THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

CHURN PREDICTION IN HYPER-CASUAL GAMES

Master's Thesis

EMRE CAN BİÇER

ISTANBUL, 2021

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL
BIG DATA ANALYTICS AND MANAGEMENT
MASTER'S PROGRAM

CHURN PREDICTION IN HYPER-CASUAL GAMES

Master's Thesis

EMRE CAN BİÇER

Thesis Supervisor: ASSOC. PROF. DR. YÜCEL BATU SALMAN

ISTANBUL, 2021

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL

…../……/…..

MASTER THESIS APPROVAL FORM

Program Name:	BIG DATA ANALYTICS AND MANAGEMENT
Student's Name and Surname:	EMRE CAN BİÇER
Name Of The Thesis:	CHURN PREDICTION IN HYPER-CASUAL GAMES
Thesis Defense Date:	

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Prof. Dr. Fatma ÖZKUL

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Signature**
Thesis Advisor's	Assoc. Prof. Yücel Batu Salman	
Member's	Asst. Prof. Tamer Uçar	
Member's	Asst. Prof. Atınç Yılmaz	

ACKNOWLEDGEMENT

I pour out my thanks to my thesis supervisor Doc. Dr. Yucel Batu Salman for his consultancy and guidance. He supported my thesis process with his knowledge and leadership. I am very glad that I finished this thesis with his support.

Istanbul, 2021 Emre Can Biçer

ABSTRACT

CHURN PREDICTION IN HYPER-CASUAL GAMES

Emre Can Biçer

Big Data Analytics and Management Master's Program

Thesis Supervisor: Assoc. Prof. Yücel Batu Salman

December 2021, 42 Pages

People's interest in mobile games is increasing day by day. Especially in recent years, when smartphones have become widespread, both the number of players and the number of investments made in this field have increased. With such an accessible area, it is possible to see millions or even billions of users for some games. As the number of users has increased, the concepts of churn and retention have become more important. This study focuses on determining the causes of churn in the hyper-casual game genre, which has been a rising value in recent years. In the study, modeling has been made with different machine learning algorithms over real data of a hyper-casual game whose company and game name are kept secret. Within the scope of churn analysis, the definition of churn was determined as seven days due to the characteristics of the data and the distribution of revenue. According to the churn definitions made for the study, the results of different machine learning algorithms were compared. As a result, important variables that cause churn for the hyper-casual game used in the study were determined.

Keywords: Churn, Retention Rate, Hyper-Casual Game, Machine Learning

ÖZET

HİPER GÜNDELİK OYUNLARDA CHURN TAHMİNİ

Emre Can Biçer

Büyük Veri Analitiği ve Yönetimi Yüksek Lisans Programı

Tez Danışmanı: Doç. Dr. Yücel Batu Salman

Aralık 2021, 42 Sayfa

İnsanların mobil oyunlara olan ilgisi her geçen gün artıyor. Özellikle akıllı telefonların yaygınlaştığı son yıllarda hem oyuncu sayısı hem de bu alana yapılan yatırımların sayısı arttı. Böylesine erişilebilir bir alan ile bazı oyunlar için milyonlarca hatta milyarlarca kullanıcı görmek mümkün. Kullanıcı sayısı arttıkça, kullanıcı kaybı ve elde tutma kavramları daha önemli hale geldi. Bu çalışma, son yıllarda yükselen bir değer olan hiper-gündelik oyun türünde kullanıcı kaybının nedenlerini belirlemeye odaklanıyor. Çalışmada, şirket ve oyun adı gizli tutulan bir hiper-gündelik oyunun gerçek verileri üzerinden farklı makine öğrenme algoritmaları ile modelleme yapılmıştır. Kullanıcı elde tutma oran analizi kapsamında verilerin özellikleri ve gelir dağılımına göre kayıp tanımı yedi gün olarak belirlendi. Çalışma için yapılan kayıp tanımlarına göre farklı makine öğrenme algoritmalarının sonuçları karşılaştırıldı. Sonuç olarak, çalışmada kullanılan hiper-gündelik oyun için oyun kaybına neden olan önemli değişkenler belirlenmiştir.

Anahtar Kelimeler: Kullanıcı Kaybı, Elde Tutma Oranı, Hiper-Gündelik Oyun, Makine Öğrenimi

TABLE OF CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AUC	: Area Under Curve
DNN	: Deep Neural Network
KNN	: Cash on Door
NN	: Neural Network
SVM	: K-Nearest Neighbor

1. INTRODUCTION

In today's world, where digitalized and smart phones are widespread, the interest in mobile games is constantly increasing and they are becoming easier to be reached. The concept of churn has become very important, with the number of million-level players and even the billions in some games. The high number of users will increase the income in the mobile gaming industry. The total number of mobile game users in 2021 is expected to exceed 2.7 billion, and revenue is expected to exceed 100 billion dollars by 2023 (Statista 2021).

With the increase in the number of users and revenue, the perspective of game developers on games also changed. How the level designs should be, when and how the advertisements shown in the game will appear, the importance of the payment transactions for the game, the campaigns given to the players have become important points. In fact, these are issues related to the concept of churn in general. User trends towards these given examples and their counterparts have become focal points in churn analysis. For this reason, the game developers started to collect the data in their hands in the best way and use them to extract insight with various tools. However, churn models led to differentiation and game developers to get to know their users better.

This study has been done on churn analysis in the gaming industry. The focus of the study is to support churn analysis with various machine learning models and make inferences over variables. Studies in the literature that may be related to this subject have been examined and included in the study. It was focused on which models were used, which variables were found important, and the study was designed accordingly. The data analysis and modeling part required for churn is made for a game belonging to a hyper-casual genre whose company and game name are kept secret. In this context, studies of all game genres and other sectors have been examined and shared within the scope of the study. Studies in every sector in the literature have shown that the machine learning algorithms applied are similar to each other, and the definition of churn changes only according to the sectors and business. For this reason, the churn analysis and applied models in this study were chosen accordingly.

Descriptive analyzes were performed first, similar to the studies in other works in the literature. Later, after the churn definition was made, models were developed with different machine learning algorithms. By comparing these models, those giving the best results were determined and important variables causing churn were determined.

In this study, the subject that is mainly examined is hyper-casual games. The dataset used in the study was also selected in this context. Although hyper-casual games are similar to casual games with some features, they have their own characteristics due to both game dynamics and gameplay styles. For this reason, they differ from both casual and other genres.

Hyper-casual games, which distinguish them from casual games in terms of game dynamics, are launched in a simplified way. In other words, hyper-casual games are the genre of games that give importance to relatively simple visuals. The most important features are that they are free, designed as never ending, and played with one hand. A lot of hyper-casual games have been released recently, as copies of each other because of their simplified structure.

Due to the large number of users, the importance of churn has increased. In this study, it is aimed to determine the causes of churn by using various machine learning models with the determined variables of our game. Open source tools are used in this study to create Churn models. Both R programming and python's libraries suitable for analysis and models are used. The game that is used in the work is selected for the study can be included in the hyper-casual game category in terms of game mechanics. Until now, there has been no study examining hyper-casual games through the concept of churn. The contribution of the study to the literature is that it is the first work done in the hyper-casual game concept and it contributes to churn analysis by analyzing hyper-casual games. Due to the increasing interest in this genre, it is quite possible that the studies to be carried out in this way will increase in the future.

1.1 HYPER-CASUAL GAME HISTORY

The emergence and popularity of the term hyper-casual game dates back to 2017. First of all, comes the laundering 'Flappy Bird' game. This simple game, which has received more than 50 million downloads and is world-famous, may have enabled many similar games in this genre. World-renowned companies such as Kwalee,

Ketchapp and Voodoo are the leading companies for this genre. In our country, many game companies that have turned their focus to the hyper-casual genre have recently been established and the game has been released. Since the first day of their release, these games have generally focused on 2D game design, have been free and presented to users without a tutorial unlike other games. Advertising revenues from hyper- casual games consist of rewarded, banner and interstitial categories.

1.2 COMPRASION of CASUAL and HYPER-CASUAL GAMES

Casual games are games that can belong to many different genres that almost everyone plays. Many games that can be classified into categories such as puzzle, racing, action can also be classified as casual. The hyper-casual genre, which appears as a different version of casual games, has become a genre whose name has been mentioned in recent years. It can differ from casual games both in terms of game dynamics and gameplay style. In addition, the most important features are the game life.
Hyper-casual games differ from casual games in that they have a simple game mechanics. They offer 'tap-to-play' or ’swipe-to-play’ or ‘hold-to-play’ and the convenience of playing with one hand. With the widespread use of such games, many world-famous companies have started to release such games. Their most important features are that almost all of them are free to play and there is no need for tutorial and instructions to play. While casual games try to target more loyal gamers who bring more revenue, hyper-casual games appeal to the mass market. They aim simple mechanics, large number of people and low number of retention days when compare to casual games. While it is

played in this way and the effect on the user is different, its important feature is that it disappears quickly. The emergence of many similar hyper-casual games has also been effective in this (GameAnalytics 2021).

