

DOI: 10.29141/2218-5003-2022-13-6-6

EDN: KBZBMO

JEL Classification: C53, D18, M21

## The study of indicators affecting customer churn in MMORPG games with machine learning models

Kaan Arik<sup>1</sup>, Murat Gezer<sup>2</sup>, Seda Tolun Tayali<sup>2</sup><sup>1</sup> Sakarya Applied Science University, Sakarya, Turkey<sup>2</sup> Istanbul University, Istanbul, Turkey

**Abstract.** Over the past two decades, the gaming industry has rapidly increased its popularity and gained a top spot in the entertainment sector. With this rise, customer relations and churn analysis have become even more prominent in gaming, as digital games are quickly integrated into the industry and fetch high revenues. The esteem for gaming and the revenue of companies has increased, which has made the concept of customer churn more critical. The behavior of most customers in the gaming industry also shows that it is a structure worth analyzing. This study focuses on the player log data of Blade and Soul game over specific periods with machine learning algorithms and tries to answer two business problems. These problems are to predict churners and player survival time, which are modeled as a classification and a regression problem, respectively. Blade and Soul is an MMORPG (massively multiplayer online role-playing game) developed by NCSOFT and widely played in the Far East. The theoretical basis of the study is the provisions of behavioral economics and churn management. For research purposes, the methods and algorithms of machine learning were used. Player log data was collected in 2016 and 2017, then anonymized and released for researchers. The research scope follows the CRISP-DM methodology and explains the process in detail by adhering to this methodology. A set of data consisting of 10,000 players with Test 1, Test 2, and Train were released separately by NCSOFT. Considering player churn and survival times, XGBoost demonstrates the highest effectiveness for classification problem and MLP and GBR for regression problem. The results show churn analysis can help businesses identify trends and patterns in customer behavior, such as when customers are most likely to leave, which features are causing them to leave, and which customer segments are most at risk of churning identifying trends and patterns.

**Keywords:** churn prediction; business intelligence; digital games; machine learning; artificial intelligence.

**Acknowledgements.** We would like to thank NCSOFT, which contributed to academic studies by sharing Blade and Soul data we used throughout the research period.

**Article info:** received August 28, 2022; received in revised form October 5, 2022; accepted October 11, 2022

**For citation:** Arik K., Gezer M., Tolun Tayali S. (2022). The study of indicators affecting customer churn in MMORPG games with machine learning models. *Upravlenets / The Manager*, vol. 13, no. 6, pp. 70–85. DOI: 10.29141/2218-5003-2022-13-6-6. EDN: KBZBMO.

## Выявление показателей оттока пользователей MMORPG-игр с применением моделей машинного обучения

К. Арик<sup>1</sup>, М. Гезер<sup>2</sup>, С. Толун Тайали<sup>2</sup><sup>1</sup> Университет прикладных наук Сакарья, г. Сакарья, Турция<sup>2</sup> Стамбульский университет, г. Стамбул, Турция

**Аннотация.** В последние годы произошел существенный рост популярности игровой индустрии, занявшей лидирующие позиции в сфере досуга и развлечений. Способность онлайн-игр быстро генерировать высокие доходы обусловила научный и практический интерес к взаимодействию с потребителями в этой сфере и выявлению причин их оттока. Исследование направлено на выявление наиболее эффективных алгоритмов машинного обучения по определению потенциальной готовности пользователей покинуть онлайн-игру (проблема классификации) и вероятной продолжительности их участия (проблема регрессии). Методологическая основа исследования представлена положениями поведенческой экономики. Для решения поставленных задач использовались методы и алгоритмы машинного обучения. Информационную базу составили анонимные игровые лог-данные 10 тысяч пользователей MMORPG-игры Blade and Soul. Согласно полученным результатам, для анализа оттока пользователей наиболее эффективен алгоритм XGBoost, а для анализа продолжительности их участия в игре – алгоритмы MLP и GBR. Наибольшую значимость для прогнозирования потенциального оттока пользователей продемонстрировали такие показатели, как продолжительность игровой сессии и жизни персонажа, частота появления пользователя в игре и др. Полученные результаты могут быть полезны компаниям – разработчикам игр при выявлении трендов и паттернов поведения пользователей, их готовности прекратить участие в игре и причин данного решения, определении сегментов игроков, наиболее подверженных риску оттока.

**Ключевые слова:** прогнозирование оттока пользователей онлайн-игр; бизнес-аналитика; онлайн-игра; машинное обучение; искусственный интеллект.

**Благодарности:** Авторы выражают благодарность компании NCSOFT за предоставление доступа к данным онлайн-игры Blade and Soul на весь период проведения исследования.

**Информация о статье:** поступила 28 августа 2022 г.; доработана 5 октября 2022 г.; одобрена 11 октября 2022 г.

**Ссылка для цитирования:** Arik K., Gezer M., Tolun Tayali S. (2022). Predictive models for customer churn analysis in MMORPG games // *Управленец*. Т. 13, № 6. С. 70–85. DOI: 10.29141/2218-5003-2022-13-6-6. EDN: KBZBMO.

## INTRODUCTION

Today's world has come under the influence of developing information and communication technologies. Data is the new oil and is increasing in size progressively. This increase has led to the emergence of the concept of 'Big Data' [Buhl et al., 2013; Sagiroglu, Sinanc, 2013; Schermann et al., 2014]. Big data's contribution to information and communication technologies has become an essential concept for all sectors. As a result of the increasing importance of technology along with the opportunities it offers, information has become an important resource for enterprises regarded as an intangible asset. [Davenport, Barth, Bean, 2012; George, Haas, Pentland, 2014; Markus, 2015; Woerner, Wixom, 2015]. Since the basic information structure is data, businesses have been storing all types of data. Therefore, businesses should retain their existing customers and gain new customers under tough competitive market conditions. In this context, customer data have become of much more particular importance.

