
PC-13	0.161819	−0.402	−1.43E−01	1.26E−01	3.37E−02	0.3078	0.3135	−8.56E−02
PC-14	−0.39824	0.3272	4.34E−02	4.93E−02	7.15E−03	−0.269	−0.268	1.36E−01
PC-15	−0.140912	−0.587	7.19E−01	−1.07E−01	−4.31E−03	−0.063	−0.058	2.69E−01
PC-16	0.202304	−0.021	−1.09E−02	2.98E−02	−1.19E−02	−0.113	−0.12	−1.88E−01
PC-17	0.377125	−0.183	9.63E−02	1.09E−01	7.23E−03	−0.358	−0.347	−3.43E−01
PC-18	−0.409904	−0.036	−1.08E−01	−1.91E−01	8.90E−04	0.1577	0.1674	1.94E−01
PC-19	−0.001526	0.0035	−9.50E−04	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