1.3 RETENTION RATES in HYPER-CASUAL GENRE

Hyper-Casual games are known for their short life span. A large part of its revenue is generated by the performance of the users in the first days. According to GameAnalytics analysis, day1 retention should be at least 40% and day7 retention should be at least 15% for this category. Otherwise, the game will not be profitable.

For this reason, it is thought that at least these rates should be achieved in order to be considered successful and continue investing (GameAnalytics 2021). For this reason, it is important to investigate the causes of churn and act accordingly. In this way, it is easier to keep the users inside and thus the revenue is increased. Considering the retention rates of day30 except day1 and day7, quite low values are seen. Because of their short life span, almost all of the users churn by the thirtieth day. The dataset to be used in this study also confirms this.

If we look at the game examined in this study from this perspective, the day1 retention rate is around forty eight percent, while the day7 retention rate is around twenty two percent. Therefore, we can say that the hyper-casual game discussed in the study is successful in retention. In line with the features of hyper-casual games, the thirtieth day retention rate is only percent 1.3.

2. RELATED STUDIES

Various studies have been done for the churn in the gaming industry. While there are online and casual studies according to game types, there is no special work belonging to the hyper-casual category. Within the scope of this study, the studies in the literature were examined under three different main titles as those related to casual games, those related to online games and those related to other sectors such as telecom, bank, insurance and e-commerce. Online games are mostly strategy, while casual games include games of different genres such as puzzle, racing, action. In this way, churn studies in different game types will be examined and analytical models will be applied accordingly. Similar studies in both the game industry and other sectors will shed light on the analysis and models to be made in this study.

The following studies have been examined within the scope of churn studies for online games.

One of the most important factors in online games, the concept of event positively affects especially old loyal players. In the study conducted on this subject in 2019, it was found that this issue was especially positive for the audience who played the game for a long time (Park et al. 2019). Another study is based on The Settlers Online game. In this study, churn analysis was performed by comparing more than 1000 variables and 8 different machine learning models (Rothmeier et al. 2020). In their work random forest model gave more than 99% AUC performance. Churn definition was determined as 14 days in this study. In another study, the effect of social network analysis on online games was examined (Otte et al. 2002). Concepts such as group and clan formation, which are frequently encountered in online games, enable the relations between players to be examined. In this study, it was determined that direct or indirect relationships established between actors can be used in churn works (Otte et al. 2002). Another churn study on online games was done for EverQuest II (EQII). The authors took the study from the theories on this subject and compared it with real data studies in the EQII game. They analyzed the player motivation they gained through different approaches and supported them with their studies. As a result of their work with real dataset on the EQII game, they concluded that the motivation of the player is the key factor on the churn analysis on that game (Zoheb et al. 2012). Kim et al. (2017) conducted a churn study by examining 3 games. One of these 3 games was an online game called TagPro. In the study where more than 10 million player logs were examined, variables such as activeDuration, consecutivePlayRatio, playCount, winRatio with high AUC values were found to be important. A churn study was conducted on Thirty-six Stratagems, another online game, in 2019. In this study, the of churn was determined as 3 days, and different machine learning algorithms such as logistic regression and random forest were used in the study. The AUC values for this game were found to be 0.711 for the logistic regression, 0.739 for the support vector machine, and 0.761 for the randomforest (Yang et al. 2019). In that work, they developed features, and they showed that developed features gave better results than the model having baseline feature. In the study conducted on the Age of Ishtaria online game, positive results were obtained by churn analysis with variables such as purchase amount, last level (Bertens et al. 2017). Their model was relied on conditional inference survival ensembles, and it shows that playtime and level degree have important impact on churn and retention analysis. Another research on this subject was conducted for the famous South Korean game Aion (Lee et al. 2020). Since most of the income in online games comes from loyal users, this research has focused on this. As many users in the game do not contribute to the game in terms of revenue, a model has been developed to help maximize revenue, a model has been developed to help maximize revenue for churn. Castro and Tsuzuki (2015) proposed a frequency analysis approach for churn estimation of features. The analytical modeling in this study was done with a 9: 1 train-test ratio, and the k-NN classification algorithm was used. In another study in this area, high churn performances were obtained by excluding purchasing variables by using only cognitive psychological features on the game called Crazy Dragon (Jeon et al. 2017). They used decision tree and random forest models. Since their dataset was imbalanced, they applied undersampling methods. For both algorithms undersampling methods gave better results. Another study was made for the game Top Eleven - Be A Football Manager using two million data (Milosevic et al. 2012). In this study, logistic regression, decision tree, random forest, Naive Bayes and gradient boosting models were used. Among the models, gradient boosting gave the best results. In this study for Top Eleven, the AUC values were 0.79 for logistic regression, 0.67 for decision tree, 0.80 for spv and 0.73 for naive bayes. Another study has been done on how online games affect usability retention.

(Papaloukas et al. 2010). In this study, the world famous "Age of Mythology" and "Seven Kingdoms" plays have been examined. It was made for the Sony Everquest II game for another online game study (Srivastava et al. 2012). for the Sony Everquest II game for another online game study (Srivastava et al, 2012). As a result of their study, they revealed that actor motivation is the most important factor affecting the churn. In another study conducted with online games, it has been revealed that establishing a social network and being included in groups is an important factor for users (Pirker 2018).

The following studies have been examined within the scope of churn studies for casual games.

The churn study for Diamond Dash and Monster World Flash, which are the world-famous casual games, was made using 4 different models (Runge et al. 2014). In this study, the definition of churn is made over 14 days, and descriptive analyzes show that top percent 7 of paying users contribute approximately 50% of the total income. Neural network gave the best AUC performances for both games. That algorithm gave 0.815 value for Diamond Dash and gave 0.93 for Monster World. However, logistic regression has worked almost as well as a neural network. Decision Tree and Support Vector Machine showed low performance especially for Diamond Dash game. The AUC values for Diamond Dash are 0.815, 0.814, 0.732 and 0.762 for neural network, logistic regression, decision tree and support vector machine algorithms, respectively. In addition, the values for Monster World Flash are 0.930, 0.924, 0.850 and 0.903 for neural network, logistic regression, decision tree and support vector machine algorithms, respectively. (Runge et al. 2014). In another study, churn analysis was performed using Discriminant Analysis, Collinearity Test, Network Analysis on features and Bayesian Belief Network methods for the Wuzzit Trouble game and the results were presented (Qui Wang 2018). In this study, the aim gives analytical research design of the churn analysis graphically. On the other hand, the author indicated that lacking data was the biggest restriction for this research. In another study, people's interest in casual game types according to age was investigated (Chesham et al. 2017). They showed that casual video games have properties that are appropriate for older people. However, this work was done with 16 healthy adult people. In a study that contributed to online games, 2 games were analyzed in the casual category (Kim et al. 2017). In this study, Dodge the Mud, a casual action type, and an unnamed casual racing game were analyzed. In the context of the work different algorithms were used such as logistic regression, gradient boosting, and random forests) and two deep learning algorithms (CNN and LSTM), and sensitivity analysis for OP and CP. The two most important variables for both casual games were the duration of activity and the play count. In this study, logistic regression, gradient boosting, random forest, CNN and LSTM algorithms were used for 3 different games. AUC values of 0.786, 0.791, 0.787, 0.774 and 0.792 were found for the first game, respectively. For game 2, AUC values of 0.711, 0.728, 0.722, 0.691 and 0.686 were found, respectively. The AUC values found for the 3rd game are 0.824, 0.824, 0.822, 0.798 and 0.803, respectively (Kim et al. 2017). In addition, the other work they also made new player analysis in their work. In another study for mobile casual games, the definition of churn was made as 10 days, and more than one technique was tried (Perianez et al. 2017). They applied survival ensemble, support vector machine, naive Bayes and decision tree. In their work survival ensemble gave the best AUC performance. Morevoer, last purchase amount and days since last purchase are the most important variables for survival ensemble algorithms. In another study, unlike other studies, a propose model was developed over the data of many games obtained from Samsung Game Launcher, and it was determined that the developed propose model gave better results than other techniques (Liu et al. 2018). They applied different machine learning algorithms and they consider churn definition as 14 days. Harrison and Roberts (2013) made an analytics-driven dynamic game adaption study for Scrabble, which is considered a casual puzzle game. Morevoer, the same authors also conducted a psychometric evaluation study for casual games (Harrison and Roberts 2015). In another study, churn estimation has been made for 6 different games. Logistic Regression (LR), support vector machine (SVM), Random Forests (RF)) and the aforementioned features were used for churn and retention prediction. In addition, within the scope of this study, in addition to the base variables, different variables were produced and it was determined that the variables produced later were more effective in churn estimation (Yang et al. 2019). In a study similar to other studies, 'Number of Sessions' and 'Number of Days' emerged as important variables from churn analysis (Hadiji et al. 2014). They applied decision tree, neural network, logistic regression and naïve Bayes algorithms, and decision tree gave the best performance. In addition to other model studies, the study was made in order to assess the churn and retention tendency of users by analyzing sentiments of users from the comments on games by implementing deep learning and word embedding models (Kilimci et al. 2020). It is seen that text-based analysis can also be done through the comments of this study. Word2vec, GloVe and FastText text based embedding algorithms were used in this study. These studies were made for 4 different games, and CNN, RNN, LSTM, as well as Glove, FastText and Word2Vec models were used. Another text-based churn research was conducted on endless runner games. The world-famous Temple Run and JetPack Joyride are the biggest examples of such games. In the study, some words for text mining were classified as 'high score' and some words as 'low score'. Continuing in this way, 3 segments were created as high score textual feedback, low score textual feedback and no textual feedback. As a result, it was determined that the high score textual feedback segment has a positive effect on churn (Padte 2019). In another casual game research, churn studies were performed with different models (Kristensen and Burelli 2019). LSTM models gave the best performance for AUC, accuracy and f1 score. In this study, it is stated that the variables showing the degree of activity in recent times are important indicators for churn. Another study was researched with a survey of 147 people who quit the Facebook-based YoWorld game (Bergstrom 2019). More than half of the participants in the study stated that they could not find time in daily life due to various reasons and had to take a break. Only 5 of the 147 people in the study stated that they quit because they were bored. Since the study in the study proceeded through the questionnaire contrary to the general churn perception, and the audience was small, it was not supported in an analytical way.