Customers have always been the key aspect for companies. The customer is a company's source of income, and loyalty is paramount. So, what is a customer? A customer is a person or organization that purchases a particular brand of property of a particular business for administrative or personal purposes [Gold, Tzu, 2020; Beyer, Holtzblatt, 1995; Meyer, Schwager, 2007]. It also helps to learn how satisfied a company's current and potential customers are with the company's and its competitor's products and services and if there are any issues they are not satisfied with. The presentation of products to users on digital platforms has expanded the volume of e-commerce, and a revolutionary increase has been achieved in product sales. With the increased number of products offered on the Internet and the level of competition, Customer Relationship Management (CRM) has become even more critical. It is crucial that the concept of the customer is at the heart of these developments and that they develop accordingly. As the concept of customer has become more important, applications for customer retention have started to be developed in different sectors, and existing applications continue to emerge. Especially in the gaming industry, the change in marketing methods in recent years has had a significant impact, and this development has also brought the concept of business analytics in gaming to the forefront. Retaining existing customers helps companies maintain total revenue and support growth. For this reason, customer churn analysis has become an essential structure in all monetized sectors.

The study aims to classify regression predictions for customer churn with machine learning algorithms. In the rapidly developing customer management domain, MMO (massively multiplayer online game) data has been used to predict who tends to leave. Also, it has focused on achieving performance results of different machine learning algorithms and contributing to customer churn analysis in gaming. The rest of the paper is organized as follows. Theoretical background mentions the research on the topic of the study and related areas. Next, Method section describes the data collection, dataset, preprocessing, and model development for game churn prediction. Then, Results and Discussion section discusses our solution in detail. Finally, Conclusion section presents our experimental results on data and summarizes our work.

## THEORETICAL BACKGROUND

Companies try to predict the future using historical data to retain current customers. These predictions are mainly based on descriptive data. One of the most emphasized aspects of customer relationship management is churn management, which has become an essential issue for businesses [KhakAbi, Gholamian, Namvar, 2010; Lejeune, 2001; Singh, Samalia, 2014]. Churn management is a critical analysis to prevent companies from losing their existing customers and bearing additional costs. Churn analysis offers performance income and has become a big part of many common sectors, such as finance, insurance, and telecommunication. Also, it allows a company to create various campaigns to keep those customers from leaving by anticipating customers who are likely to quit working with the company. Analyzing churn rate allows seeing more than just a set of percentiles. This will also allow understanding that customers tend to cancel the contract and take preventive action before they leave the company. In this way, a company can increase the proportion of loyal customers by preventing possible losses [Almana, Aksoy, Alzahrani, 2014; Çelik, Osmanoğlu, 2019; Dahiya, Bhatia, 2015; Vyas et al., 2018; Payne, 2008]. In particular, companies predict loss rates by analyzing data on customer metrics, named for customer churn analysis. It helps to identify metrics such as customer retention, churn rate, revenue reacquisition, and customer lifetime value [Lazarov, Capota, 2007]. Thus, CRM departments can contact probable churners and make new offers to keep them in their portfolio. Another important factor why churn management is critical is that companies' cost incurred in acquiring a new customer is nearly five times

the cost they spend on current customers. Again, it is worth noting that this information is known, especially in the finance and telecommunication sectors today [Gold, Tzu, 2020].

The digital industry is growing extremely fast, and its employees are expected to specialize in their interests. One of the prominent and prevalent structures of the digital industry in recent years has been games. The gaming industry has started to create a significant employment income worldwide, and interestingly, it has begun to be called the 'Gaming Industry'. However, even the name game industry has remained so general that it has been divided into different categories, such as games played on social networks and console games. Games are offered to gamers on different platforms and genres, categorized based on camera angles and platforms. Some of them need an internet connection while others do not. One of these game genres is the MMO, which is also within the scope of our research. MMO is an online game in which perpetual interaction is at a high level, and progress is achieved by performing specific missions in-game [Adams, Dormans, 2012; Ijsselstein et al., 2007; Schell, 2019]. Therefore, the function of being a community member, where mutual tasks are usually completed, is generally more common in online games. Especially in MMOs, since players can communicate via chat-box, interaction remains continuous and high-level [Crawford, 2003; Duke, 1980]. In order to upgrade in-game, players usually need to devote most of their time and perform necessary missions and tasks. MMOs have been one of the most played genres in the gaming industry in recent years.

Churn analysis in gaming has become a popular research area recently, and models have been developed on different datasets with various types of games. Table 1 shows related works on customer churn analysis for gaming in the literature.

According to Table 1, customer churn analysis has been performed in MMORPG (massively multiplayer on-

line role-playing game), MOBA (multiplayer online battle arena) and online social game genres. While most of these studies are mainly classification, regression and clustering analysis solutions have been produced. Furthermore, due to the unique structure of each game data, different models have been tried and high accuracy ratios have been achieved.

It is important to get high-performance metrics, especially in an MMORPG game that is difficult to predict. In 2011, Borbora, Srivastava, Hsu, and Williams conducted game churn prediction in MMORPGs using theories of player motivation and a community approach [Borbora et al., 2011]. Different approaches were implemented in customer churn prediction in the EverQuest II game, and a high level of success was achieved [Borbora et al., 2011]. Another study predicts player behavior during customer churn in MMORPGs [Borbora, Srivastava, 2012]. Researchers also carried out on other MMORPG game data and similar results were achieved [Ding, Gao, Chen, 2015]. The best final score was 0.61 in the study [Lee et al., 2019] and the 0.72 final score was mentioned by Kummer, Nievola, and Paraiso [Kummer et al., 2018]. However, it's seen that final scores found in studies do not up to 0.72 for Blade and Soul data.