The following studies have been examined within the scope of churn studies for other sectors.

One of the churn studies in the Telecom industry was done in 2019 (Hargreaves 2019). In this study, churn and revenue effects are discussed together. Revenue was ten times the investment made in the Churn strategy. In the study, logistic regression provided the highest accuracy value with 75.3%. Another study in the telecom sector was carried out in 2020 (Lalwani et al 2020). In this study Regression, naive Bayes, support vector machine, random forest, decision trees, boosting methods are used. Adaboost and XGBoost Classifiers methods gave the best results with an accuracy of 84%. In this study, train test ratios were used as 4: 1. Another study in the telecom industry was conducted in 2018 using logistic regression, support vector machine, random forest and gradient boosting algorithms (Gaur and Dubey 2018). Gradient boosting algorithm gave the best accuracy value with approximately 84.6% in the study. For the dataset in the study, the train test ratio was taken as 3: 1. Another study in the telecom sector was discussed in 2015 (Vafeiadis et al 2015). In this study, firstly algorithms such as logistic regression, decision tree, naive Bayes are applied. Later, performance enhancing boosting algorithms were developed. The best overall classifier was the SVM-POLY using AdaBoost with accuracy of approximately 97% and F-measure over 84%. Another study in the telecom sector was conducted in 2019 (Ullah et al 2019). The random forest model applied in this study performed well with an accuracy of 88.6%. Another churn study in the telecom sector was carried out on customer service in 2018 (Sabbeh 2018). Among the machine learning algorithms used in this study, random forest and ADA boost gave the best performances with 96% accuracy values. In addition to these models, multi-layer perceptron, support vector machine, decision tree, naive Bayes, linear discriminant analysis and logistic regression algorithms have also yielded good results. Another study was done in the banking sector in 2016 (Kumar and Chandrakala 2016). In this study, Chinese bank dataset was used and Decision tree, ANN, SVM and hybrid models were used. Improved balance random forest model gave the highest accuracy results. Another study was carried out with a dataset from kaggle for the banking sector (Rahman and Kumar 2020). KNN, SVM, Decision Tree, and Random Forest algorithms were used in this study. The Random Forest model has been the best machine learning algorithm when comparing accuracy values. Another study in the field of banking was conducted in 2008 (Jing and Xing- hua 2008). In this study, a churn analysis has been made on commercial bank VIP customers. Machine learning algorithms such as artificial neural network, decision tree, logistic regression, support vector machine and naive Bayesian classifier were used in the study. While the accuracy values of all the algorithms applied in the study remained below 60%, the SVM algorithm gave the best result with an accuracy of 59.7%. A study in the field of retail banking was conducted with a data from a bank in Indonesia in 2019 (Karvana et al 2019). The models applied in the study were made with 57 variables. Five different machine learning algorithms were used in the study, and the support vector machine algorithm gave the best result. Unlike other studies, the train-test rate in the study was 50-50. Another study in retail banking was done in 2008 with 5000 records (Popovic 2008). Canonical discriminant analysis, k-means, fuzzy methods were used in the study. With the techniques used in the study, accuracy rates around 63% were calculated. Unlike other studies, another study was conducted on the insurance industry (He et al 2020). The dataset used in this study was provided by Markel Corporation. In the study, logistic regression, random forest, extremely randomized tree (extra trees) classifier, support vector machine, automated machine learning, neural network and gradient boosting algorithms were applied. While the models gave similar accuracy results, they were all below 70% accuracy. Another study in the insurance sector was conducted with a different technique in 2004 (Morik and Köpcke 2004). Algorithms were applied with the TF-IDF technique and high accuracy results were obtained. In the motor insurance sector, studies were conducted using four different models in 2016 (Bolance et al 2016). In the study, logistic regression, conditional tree, neural network and support vector machine algorithms were used. In terms of accuracy, the support vector machine algorithm gave the best result. A study within the scope of health insurance was conducted as a master thesis in 2015 (Huigevoort 2015). In the comparison of accuracy values of models, logistic regression and neural network algorithms are the methods that give the best results. Studies examining churn and retention models in the field of e-commerce have also been conducted in other sectors (Yu et al 2011). In this study, artificial neural network (ANN), decision tree, support vector machine and extended support vector machine algorithms are used. When the accuracy rates were compared, the extended support vector machine was the model with the best performance. Another study in the field of e-commerce was made with an e-commerce dataset in China in 2018 (Yayun 2018). Logistic regression, neural network, support vector machine and xgboost algorithms were applied over the dataset in this study. Although the models gave similar results from the accuracy calculations, the xgboost algorithm gave the best performance with an accuracy of 0.88. Another study was conducted with the data of a company working in the food delivery service from Iran (Raeisi & Sajedi 2020). The purpose of this study is to make churn estimation on customers by examining online features and user behavior. Various machine learning algorithms were applied in the study, and the algorithm using gradient boosted tree with an accuracy of 86.9% gave the best result. Another e-commerce churn modeling was done using support vector machine and neural network algorithms (Zhao 2014). In this study, firstly, predictions were made with SVM and NN models. Later, the SVM model was applied again by combining the estimates obtained different from other models. The last combined SVM model applied gave the best performance. A churn survey study, which can be considered as a summary of his work in other sectors, was discussed in 2018 (Santharam and Krishnan 2018). The aim of this study is to reveal which algorithms are used for churn models applied in all sectors. This study can also be called a summary of churn models according to sectors. In this study, it was shared that models such as logistic regression, random forest, decision tree, neural network, and support vector machine are used for churn models in almost all sectors. It is seen that fuzzy, k-means algorithms are used for patient churn analysis, and xgboost algorithms are preferred in studies involving subscriber data. It is stated that 18 different algorithms are used in various churn models in total, and the reasons causing churn are investigated with these models.

In the literature review, various studies belonging to both the game industry and other sectors have been examined. As a result of these investigations, although there are certain differences, it has been observed that the algorithms used in churn models are similar in many studies. The different points are related to the general definition of churn. It is possible to see different churn definitions in the sector, even within the game genres. For this reason, it would be more correct to examine the dataset to be used in the churn definition. Because each dataset may have its own characteristics, this should not be ignored when creating the target variable. The models applied in the studies and the variables used were examined and used appropriately in this study. Almost all of the models applied in similar studies in the literature were also used on the dataset in this study. Similar to the studies in the literature, the models were compared, and the models with the best results were specified as accuracy values. Although various analytical tools were used in the literature for the models, R and python tools were used in this study.

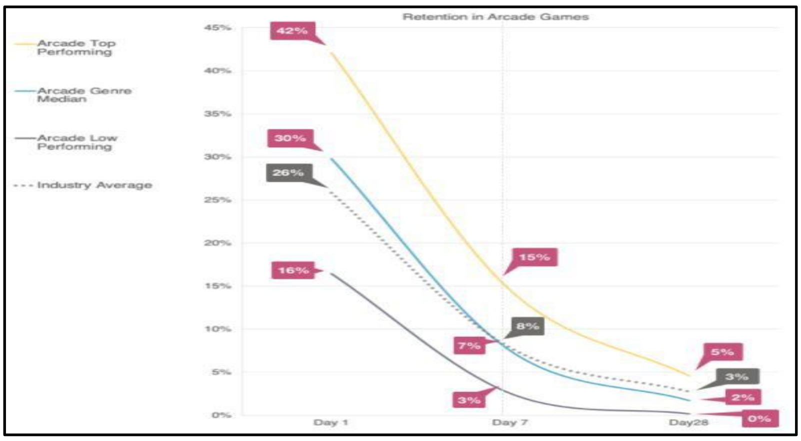
3. DATA ANALYSIS AND MODELLING

3.1 DESCRIPTIVE ANALYSIS

Before starting analytical studies, descriptive analyzes have been examined in this section to get to know the data and to reveal the features of hyper-casual games. Since the distribution of the variables to be used in the data according to the target variable, the characteristics of the variables such as level and revenue are important both in generation of new variables and selection of the target variable, their analysis has been made. Firstly, the features of the hyper-casual games in the sector have been revealed. Afterwards, it was compared whether it gave results compatible with the data used in the study.

As mentioned before, according to GameAnalytics data, day1 retention should be at least 40 percent in this category, and day7 retention should be at least 15 percent, let's say the game can earn money. These values are provided for the hyper-casual game examined in this study. It can be said that analyzed game achieved success according to the hyper-casual retention metrics. Below you can see the retention rates of arcade games shared by GameAnalytics.It is not officially named as Hyper-casual on store, it's being examined on the site under the name arcade.

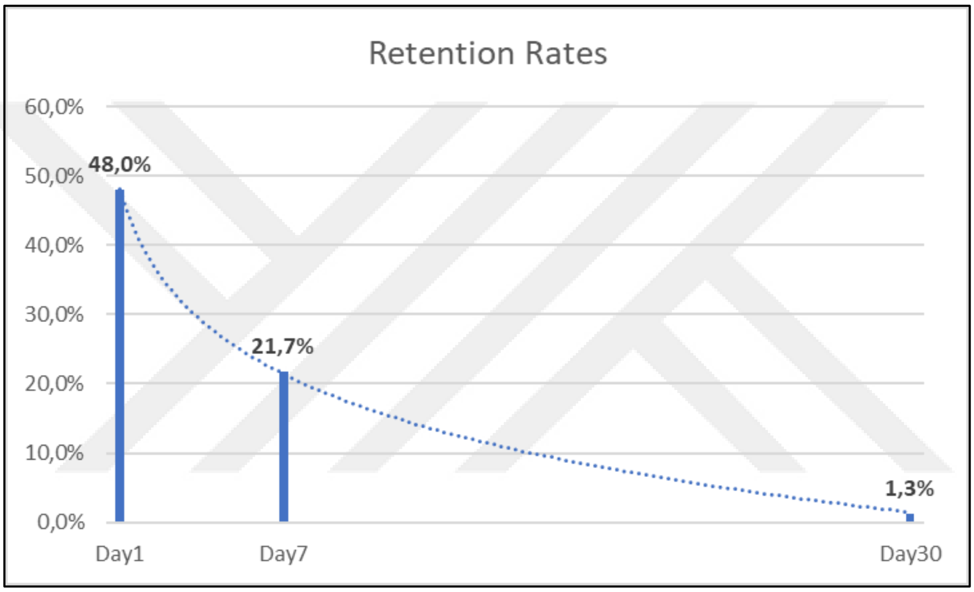
Figure 3.1: Retention rates in hyper-casual games over time



Source: GameAnalytics, 2019 report

The graphic from GameAnalytics above shows the retention rates for hyper- casual games. In the graph below, this study shows the retention rates of the game examined. The game examined in the study has the characteristic of being a typical hyper-casual game. In the first days after starting the game, the retention rate decreases rapidly and there is almost no user left at the end of a month. It is even one of the successful games according to the metrics given.

Figure 3.2: Retention rates in dataset used



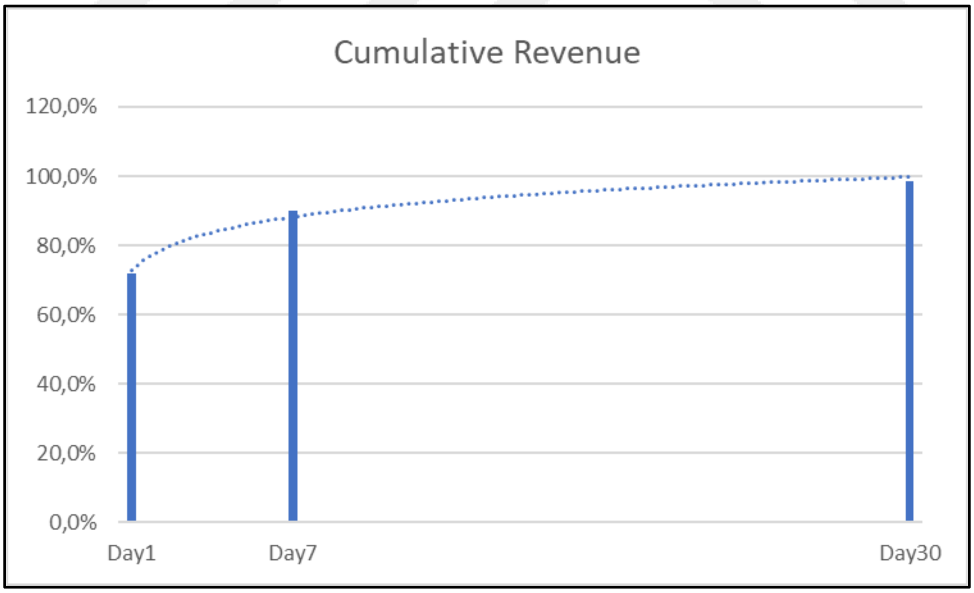
According to GameAnalytics data, as hyper-casual games are compared with other genes, day1 is among the best 5 genres according to retention rates. In this respect, it performs better than casual games. But when it comes to day7 retention rates, it drops to 10th (GameAnalytics, 2019). As we understand from here, users who play hyper-casual games tend to leave the game in a short time. This further increases the importance of churn and retention analysis on this genre.

Various descriptive analyzes have been made for the dataset to be used in the next part of the study. Before the modeling studies, this section was made to recognize the data and to see how the variables are positioned according to the target retention rates.

3.1.1 Revenue

There are three different types in the data as Revenue. All three of these are one of the sources of income for many companies in the gaming industry. These come from the banner, rewarded and interstitial categories, respectively. When we looked at dataset, we saw that the majority of the income from the game is coming from the interstitial category. Interstitial impressions are opened as images in full screen. These mostly occur in the natural course of the game. It can be confronted by the user when switching between levels or when the game is paused. Therefore, the income from here is more likely to be higher. Banner impressions are usually rectangular, on the right, left, above or below the screen while the game is playing or pausing. It is likely to receive fewer clicks compared to interstitial impressions. Rewarded impressions are followed by the promise of a certain response to users. This could be an extra life or extra time.