**Churn analysis and gaming.** The gaming industry is a \$100 billion industry [Sydow, 2020], and it is growing at an exponential rate. Churn analysis helps game studios understand where they are losing their players and is mentioned as "Game Churn Prediction" in the gaming industry. It can help them identify the reasons for this loss, and work on strategies to improve the player experience, thus reducing churn rates. Given market and business intelligence issues associated with game development, primarily online computer games, one of the most critical factors for success is the ability to detect and identify the subscribers of a system (for example, players of a game) that will soon leave. Typically, the churn rate is measured in months, quarters, or years, depending on the industry

Table 1 – Related works on online game churn analysis

Таблица 1 – Смежные исследования по тематике оттока пользователей в игровой индустрии

| Authors, year          | Platform-game   | Problem                         |
|------------------------|---|---------------------------------|
| Borbora et al., 2011   | MMORPG-EverQuest II   | Classification                  |
| Long et al., 2012      | Social Media Platform – Pengyou                               | Clustering                      |
| Hadiji et al., 2014    | MMORPG-Blade and Soul   | Classification-regression       |
| Tamassia et al., 2016  | MOBA  | A hidden Markov model approach  |
| Castro, Tsuzuki, 2015  | Online games – RF Online, APB: Reloaded, Heavy Metal Machines | Classification and RFM analysis |
| Drachen et al., 2016   | Online FPS – Destiny  | Cluster analysis                |
| Lee et al., 2019       | MMORPG – Blade and Soul                                       | Classification                  |
| Zheng et al., 2020     | Online MMO – Ghost II, Tianxia 3, and Fantasy Westward        | Classification                  |
| Rothmeier et al., 2021 | Online Strategy – The Settler Online                          | Classification                  |
| Zhao et al., 2022      | Action game-collectable card game                             | Regression – predict CLV        |

and product. Therefore, a time-sensitive action determines whether a player who leaves should be considered a churning. To better understand why a particular player is churning, features derived from historical telemetry data that might affect action should be clean, persistent, and independent [Ahn et al., 2020].

For both corporate and profit growth, the gaming industry must be able to foresee and avoid user misunderstanding as it develops in size. There is a change in usage time and in-game behavior before an online gaming user departs a game. Players can make an effortless sly transition from one game to another in a game ecosystem. This phenomenon is frequently seen on mobile devices rather than indie games. Predicting user attrition effectively is crucial since it helps firms better comprehend their expected revenue. Gaming businesses are concentrating on retaining many devoted players who can guarantee a steady stream of cash [Lovato, 2015]. Fig. 1 shows that academic publications in game churn analysis have increased recently. The infertility of finding data on game churn analysis, the difficulty of understanding the data, and the fact that not every researcher has dynamic and mechanical knowledge about the structure of the games can be listed as the limitations of the studies. Despite all limitations, game churn analysis is one of the main topics for academic research in Business Intelligence.

Players may sometimes drop out of games for unexpected reasons, which vary based on gaming platforms and genres. For example, some players lose attention and interest over time and tend to stop playing. Especially in mobile games, leaving occurs in the early stages, while the period is longer in online computer games. However, there is no definitive list of why players leave a game, and it is nearly impossible to predict the reason. Some of the possible reasons are wasting time and money, feeling of failure in the real world, causing social isolation, distracting people from what they need to do, lack of technological devices, poor internet connection, long game sessions, game inputs, and UI effects, charge or subscription policy, etc.

In most media usage, a company aims to entertain its customers. Therefore, delivering content to users is just as important as watching, liking, or sharing the content. Generally, the purpose of playing games is to have fun. As with media, subjective emotions are challenging to measure, so most critical events may be reaching scores and levels or social interactions with friends. Social interactions have become a part of online games. Game development is easier than ever, with the developments in programming tools. However, user acquisition and retention have become significant challenges due to increased rivalry. Telecommunications and finance businesses have been applying churn analysis to predict customer churn for a long time. As a result, by investing a large percentage of their budget on incentives for their whole user base, companies can concentrate their actions and resources on those at greater risk. As mentioned above, it is difficult to determine why players leave a game. Customer churn analysis is a logical solution to determine the rate of players likely to churn and an indicator of the actions that can be implemented to prevent it. This study does not focus on the reasons why players quit a game. Mainly, the emphasis is customer churn prediction in MMO gaming with machine learning models.

The gaming industry does not have churn customers like telecom, finance, and insurance companies. In these sectors, the user/customer chooses another company and cancels the subscription. However, even if the player does not log in to the game for a long time, it is considered inactive in gaming. Unless the player deletes the account, they are still a potential customer. Even if they have no intention of continuing the game, online gaming players rarely deactivate or unsubscribe from their accounts. Only 1% of players who have been idle for more than a year leave the game explicitly by deleting their accounts [Lee et al., 2019]. However, how long does inaction qualify as a churn? This is a challenging question compared to different reasons for it. Because they only play on weekends, some players may be inactive for a few days due to traveling or having a major exam; others may appear inactive

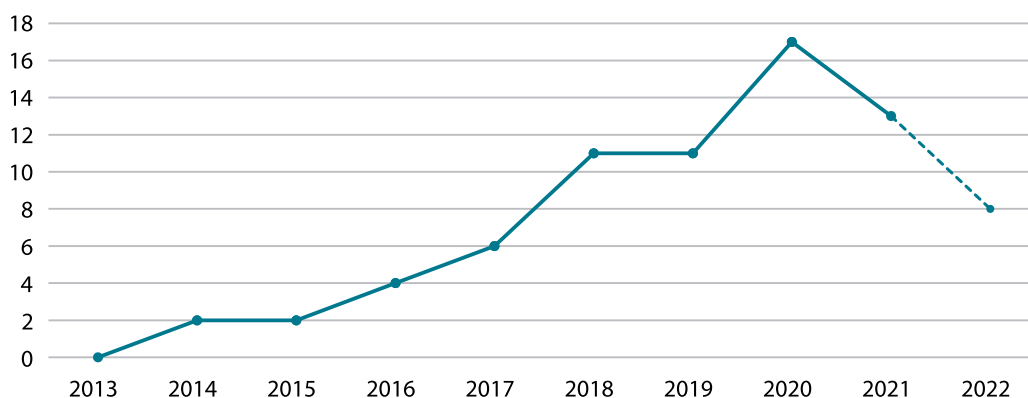


Fig. 1. Number of publications on game churn prediction (from Google Scholar using "Game Churn Prediction" keyword)

Рис. 1. Количество публикаций по прогнозированию оттока пользователей онлайн-игр, ед. (поиск в Google Scholar по ключевому слову Game Churn Prediction)

for several weeks. If the observation period is too short, the misclassification rate will be high. If it is too long, the misclassification rate will be lower, but it will take longer to determine if a player has left the game. As a result, once a player is found to be in a churn, there will not be enough time to persuade them to return [Kim et al., 2017].