Figure 3.3: Cumulative revenue as time passes



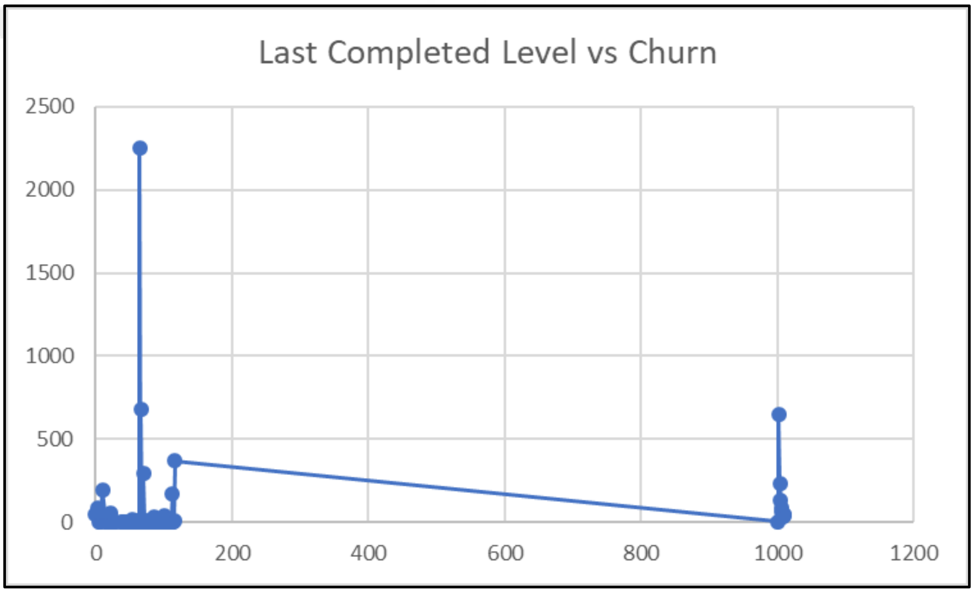
By looking at the graph above, the cumulative distribution of the income collected in the game can be monitored. The striking point here is that users participating in the game bring a large part of the income they will bring in the first few days. When it comes to

Day7 retention, approximately 90 percent of the total revenue is obtained. For this reason, hypercasual games are short-lived games that should constantly be included in new users.

3.1.2 Level

One of the reasons that causes churn in hypercasual games is level design. Factors such as the level being too difficult or too easy, taking a long time to complete may cause the user not to continue the game. Therefore, in this study, the last completed level variable within the scope of churn analysis will be included in the modeling.

Figure 3.4: Last completed level analysis



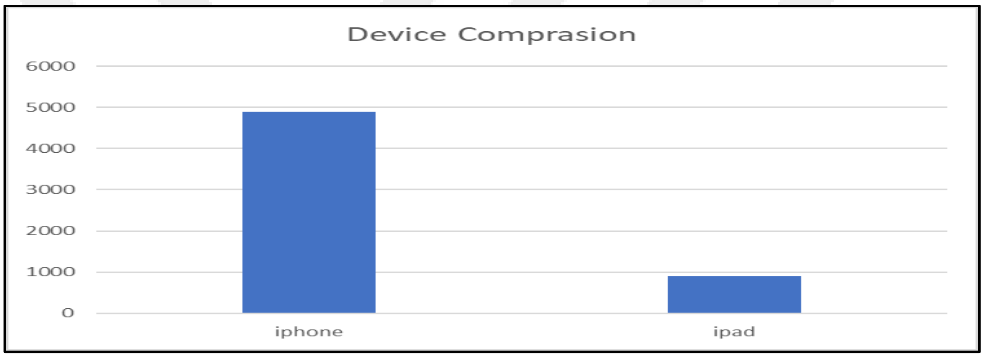
The graphic above shows which levels are completed and users churn after completion. Looking at some points, it is possible to see remarkable levels. There are high churn counts, especially up to level 117. Afterwards, there is no point that draws attention until the levels of 1000. Therefore, it is very important for users to reach 117 departments.

Firstly, 3.3 percent of the users quit the game after completing the 10th level. There is no such ratio until the 65th level. Therefore, it may be necessary to review level 11. The next big point is level 66. Because approximately 39 percent of those who completed the 65th level quit the game. This is a fairly high rate, indicating a problem with level 66. Probably this level is quite difficult and users cannot pass. Later, approximately 12 percent of those who completed the 67th level quit the game. This shows that level 68 needs to be re-examined.

3.1.3 Device

Nowadays, with the spread of tablets, games are designed for both smartphones and tablets. For this reason, it may be important to follow users in this way, too. The devices of the users who play the game are categorized as iPad and iPhone. When we look at this breakdown, approximately 85 percent of the users in the dataset used in the study play the game on the iPhone. The rest are progressing in the game with their iPads.

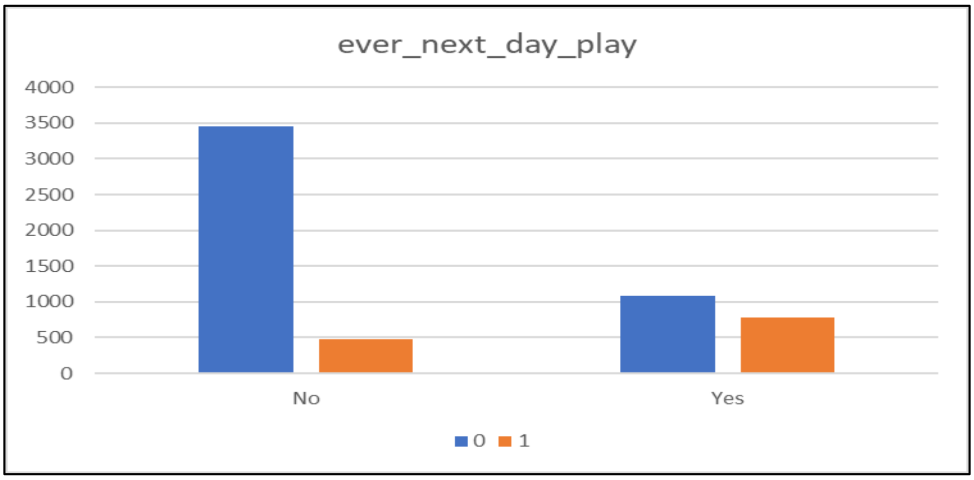
Figure 3.5: The used device comparison



3.1.4 Ever_Next_Day_Play Variable

If we look at the variable whether the users have ever arrived the next day or not, we can see the significant difference. The retention rates of users who tend to come at least once the next day are much higher than those who never come the next day. The game we will remove from here is to trigger the incoming user quickly and make them come to the next day at least once. In this way, it may be possible to talk about the user's loyalty to the game. In parallel with this, the variable count_active_days, which shows the minimum number of users arriving on different days, also validates this variable.

Figure 3.6: Distribution of the ever_next_day_play feature



3.2 MODELLING

3.2.1 Methods Used

After descriptive analysis, 7 different machine learning algorithms were applied for the data of the study. These are linear regression, logistic regression, support vector machine, k-nearest neighbor, decision tree, random forest and deep neural network, respectively. While choosing the models, it was aimed to look at the performances with different types of algorithms by choosing both regression, classification and deep learning techniques. In the next part of the study, the working methodologies of these models are explained briefly. To be able to use the mentioned methods, the dataset has been made suitable for R and python modeling. Afterwards, models were applied through appropriate libraries.

3.2.2 Linear Regression

Linear regression evaluates the linear relationship between two continuous variables to estimate the value of a dependent variable based on the value of an independent variable. While there is only one input feature in linear regression called simple, there is more than one feature in multiple linear regression (Freedman, 2009).

3.2.3 Logistic Regression

Logistic regression was first mentioned to literature in 1967 (Walker and Duncan, 1967). Briefly, logistic regression is a statistical method used to analyze a data set with one or more independent variables that determine a result. The result is measured by a binary variable (there are only two possible outcomes).

3.2.4 Support Vector Machine (SVM)

This method was used in literature both as classification and regression studies in 1995 (Cortes and Vapnik, 1995). Support Vector Machines (SVM) are the separation and classification of points on the plane by a straight or hyperplane. It is suitable for small or medium data sets. It is sensitive to scale. It must be scaled. Morevoer, it works on numerical variables.

3.2.5. K-Nearest Neighbor (KNN)

This algorithm can also be used for both classification and regression studies like SVM (Coomans and Massart, 1982; Altman, 1992). With KNN, basically, the closest points to the new point are searched. K represents the amount of the closest neighbors of the unknown point. The amount of k (usually an odd number) of the algorithm is chosen to estimate the results. K value should be chosen carefully to get good results. While small K values show less stability, higher K values can increase the error rate.

3.2.6 Decision Tree

Decision trees is a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes by feature and target. Decision tree algorithm is developed by breaking the data set into smaller or even smaller pieces (Quinlan, 1986).

3.2.7 Random Forest

It is an algorithm that aims to increase the classification value by generating more than one decision tree during the Random Forest classification process. Decision trees created individually come together to form a decision forest. The decision trees here are randomly selected subsets from the data set they are linked to (Ho, 1995).

3.2.8 Gaussian Naïve Bayes

This algorithm is basically based on Bayes' theorem and expressed as explained below (Friedman, Geiger, & Goldszmidt, 1997).

Figure 3.7: Bayes rule

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Definitions and rules in Bayes theory are as stated;

a. The events to be included in the probability calculation are shown with variables A and B.
b. P (A) shows the probability of the A event while P (B) shows the probability of the B event.
c. The probabilities are independent of each other and the probability of B variable can never be equal to zero.
d. When an event is known to occur, the probability of another happening is called conditional probability. For example, knowing B, the conditional probability of event A is expressed as P (A | B). Likewise, the conditional probability of event B when A is known is expressed as P (B | A).

3.2.9 Deep Neural Network

The deep neural network algorithms used in the study are multi-layered versions of artificial neural networks that are inspired by the information processing method of the human brain. Multi-layer perceptron models work with feed forward method and back propagation algorithm (Schmidhuber, 2015).

4. PROPOSED WORK

4.1 CHURN DEFINITION

The definition of churn is the time that varies from sector to sector, with the assumption that the user does not come back to the game. While there may be long periods in sectors such as telecom, banking, insurance, this concept is lower in the gaming world. In fact, care should be taken when evaluating the genres of the game within themselves. Especially in the hyper-casual game category where retention rates are low, low-time churn definitions should be made. Because in a short time, most of the players are lost and a large part of the retention is obtained.

Since the dataset used in the study was examined as revenue and retention as shown in the descriptive analysis, it was decided to define the churn as seven days. Because the game monetization reaches very low levels after the seventh day and the retention rate is very low after the seventh day.

4.2 FEATURES

Work has been done with the algorithms briefly mentioned above. As a result, metrics associated with algorithms were compared. The data used in the study contains information of 5790 users. Studies have been carried out to determine how these users churn after starting the game.

The data set in the study includes the following variables:

a. User_id (it shows the unique user id),
b. Banner_revenue (it shows the adversiment shown type 1),
c. Interstitial_revenue (it shows the adversiment shown type 2),
d. Rewarded_revenue (it shows the adversiment shown type 3),
e. Total_revenue (it shows the sum of first three revenue features),
f. Count_active_days (it shows the number of different days actively seen)
g. Ever_next_day_play (This is the variable that indicates whether the user played the game again the next day or not, 1: yes, 0: no)

h. Last_completed_level (it shows the last level the user has successfully completed),

i. Iphone_ipad (which device is used, 1: iPhone, 2: iPad),

j. Day7_retention (target variable: whether the user has churned for 7 days after starting the game, 0: churn, 1: keep playing)

Since most of the rapid churn rate and income in the study came on the 7th day, the definition of churn was made as 7 days. For the applied models, the train-test data is divided into 70:30 ratio. Since the user_id column does not make sense, and the day7_retention column is the target variable, the models are applied over 8 columns.

The lack of payment variables specific to variables is an obstacle. As can be seen in other studies in the literature, it has been revealed that purchasing for the game is an important factor over churn. For this reason, modeling algorithms without these variables were applied in this study.

5. RESULTS and DISCUSSIONS

Different models were made with various algorithms for the data used in the study. The results of these were also categorized within themselves and shared in the following part of the study.

5.1 PERFORMANCE MEASUREMENTS

Since the target variable for the models in the study is a binary value, confusion matrixes were created and accuracy values were examined. Also, the churn ratios in the dataset do not generate the imbalance dataset on the seventh day. Therefore, we can look at the accuracy rates. If we were modeling according to day30 retention rates, we would not be able to look at the accuracy rates, because the retention rate of the data is very low, so the accuracy values would always be high (Visa and Ralescu, 2005). For this reason, comparisons of models can be made by looking at accuracy rates in current churn models.

The model can be evaluated according to the table and formula below. Accuracy value ranges from 0 to 1. The closer to 1 indicates the higher the accuracy.

Figure 5.1: Confusion matrix table

Real churn	Predicted churn	
	1	0
1	True Positive (TP)	False Negative (FN)
0	False Positive (FP)	True Negative (TN)

Calculations are made with 4 outputs obtained from the table above. The meanings of these 4 outputs are briefly defined;

True Positives

It is a definition showing the class that is estimated to be 1 as a binary value and actually has a value of 1.

True Negatives

It is a definition that shows the class that is estimated to be 0 as a binary value and actually has a value of 0.

False Positives

It is the definition that shows the class that is predicted to be 1 but has an actual value of 0. It is also expressed as 'type 1 error' in statistical terms.

False Negatives

It is the definition that shows the class that is predicted to be 0 but has an actual value of 1. It is also expressed as 'type 2 error' in statistical terms.

Calculation of the formula according to the confusion matrix and above definitions shown above is as follows;

Figure 5.2. Accuracy formula

$$\frac{TP+TN}{TP+FP+TN+FN}$$

5.2. REGRESSION ALGORITHMS

Dataset was controlled according to the R-Square value by linear regression technique. R-Square analytically shows the closeness of fitted to the regression line. This value was calculated as percent 45.7. When we look at it as a threshold, it is an unacceptable value. However, this is a normal result because the dataset used in the study is not suitable for linear regression.

5.3 CLASSIFICATION ALGORTIHMS

As classification methods, 6 different models were studied. Accuracy rates for these models were calculated and compared with each other. While these models were established, day7 retention figures were used. The day7 churn rate used in our study is 78.3%. Since the churn rate on the seventh day is not very high, the accuracy rate gives meaningful results.

When we compared the results of the models, the SVM model remained inferior to the others. In the other 4 models, accuracy rates vary between 84.1% and 88.7%.

Random forest model gave 88.7% accuracy, decision tree gave 88.6%, k-nearest neighbor gave 87.6%, logistic regression gave 88.1%, support vector machine gave 77.7% and Gaussian Naïve Bayes gave 84.1%. If we look at them as variables,

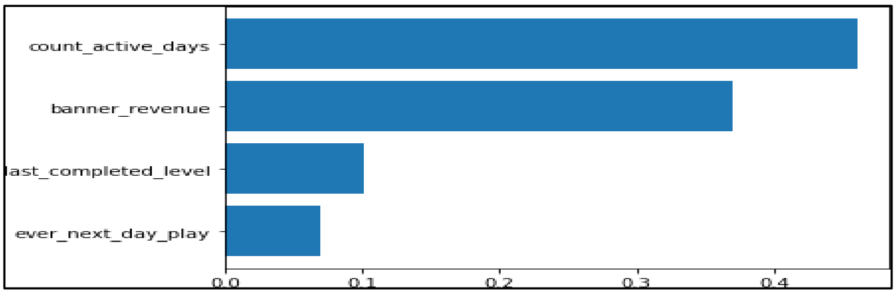
count_active_days, banner_revenue and last_completed_level come as the 3 most important variables. count_active_days came as the most important variable in all models.

Figure 5.3: Model Results

Used Algorithm	Accuracy (%)
Logistic Regression	88,1
Decision Tree	88,6
Random Forest	88,7
Support Vector Machine (SVM)	77,7
K-Nearest Neighbor (KNN)	87,6
Gaussian Naive Bayes	84,1

The model with the highest accuracy was the random forest algorithm. The graph showing the importance level of the variables taken from this model is shared below. The fourth important variable for this model is ever_next_day_play. However, this variable remained below threshold values indicating its degree of significance. Also, there is moderate correlation between count_active_days and ever_next_day_play.

Figure 5.4. Important feature from randomforest model



For the table mentioned in the Confusion Matrix section, the confusion matrix obtained from the random forest model is as follows;

Figure 5.5: Confusion matrix from randomforest model

```
In [23]: print(confusion_matrix(y_test,y_pred))
[[1292   81]
 [ 113  251]]
```

5.4 DNN

Deep neural network with deep machine learning on dataset has also been implemented. Accuracy ratio was calculated as percent 78.7. Although it is a relatively acceptable level, it fell below the classification algorithms. This model also confirmed the important variables that are coming in the classification algorithms.

5.5 COMPRASION of SIMILAR WORKS

According to the literature review made in this study, the results and sectors of the study using similar machine learning algorithms used in the study are given in the table below.

Table 5.1: The AUC values of similar Works

Sector	Logistic Regression	Decision Tree	Support Vector Machine	Randomforest	Naive Bayes	KNN
Casual Game	81,4	73,2	76,2			
Casual Game	92,4	85	90,3			
Casual Game	78,6			78,7		
Casual Game	71,1			72,2		
Casual Game	82,4			82,2		
Casual Game	66	59	61	75		
Casual Game				84		
Casual Game		90,4		93,7		
Online Game	71,1		73,9	76,1		
Online Game		94,8		98		
Online Game	68,2	59,5	71,7	73,1		
Online Game				93,6		
Online Game				73,1		
Online Game	79	67	80		73	
Telecom	76,7					
Telecom	82,9		79,7	81,3		
Telecom	85,6		94	94,9	86,5	90,4
Banking		78,3	79,6	83,6		87,6
Banking	58,9		59,7		55,5	

Only modeling studies including algorithms used in this study are shown in the table. Studies carried out in addition to these algorithms are also examined in detail in the literature review section. As can be seen from the table, the most frequently used among the 7 models are logisctic regression, decision tree, support vector machine and randomforest algorithms. Naive Bayes and KNN algorithms are included in this study, although similar studies are not frequently preferred. Especially in the literature, studies using the KNN algorithm belonging to the game industry were not found. However, it can be seen in the studies in the literature that it gives good results in churn studies for telecom and banks. For this reason, KNN was also included in the churn models made for the hyper-casual game used in this study and good results were obtained.