## RESEARCH METHOD

This study follows the CRISP-DM methodology, and this section explains the data collection process, dataset, preprocessing steps, research method, and model performance metrics. The data is subjected to an operation to give meaningful and future-oriented outputs. These methodologies are listed as KDD, SEMMA, CRISP-DM, etc.

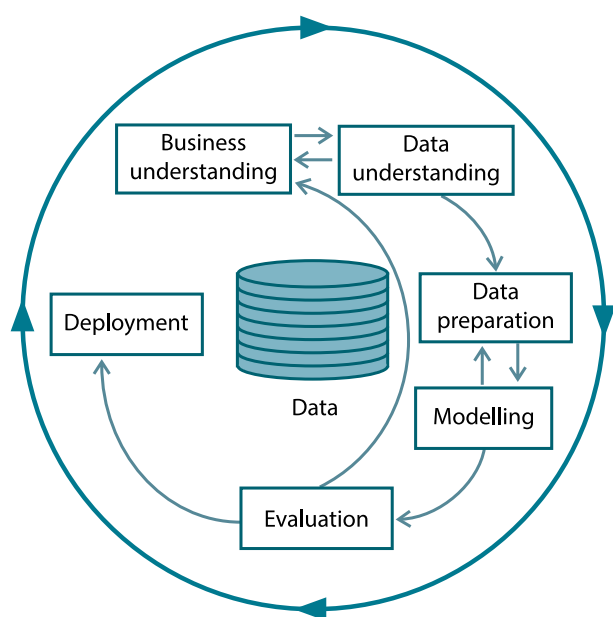


Fig. 2. CRISP-DM methodology life cycle

Рис. 2. Модель жизненного цикла исследования данных CRISP-DM

As depicted in Fig. 2, we guided our research by applying the CRISP-DM model, which stands for Cross Industry Standard Processing for Data Mining. This is a robust and analytical methodology that has been tested for data mining processes. Before formalizing the business methodology, CRISP-DM demonstrates the necessary steps, from internalizing data to implementing solutions [Shearer, 2000]. Our research is based on dual main problems:

*Problem 1:* predicting whether players are going to churn or not – Binary Classification;

*Problem 2:* predicting survival time – Regression.

**Data collection.** Blade & Soul is an MMORPG game played with many players on a server where players can interact with each other and manage their chosen characters in the game. It was released in June 2012 by NCSOFT.

The purpose of our research was to encourage game data mining research by providing researchers with commercial game logs. NCSOFT, one of the largest gaming companies in South Korea, approximately 100 GB of game log data from Blade & Soul released to %10 with researchers. Data were collected in specific periods from 28 March 2016 to 25 August 2017. Table 2 shows the data collection information on Blade and Soul dataset.

The raw data of 10,000 players is processed into a single form. In addition, the company provides churn and survival-time columns, which are assigned as the target attributes for the classification and regression problems, respectively. Test 1 and Test 2 data were presented separately, and the study aimed to achieve classification accuracy on both Test 1 and Test 2 data. The company changed its concept after December 2016 and switched from P2P (PaytoPlay) to F2P (FreetoPlay) in marketing. While it was monetizing P2P for Test 1 and F2P for Test 2 collection. Test 1 and Test 2 performance metrics are given in separate tables. The aim on point is to compare the performance metrics of the model to be developed for P2P and F2P.

**Dataset.** Log data records a character's behavior, event, and status change for game operation or analysis. The base information for finding the phenomenon, cause, and solution is a problem in the game. The dataset consists entirely of numeric attributes. Common, actor, object, and target attributes are given in Fig. 3.

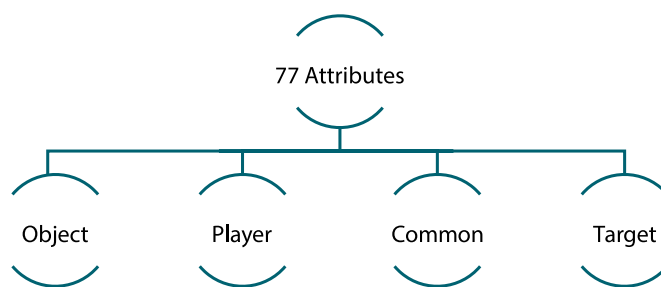


Fig. 3. Attribute groups

Рис. 3. Группы атрибутов

Table 2 – Data collection on Blade and Soul dataset  
Таблица 2 – Массив данных онлайн-игры Blade and Soul

| Dataset  | Data collection period               | Weeks | Business model | Number of players | Size (GB) |
|----------|--------------------------------------|-------|----------------|-------------------|-----------|
| Training | April 1, 2016 – May 11, 2016         | 6     | Subscription   | 4,000             | 48        |
| Test 1   | July 27, 2016 – September 21, 2016   | 8     | Subscription   | 3,000             | 30        |
| Test 2   | December 14, 2016 – February 8, 2017 | 8     | Free-to-play   | 3,000             | 30        |

Common fields include information that is common to all action log, such as log creation time, LogID. Actor fields include information about the subject of the act, such as character ID, account ID, race, level, and class. Object fields include information linked to the action of the actor, such as item ID, grade, and quantity. Target fields include information such as the target's character ID, account ID, race, level, and class.

The 'LogID' attribute in Table 3 corresponds to codes to player in-game activity with 86 sub-codes. The LogID is distributed to the action that may occur while playing Blade and Soul, and 82 main action logs are selected for the analysis that will be covered in this competition. Raw data is mainly integer and float data types. Also, the 'time' attribute is in Year-Month-Day-Hour-Minute-Second format. For example: 2016-07-13 18:06:13. Each raw data has an attribute called 'actor\_account\_id' in itself. It helps to publish the data by anonymizing the user's name.

The game's structure consists of several sub-codes with different percentages, as depicted in Fig. 4. MMORPG games are based on continuous character development, customizing in a community, developing by constantly performing missions, and playing by different audiences via the Internet.

**Steps for data preprocessing.** Data preprocessing, which covers the actions to get the data ready for modeling, such as completing the missing data, integrating, cleaning, etc., is an essential step of the CRISP-DM methodology that needs to be handled prior to model building [Wirth, Hipp, 1999]. The study used the "Scikit learn" library in Python for data preprocessing, data preparation, model performance evaluation, and application steps, and Tableau for data visualization and data manipulation. Pre-processing is counting the values in the "LogID" in the dataset. This method is used to draw the most meaningful conclusions. The steps for preprocessing are listed below.

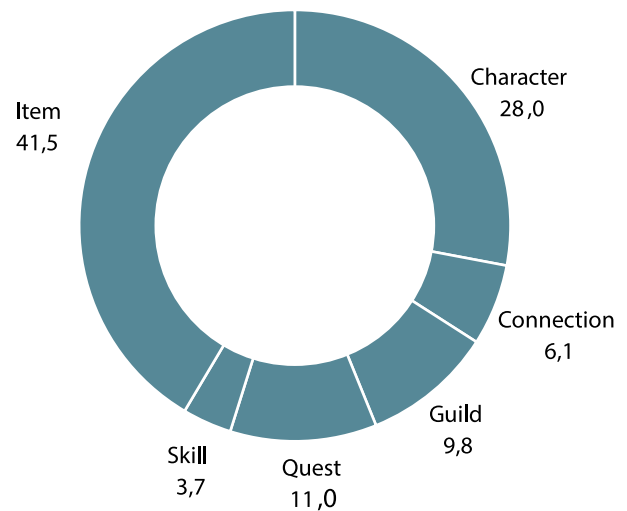


Fig. 4. Sub-code distribution for Blade and Soul, %

Рис. 4. Распределение субкодов в онлайн-игре Blade and Soul, %

1. Since the data is about 100 GB, it was presented to the researchers using the double compression method. When the data is unzipped, the individual files become insurmountable. For this reason, the data was preprocessed without opening targz files in the data preprocessing step by unzipping them once. 1000 different CSV files were converted into a single 10000-line CSV file with preprocessing.

2. Columns with too many 'NaN' data were excluded.

3. Data inconsistencies were evaluated.

4. We did not perform any filling operations on data because of *sparsity*.

5. By the researchers, 0's are seen as not preferred because the player does not do this action even though could do it. Preferably, we did not assign "NaN" value to unused parts, on the contrary, we assigned it as 0, considered *sparse data*.

Table 3 – An example set of sub-code  
Таблица 3 – Пример набора субкодов

| Category   | LogID | LogName       | Description   |
|------------|-------|---------------|---|
| Connection | 1003  | EnterWorld    | When the actor entered the game-server  |
|            | 1004  | LeaveWorld    | When the actor left the game-server   |
| Character  | 1017  | GetMoney      | When the game money increased in the inventory  |
|            | 1022  | GetItem       | When the items in the inventory are increased   |
| Item       | 2004  | DestroyItem   | When the item was destroyed   |
|            | 2113  | RepairItem    | When the actor repaired the item  |
| Skill      | 4001  | AcquireSkill  | When an actor acquires skill  |
|            | 4002  | SkillLevelUp  | When skill was leveled up   |
| Quest      | 5001  | AcquireQuest  | When the actor acquired the quest   |
|            | 5004  | CompleteQuest | When the actor completed the quest  |
| Guild      | 6001  | CreateGuild   | When the actor found the guild by paying an application fee to faction guild supervisor |
|            | 6002  | DestoryGuild  | When the guild was destroyed  |

6. Min-max normalization between 0-1 was applied to data in the classification problem.

7. Standardization was applied to data in the regression problem.

8. New features were generated using feature extraction based on Formulas 1–4 below.

$$\sum_i^j \text{NumberOfPlayedDaysPerPlayer} = \text{survival time} + (\text{last day} - \text{first day}). \quad (1)$$

Since we cannot use the survival time variable directly, we present an approach to use it by generating new attributes in the data. Survival time represents the amount of time the player has spent with the company, while last day and first day represent the first login and last login timestamps as NumberOfPlayedDaysPerPlayer in Formula 1.

$$\sum_i^j \text{TotalLevel} = \text{actor level} + \text{actor mastery level}. \quad (2)$$

In order to find the TotalLevel reached by the player, the level and mastery level variables are included under a single variable. It represents the total level reached by the player in Formula 2.

$$\sum_i^j \text{SurvivalTimeForPerLevel} = \text{survival time} / \text{TotalLevel}. \quad (3)$$

$$\sum_i^j \text{mean\_st\_}[i] = \text{survival time} / \log \text{ID columns } [i]. \quad (4)$$

As shown in Formula 3, we created the elapsed time per level by dividing the total level reached by the player by the time spent in the game variable, and then, as shown in Formula 4, we divided each survival time variable by the columns generated from the 'LogID' variable and generated the variable 'mean\_st\_[i]' for each of them.

To our knowledge, there is no standard threshold in the literature for how the classes should be distributed – also, already known that churn prediction works on more unstable datasets. The data is presented to the researchers in a specific format by NCSOFT, columns with churn labels for classification and survival times for regression are present in the data as seen in Table 4, and no resampling method was used to obtain test sets.