In addition to linear regression, all 6 machine learning algorithms shown in the table in this thesis were used and their results were evaluated. Only 3 models in this study gave AUC values above 88. These are 88.7 for the random forest, 88.6 for the decision tree, and 88.1 for the logistic regression, respectively. The KNN algorithm gave a value of 87.6 AUC as the 4th best model. While Naive Bayes gave 84.1 AUC, the lowest AUC value was 77.7 by the Support Vector Machine algorithm.

Although the KNN algorithm was used in only 2 similar studies, the AUC values were quite high, which was the reason why it was tested in this study. Although the KNN AUC value in this study was not the best result, it came out at a satisfactory level with 84.6. Therefore, its use in similar studies may be beneficial in terms of model diversity. As can be seen from the table above, the random forest model has been the algorithm that gives the best AUC value in many studies. It has given the best AUC value in many studies in churn models made in both gaming and other sectors. Although it is slightly better than the decision tree algorithm for the hyper-casual game used in this study, random forest was the best model. The conclusion we can draw from this is that the algorithm that gives the best AUC value for casual games, online games, telecom and banking sectors in the literature also gave the best result for a game in the hyper-casual category.

When we examine it as the game industry, Naive Bayes and KNN algorithms have not been a preferred method in churn model studies for both casual games and online games. When we look at the studies in the literature, these two models are preferred in the telecom and banking sectors. However, as can be seen in this study, good AUC values were obtained for a game in the hyper-casual category. In particular, the KNN algorithm contributed to the model studies by giving a value of approximately 88 AUC. Moreover, while the Support vector machine gave AUC values of around 90 in some studies, it gave the lowest AUC value of 77.7 in this study.

When we look at the literature, in addition to the model results in online games, many variables have been examined. In particular, factors such as competition and event setting seem to be important factors in increasing user loyalty in online games. It is possible to see the level and game dynamics of casual games as in the hyper-casual review, which we examined in this context. In this study, detailed information about this is included in the recommendations section.

6. RECOMMENDATIONS

In the first parts of the study, it was mentioned that hyper-casual games have a short life and high churn rates. For this reason, the definition of churn, which is usually chosen for 14 days in casual games, was chosen as 7 in this study. Because at the end of the 7th day, a large part of the user was churning and a large part of the revenue to be obtained was earned. Analytical models were applied according to the churn definition. In the study, linear regression had very low accuracy, while other algorithms gave good results, while SVM and DNN remained below 80%, while other algorithms gave higher results.
Retention and revenue distributions specific to variables were laid out in the descriptive analysis sections of the study. Important variables coming from the model result also support these inferences. Therefore, it is possible to evaluate the results in three different ways. First of all, getting the users who started the game back to the game somehow gives a very positive effect. An option may be to send notifications to users with reminders about the game. In addition, opportunities to return the user to the game or extra lives can be good options. Apart from that, it would be good to review the levels again. Because some levels remarkably caused users to quit the game. Inferring that these levels are difficult and redesigning will enable users to continue the game. It may be useful to take a special look at the levels mentioned in the descriptive analysis section. In addition to these, we see that the users who stay in the game generate income through the banner ad. This is actually a part that affects income, slightly different from the first 2 variables. Since users who continue the game are clicks on banner impression, paying attention to these will be a factor that increases revenue. The point that draws attention here is about rewarded advertisements. Before the study, we could expect rewarded impressions to be more important. But the revenue from here is very low. This may be because the rewarding ads are not being properly delivered. Because we can expect users to be more inclined to rewarded ads. For this reason, it will be better for users to review rewarding ads and test them in different ways.

With the studies to be done in this way, it is possible to reduce both churn and increase the income. Since the gains of hyper-casual games are in the first days, it is very important for new users to join the game. One of the parts of user acquisition in the gaming industry is the concept of virality. The more we can keep the players in the game in, the more chance we have to see the positive effect of virality.

7. FUTURE WORK and LIMITAIONS

The biggest restriction was related to the data set provided. If the dataset of different periods could be obtained, the comparison of the periods could also be made. In addition, it would be good to add some variables to make the modeling and analysis in more detail. For example, the expenditures made for the game, or the time spent in the game when active, would be effective in order to make detailed inferences. According to the studies in the literature, these variables may be important for the churn concept. It can be provided to support working with extra features in future studies.

8. CONCLUSION

In today's world, which is formed by the proliferation of digitalized and smart devices, a new type of mobile games, the hyper-casual category, has emerged. This is a genre that is similar to casual games but has its own dynamics in terms of game dynamics. Its main characteristics are high churn rates and short life span. The rapid churn of the users increases the importance of the studies on this subject. In this study, the churn and retention rates of a hyper-casual game that can be considered successful according to the retention rates were examined. Before starting the analysis and modeling, a literature review was made on this subject. In the literature, similar studies in the gaming industry and similar studies in other sectors have been examined. As a result of this review, models applied in other studies were also implemented in this study. Later, a churn definition was made in accordance with the dataset in this study. While doing this, both retention rates and income distribution were taken into consideration. After deciding on the definition of churn, the reasons causing churn have been examined with various analytical models. Models developed in this context were compared and important variables were determined. While the applied models provide results that support each other, accuracy rates are calculated high. It has been observed that the determined variables and level design and difficulty levels are important, the user needs to be triggered quickly after starting the game, and the advertisements shown affect the user's continuing the game. For this reason, strategies should be reviewed according to the results obtained, and this should be done in dynamic and frequent periods. Because it is seen again with the dataset used in this study that the life of the users who come to hyper- casual games is very short. While there is a large proportion of user retention rates on the seventh day, by the thirtieth day there is almost no user left. In studies, these should be taken into consideration and churn and retention rates should be done regularly in frequent periods. Because the life span is short, it is possible that users coming in different periods will have different tendencies. The models applied in this study and the variables examined were investigated in the literature and similar studies were examined. This study has been conducted for the hyper-casual genre based on churn and retention studies in both casual games and online games. The lack of a similar study in this field is the most important contribution of this study to literature.

REFERENCES

Books

Cortes, C. and Vapnik, V., 1995. Support-vector networks. Machine learning, 20, pp. 273-297.

Coomans, D., & Massart, D. L., 1982. Alternative k-nearest neighbor rules in supervised pattern recognition: Part 1. k-Nearest neighbor classification by using alternative voting rules. Analytica Chimica Acta, 136, pp. 15-27.

Periodicals

Bergstrom, K., 2018. Moving Beyond Churn: Barriers and Constraints to Playing a Social Network Game. Games and Culture 2019, Vol. 14(2), pp. 170-189

Bertens, P., Guitart, A. & Perianez, A., 2017. Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model

Bolance, C., Guillen, M. & Padilla-Barreto, A.E., 2016. Predicting Probability of Customer Churn in Insurance. pp. 82-89.

Borbora, Z.H. & Srivastava, J., 2012. User Behavior Modelling Approach for Churn Prediction in Online Games, 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, pp. 51-60.

Castro, E.G. & Tsuzuki, M.S.G., 2015. Churn Prediction in Online Games Using Players' Login Records: A Frequency Analysis Approach. IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, VOL. 7, NO. 3, pp.255-264

Chesham, A., Wyss, P., Müri, R.M., Mosimann, U.P. & Nef, T., What Older People Like to Play: Genre Preferences and Acceptance of Casual Games. JMIR Serious Games 2017

| vol. 5 | iss. 2 | e8 | p. 1

Freedman, D.A., 2009. Statistical models: theory and practice. cambridge university press.pp. 26.

Friedman, N., Geiger, D. and Goldszmidt, M.,1997. Bayesian network classifiers. Machine learning, 29, pp.131-163.