Table 4 – Churn column distributions for dataset  
Таблица 4 – Распределение столбцов оттока игроков для набора данных

| Dataset  | Churned        | Not Churned   | Total  |
|----------|----------------|---------------|--------|
| Training | 2,650 (%66.25) | 1350 (%33.75) | 4,000  |
| Test 1   | 2,092 (%69.73) | 908 (%30.27)  | 3,000  |
| Test 2   | 2,102 (%70.03) | 898 (%29.97)  | 3,000  |
| Total    | 6,844 (%68.4)  | 3156 (%31.6)  | 10,000 |

While the week for observation period in online MMORPG games is predominantly determined as 4 to 5

weeks, this period varies for mobile and console games. Accordingly, inactiveness period was defined as 5 weeks by the company providing the data. Hence, game churn analysis is formed through observation and churn. The observation period starts from the last log-in date to the game and if a player does not enter the game after five weeks of inactiveness, a churn tag is assigned. Even though the player logs in to the game for a short time during these five weeks, the player will remain as a potential customer.

**Data modeling.** As we mentioned in the article, when reviewed studies in the literature, studies are not only predominantly classification but also regression and cluster analysis. Throughout study, we have tried to find a solution to the classification and regression problem. So, what are classification and regression? Classification is a supervised learning method that specifies the class to which the data items belong and is used when the output attribute takes finite and discrete values [Géron, 2019]. Regression is basically a statistical approach to finding the relationship between variables. It is used in machine learning to predict the outcome of an event based on the relationship between the variables obtained from the dataset [Altman, 1992]. This section gives the theoretical content of classification and regression algorithms used in the study with a simple definition.

XGBoost is a powerful and popular open-source software library for gradient boosting in machine learning. In machine learning, boosting is an ensemble learning technique that combines the predictions of multiple weak learners (e.g., decision trees) to create a strong learner that is more accurate than any of the individual weak learners. Boosting algorithms work by sequentially building models, each of which attempts to correct the mistakes of the previous model. XGBoost is particularly effective for classification problems, and the XGBoost classifier is a specific implementation of the XGBoost library that is optimized for classification tasks [Chen, Guestrin, 2016]. Logistic regression is a statistical method used for classification tasks, particularly when the output variable is binary (e.g., 0 or 1, true or false). In logistic regression, the goal is to predict the probability that a given input belongs to a particular class [Wright, 1995]. Support vector machines (SVMs) are a type of supervised learning algorithm that can be used for classification tasks. The goal of an SVM classifier is to find the hyperplane in a high-dimensional space that maximally separates the data points belonging to different classes. The classifier will then fit an SVM model to the data by adjusting the model's parameters to maximize the margin between the hyperplane and the nearest data points of each class [Cortes, Vapnik, 1995].

Random forest is a powerful and popular ensemble learning method for regression tasks, which involves training multiple decision trees and combining their predictions to make a more accurate and stable prediction

[Breiman, 2001]. A random forest regressor is a specific implementation of the random forest algorithm that is optimized for regression tasks, where the goal is to predict a continuous value (e.g., a price or a probability) rather than a discrete class label. Multilayer perceptron (MLP) is a type of artificial neural network (ANN) that can be used for regression tasks, which involves predicting a continuous value (e.g., a price or a probability) rather than a discrete class label. A multilayer perceptron is a class of fully connected feed-forward artificial neural networks called perceptrons that feed input data to the input layer and get output from the output layer [Glorot, Bengio, 2010]. A gradient boosting regressor (GBR) is a specific implementation of the gradient boosting algorithm that is optimized for regression tasks [Friedman, 2001; Hastie, Tibshirani, Friedman, 2009].

**Model performance metrics and confusion matrix.** After implementing a machine learning algorithm, the next step is determining the model's effectiveness based on the datasets. Since the fundamental of machine learning is data, statistics define the functionality in this field. For this reason, some metrics are used to measure the performance of the developed model. Our research discusses machine learning metrics in Table 5 for classification and Table 6 for regression. Accuracy is the percentage of samples that are classified correctly. Recall is a metric that shows the number of true positives that are correctly identified. Finally, precision shows the number of true negatives that are correctly identified [Alpaydin, 2020].

The final score was created by averaging the arithmetic sums of Test 1 and Test 2 resulted in Formula 5.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_i}}, \quad (5)$$

where  $H$  is harmonic mean,  $n$  is the number of model metrics ratio,  $x_1, x_2, \dots, x_i$  are attributes for classification model. The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations.

RMSE (Root Mean Squared Error) value was considered in the regression problem, while RMSE represents the MSE value rooted form as seen in Formula 6.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}, \quad (6)$$

where RMSE is root of the MSE,  $\hat{y}_i$  is the corresponding prediction in the data set,  $y_i$  is the  $i$ -th measurement,  $n$  is the number of data points for regression model. Also, RMSE is commonly used over standardized data.

Considering previous studies, harmonic mean of Precision, Recall, and F1 scores was considered to measure the performance of the model. It is worth noting that model performance metrics and confusion matrix for classification and regression are given in Table 5 and 6.

## RESULTS AND DISCUSSION

This study uses machine learning algorithms to solve classification and regression for MMO game data. Also, predict churners and player survival time by focusing on player log data of Blade and Soul game over specific periods. The performance of a machine learning algorithm can vary depending on how the results are interpreted. However,

Table 5 – Model performance metrics and confusion matrix for classification  
Таблица 5 – Показатели качества модели и матрица ошибок для классификации

|              |          | Predicted class             |   |   |
|--------------|----------|-----------------------------|---|---|
|              |          | Positive                    | Negative                                    |   |
| Actual class | Positive | True Positive<br>$TP$       | False Negative<br>$FN$                      | Recall (Sensitivity)<br>$TP/(TP + FN)$      |
|              | Negative | False Positive<br>$FP$      | True Negative<br>$TN$                       | Precision (Specificity)<br>$TN/(TN + FP)$   |
|              |          | Precision<br>$TP/(TP + FP)$ | Negative Predictive Value<br>$TN/(TN + FN)$ | Accuracy<br>$(TP + NP)/(TP + FP + FN + TN)$ |

Table 6 – Model performance metrics and confusion matrix for regression  
Таблица 6 – Показатели качества модели и матрица ошибок для регрессии

| Metrics                        | Definition  |
|--------------------------------|---|
| $R^2$                          | $R^2$ is a statistical measure of how close the data are to the fitted regression line                              |
| Adjusted $R^2$                 | Adjusted $R^2$ is the value that makes $R^2$ more explainable. It is an emerging metric for the overfitting problem |
| Mean Absolute Error (MAE)      | Absolute error is the difference between the predicted values and the actual values                                 |
| Mean Squared Error (MSE)       | MSE is the most used regression loss function.<br>MSE is the average frame loss per sample over the entire dataset  |
| Root Mean Squared Error (RMSE) | RMSE is the root of the MSE   |



after a successful model is developed in machine learning, it is expected to achieve an accuracy ratio on both positives and negatives. In this direction, Precision, Recall, and F1 scores come to the fore. For this reason, comparing several evaluation metrics before commenting on the research results is safer. Therefore, although the dataset is not considered imbalanced, the accuracy ratios and other performance metrics are presented for the classification model.

**Classification problem.** Fig. 5 shows XGBoost, RF (Random Forest), SVM (Support Vector Machine), LR (Logistic Regression), and KNN (K-Nearest Neighbor) algorithms performed close to each other on Test 1 and Test 2 data in predicting accuracy.

Among the algorithms, XGBoost has the highest accuracy rate and minimizes the difference between Test 1 and Test 2 sets. NCSOFT switched from the P2P (Pay to Play) to F2P (Free to Play) marketing strategy while collecting the Test 1 and Test 2 datasets, respectively. For this reason, the accuracy ratios are not consistent result in the Test 2 dataset of the F2P model, where the difference is relatively high in some algorithms (XGBoost, Logistic Regression, and Support Vector Machine). This is because, in the P2P model, the fact that players are included in the game by paying a specific price show that there are players who love and play the game by filtering specific factors for customer churn analysis in gaming. However, in the F2P model, the player has downloaded the game to try it, played it once, and is likely not to log in again. Therefore, applying campaigns to such players makes

more sense, knowing that the retention rate will not be as low as in the F2P model.

Accordingly, conducting customer churn analysis in the F2P model is more difficult. Therefore, the difference in performance between the models arises. The dataset analysis aims to establish a consistent model for both the P2P and F2P models and can produce meaningful results in the datasets. This is included in the research because the accuracy metric only focuses on positive actual and predicted values. However, negative predicted and actual values are just as important in customer churn analysis in gaming. For this reason, the best final score was achieved with the XGBoost algorithm at 0.94. In these final scores, we can say that XGBoost, LR and SVM algorithms come to the fore. The final score was created to establish the model with a percentage of negative data. In this direction, the final scores of the algorithms were included in Tables 7–9.

Choosing a maximum depth = 6 value provides better results as shown in Table 7. Since XGBoost is an ensemble algorithm, other parameters like number of estimators and binary mode are also critical to model selection and overfitting. Therefore, for parameters, final score is 0.94. XGBoost has a number of optimization techniques that make it more efficient, such as sparsity-aware learning, weighted quantile sketch, and block structure to parallelize tree construction. It has a wide range of hyperparameters that can be tuned to improve performance, such as the learning rate, the maximum depth of the trees, and the number of trees in the ensemble. Also, it is able to handle missing

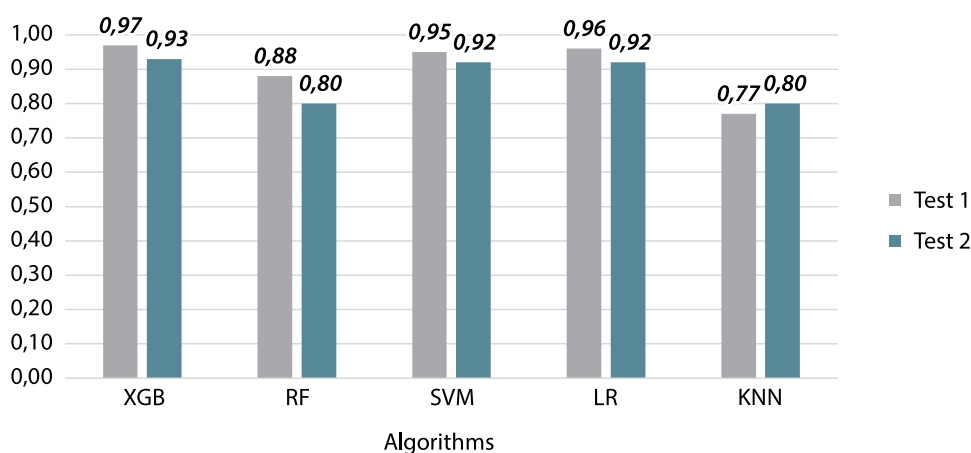


Fig. 5. Accuracy ratios on Test 1 and Test 2

Рис. 5. Соотношение точностей теста 1 и теста 2

Table 7 – Performance metrics for XGBoost

Таблица 7 – Показатели качества для алгоритма XGBoost

| XGBoost | Precision | Recall | F1 Score | Accuracy | Final |
|---------|-----------|--------|----------|----------|-------|
| Test 1  | 0.95      | 0.98   | 0.96     | 0.97     | 0.94* |
| Test 2  | 0.94      | 0.90   | 0.92     | 0.93     |       |

\*Best parameter: max\_depth=6, n\_estimators=5, binary='logistic'

values in the input data, which can be a major challenge in many real-world datasets.

According to the attributes included in Fig. 6, the company could add easier tasks for players at the beginning of game, balance the power of in-game enemy characters, as well as create a specific area for joining a party, creating an environment that can bring players together. On the other hand, game designers could apply statistical models that favor the player in decomposing items and directing them to quests that can give them more experience points for leveling up in game. Also, some changes can be suggested to make the player perspective. It is seen that the importance of the attributes of joining a clan, taking on a new task, forming a party, performing a quest and dying quickly in 1v1 games are relatively higher for players than the others. At the same time, the time spent in the game and its time per player are the attributes with the highest importance. Within the scope of the model, it is suggested that in defining new missions for players, adjusting the difficulty of the missions according to the level, feeding 1v1 with more balanced mechanics, protecting the player from other players in performing the missions, and facilitating the player to enter a clan where player can rise more easily.