Gaur, A. & Dubey, R., 2018. PREDICTING CUSTOMER CHURN PREDICTION IN TELECOM SECTOR USING VARIOUS MACHINE LEARNING TECHNIQUES

Hadiji, F., Sifa, R., Drachen, A., Thurau, C., Kersting, K. & Bauckhage, C., 2014. Predicting Player Chum In the Wild

Hargreaves, C.A., 2019. A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry, International Journal of Future Computer and Communication, DOI: 10.18178/ijfcc.2019.8.4.550

Harrison, B. & Roberts, D.L., Analytics-Driven Dynamic Game Adaption for Player Retention in Scrabble

Harrison, B. & Roberts, D.L., 2015. An Analytic and Psychometric Evaluation of Dynamic Game Adaption for Increasing Session-Level Retention in Casual Games. IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, VOL. 7, NO. 3, SEPTEMBER 2015

Harrison, B. & Roberts, D.L., Analytics-Driven Dynamic Game Adaption for Player Retention in Scrabble

Harrison, B. & Roberts, D.L., 2015. An Analytic and Psychometric Evaluation of Dynamic Game Adaption for Increasing Session-Level Retention in Casual Games. IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES, VOL. 7, NO. 3, SEPTEMBER 2015

Hargreaves, C.A., 2019. A Machine Learning Algorithm for Churn Reduction & Revenue Maximization: An Application in the Telecommunication Industry, International Journal of Future Computer and Communication, DOI: 10.18178/ijfcc.2019.8.4.550

He, Y., Xiong, Y. & Tsai, Y., 2020. Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company

Ho, T. K. 1995. Random decision forests. International conference on document analysis and recognition, 1, pp. 278-282. IEEE, Montreal, Quebec, Canada. Paper. Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, pp 67–73.

Huigevoort, C., 2015. Customer churn prediction for an insurance company. Thesis for master degree. Eindhoven: TUE

Jaideep, Z.B., Hsu, K. & Wiliams, D., 2012. Churn Prediction in MMORPGs using player motivation theories and an ensemble approach

Jeon, J., Yoon, D., Yang, S. & Kim, K., 2017. Extracting Gamers' Cognitive Psychological Features and Improving Performance of Churn Prediction from Mobile Games. IEEEConference on Computational Intelligence and Games 2017, pp. 150-153

Jiang, X. & Safaei, F., 2011. Churn Performance Study of Structured Peer-to-Peer Overlay in Supporting Massively Multiplayer Online Role Playing Games (MMORPGs). pp.

183-188.

Jin, Z. & Xing-hua, D., 2008. Bank Customer Churn Prediction Based on Support Vector Machine: Taking a Commercial Bank's VIP Customer Churn as the Example

Karvana K.G.M, Yazid, S., Syalim, A. & Mursanto, P., Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. 2019 International Workshop on Big Data and Information Security (IWBIS)

Kilimci, Z.H., Yörük, H. & Akyokuş, S., 2020. Sentiment Analysis Based Churn Prediction in Mobile Games using Word Embedding Models and Deep Learning Algorithms.

Kim, S., Choi, D., Lee, E. & Rhee, W., 2017. Churn prediction of mobile and online casual games using play log data. PLoS ONE 12(7): e0180735. https://doi.org/10.1371/journal. pone.0180735

Kim, K.H. & Kim, H.K., 2019. Oldie is Goodie: Effective User Retention by In-game Promotion Event Analysis

Kristensen, J.T. & Burelli, P., 2019. Combining Sequential and Aggregated Data for Churn Prediction in Casual Freemium Games

Lalwani, P., Mishra, M.K., Chadha, J.S. & Sethi P., 2020. Customer churn prediction system: a machine learning approach, Mathematics Subject Classification 68T01 · 68T05

Lee, E., Kim, B., Kang, S., Kang, B., Jang, Y. & Kim, H.K., 2018. Profit Optimizing Churn Prediction for Long-Term Loyal Customers in Online Games

Liu, X., Xie, M., Wen, X., Chen, R., Ge, Y., Duffield, N. & Wang, N., 2018. A Semi-Supervised and Inductive Embedding Model for Churn Prediction of Large-Scale Mobile Games

Milosevic, M., Zivic, N. & Andjelkovic, I., 2017. Early churn prediction with personalized targeting in mobile social games. pp. 326-332.

Morik, K. & Köpcke, H., 2004. Analysing Customer Churn in Insurance Data – A Case Study, pp. 325–336.

Quinlan, J. R., 1986. Induction of decision trees. Machine learning, 1(1), pp. S1-S106.

Papaloukas, S., Stoli, C., Patriarches, K. & Xenos, M., 2020. A Survey on How Online Games Usability Affects the Retention of Information. A Survey on How Online Games Usability Affects the Retention of Information, pp. 209-212.

Perianes, A., Saas, A., Guitart, A. & Magne, C., 2017. Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles

Pirker, J., 2018. Social Network Analysis in Games User Research In Games User Research. Oxford University Press. https://doi.org/10.1093/oso/9780198794844.003.0029, pp. 491-499.

Popović, D., 2008. Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm, pp.243-247.

Raeisi, S. & Sajedi, H., 2020. E-Commerce Customer Churn Prediction By Gradient BoostedTrees. 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)

Rahman, M. & Kumar, V., 2020. MACHINE LEARNING BASED CUSTOMER CHURN

PREDICTION IN BANKING, Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020), pp. 1196-1200.

Rothmeier, K., Pflanzl, N., Hüllman, J.A. & Preuss, M., 2018. Prediction of Player Churn andDisengagement Based on User Activity Data of a Freemium Online Strategy Game

Runge, J., Gao, P., Garcin, F. & Faltings, B., 2014. Churn Prediction for High-Value Playersin Casual Social Games

Sabbeh, S.F., 2018. Machine-Learning Techniques for Customer Retention: A Comparative Study, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 2, 2018, pp. 273-279.

Santharam, A. & Krishnan, S.B., 2018. SURVEY ON CUSTOMER CHURN PREDICTION

TECHNIQUES. International Research Journal of Engineering and Technology (IRJET)

Saran Kumar, A. & Chandrakala, D, 2016. A Survey on Customer Churn Prediction using Machine Learning Technique International Journal of Computer Applications (0975 –8887)

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.

ULLAH, I., RAZA, B., Malik, A. K., IMRAN, M., ISLAM, S. & KIM, S.W., 2019. A Churn

Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, pp. 60134-60148.

Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. & Chatzisavvas, K.Ch., 2015. A comparison of machine learning techniques for customer churn prediction, Simulation Modelling Practice and Theory, pp. 1-9.

Viljanen, M., Airola, A., Pahikkala, T. & Heikkonen, J., Modelling User Retention in Mobile Games

Visa, S. & Ralescu, A., 2005. Issues in Mining Imbalanced Data Sets - A Review

Walker, S.H. & Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. Biometrika, 54, pp. 167-179.

Yang, W., Huang, T., Zeng, J., Yang, G., Cai, J., Chen, L., Mishra, S. & Liu, Y., 2019. Mining Player In-game Time Spending Regularity for Churn Prediction in Free Online Games

Yang, W., Huang, T., Zeng, J., Chen, L., Mishra, S. & Liu, Y., 2019. Utilizing Players' Playtime Records for Churn Prediction: Mining Playtime Regularity

Yang, W., Huang, T., Zeng, J., Yang, G., Cai, J., Chen, L., Mishra, S. & Liu, Y., 2019. Mining Player In-game Time Spending Regularity for Churn Prediction in Free Online Games

Yang, W., Huang, T., Zeng, J., Chen, L., Mishra, S. & Liu, Y., 2019. Utilizing Players' Playtime Records for Churn Prediction: Mining Playtime Regularity

Yayun, Z., 2018. Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. Management Science and Engineering Vol.12, No. 3, 2018, pp. 51-56

Yu, X., Guo, S., Guo, J. & Huang, X., 2011. An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Applications, pp.1425-1430.

Zhao, X., 2014. Research on E-Commerce Customer Churning Modeling and Prediction. The Open Cybernetics & Systemics Journal, 2014, 8, pp. 800-804.

Other

GameAnalytics, 2019. Hyper-Casual Gaming: The Latest Stats and Trends You Need to Know. https://gameanalytics.com/blog/hyper-casual-gaming-stats/

GameAnalytics, 2020. 4 Key Steps to Making a Successful Hyper-Casual game. https://gameanalytics.com/blog/4-key-steps-to-making-a-successful-hyper-casual- game/

Padte, R., 2019. Impact of Textual Feedback on Player Retention in Endless Runner Mobile Games. Thesis for master degree. Boston: Northeastern University.

Statista, 2021. Mobile gaming market in the U.S. - statistics & facts. https://www.statista.com/topics/1906/mobile-gaming/

Wikipedia, 2020. https://en.wikipedia.org/wiki/Hyper-casual_game#cite_note-:0-3

Wang, Q., 2018. In Partial Fulfillment of the Requirements for the Degree of Master of Science in Game Science and Design, Thesis for master degree. Boston: NortheasternUniversity