As shown in Table 8,  $C = 5$  and 'liblinear' solver is used for Logistic Regression classifier. Although the LR algorithm works stably, it has a minimal lower final score compared to XGBoost in final score. In logistic regression, the  $C$  parameter is a regularization term that controls the strength of the regularization. In general, a larger  $C$  value means that the model is less regularized and may be more prone to overfitting. A smaller  $C$  value means that the model is more regularized and may be more prone to underfitting. The solver parameter specifies the algorithm used to fit the logistic regression model. Different solvers may be better suited for different types of data and different problem characteristics. 'Liblinear' is a linear solver that is based on the Coordinate Descent algorithm. It is generally faster than the other solvers for small to

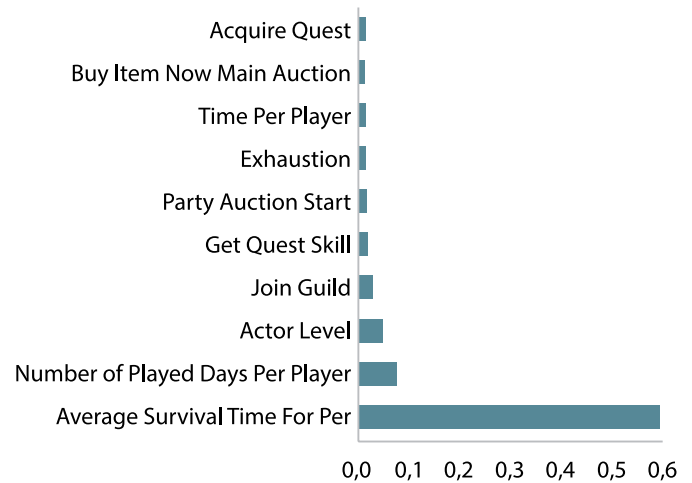


Fig. 6. Feature importance of XGBoost algorithm  
Рис. 6. Значимость переменных алгоритма XGBoost

medium-sized datasets but may not perform as well on very large datasets.

Better results are obtained when the 'C' and minimizing gamma ratio for Support Vector Machines in Table 9. On the other hand, SVM is the algorithm that stands out like Logistic Regression in final score. The Radial Basis Function (RBF) kernel is a common choice for support vector machines (SVMs) because it can handle non-linearly separable data and often yields good performance in practice. Value of parameter  $C$  has not been set too high to avoid overfitting and underfitting. The RBF kernel is a universal kernel, which means that it can approximate any continuous function to any desired accuracy, given enough data. This makes it a very flexible kernel that can handle a wide range of data types and problem complexities. It is well-suited for high-dimensional data, as it can capture complex relationships between the input variables without requiring an excessively large number of model parameters. Also, this kernel is computationally efficient, as it only requires a single dot product between the input vectors at each evaluation. Overall, the RBF kernel is a popular choice for SVMs because it can han-

Table 8 – Performance metrics for Logistic Regression

Таблица 8 – Показатели качества для алгоритма Logistic Regression

| LR     | Precision | Recall | F1 Score | Accuracy | Final |
|--------|-----------|--------|----------|----------|-------|
| Test 1 | 0.94      | 0.97   | 0.95     | 0.96     | 0.93* |
| Test 2 | 0.92      | 0.88   | 0.90     | 0.92     |       |

\*Best parameter:  $C=5$ , solver='liblinear'

Table 9 – Performance metrics for Support Vector Machines

Таблица 9 – Показатели качества для алгоритма Support Vector Machines

| SVM    | Precision | Recall | F1 Score | Accuracy | Final |
|--------|-----------|--------|----------|----------|-------|
| Test 1 | 0.94      | 0.97   | 0.94     | 0.95     | -     |
| Test 2 | 0.93      | 0.89   | 0.90     | 0.92     |       |

\*Best parameter:  $C=100$ , gamma=0.01, kernel='rbf'.

dle non-linearly separable data and often performs well in practice, making it a useful tool for many types of machine learning tasks.

**Regression problem.** Fig. 7 shows the results of the regression algorithms applied to the survival time given. If all data points were on a straight line, it would form a perfect regression curve, but this is practically impossible. Therefore, any point that does not coincide with the curve is considered a residual for us. The sum of these residuals represents the error of the model.

Besides the classification problem, the second problem in the data is based on predicting the survival times of players. This metric provides information on how many days apart players will spend in-game. Knowing such metrics, especially in customer churn analysis, is essential in making sense of the analysis. As we mentioned earlier, regression aims to minimize residual values. Our regression models consist of linear and non-linear models. In linear regression models, model parameters have a linear structure. However, in nonlinear regression models, the model parameters are not linear, but the independent variables can be linear or nonlinear.

Fig. 7 and Table 10 show all regressor performances and best parameters, respectively.

RF, MLP and GBR regressor for Test 1 and Test 2 data performed and minimized the error rate for linear and non-linear models. In this part, different regression algorithms try to minimize the error rates of Test 1 and Test 2

because the data was collected in different periods, while regression algorithms such as GBR, MLP and RF work well in the F2P and P2P system. Therefore, the difference between the applied algorithms might not be significant for the regression problem.

## CONCLUSION

Technological developments have created differences in our lifestyles and habits and how we do business. Data plays an essential role in all these approaches. Data constitutes the indispensable building blocks of roadmaps for institutions in the long and short term. This situation has increased the data needs of institutions and caused them to acquire goals such as revealing the unknown from the data they have obtained, deriving inferences from the facts that have taken place, and shaping their profit-making strategies. It has become a science that creates added value for more sectors and different industries [Han, Kamber, Jai, 2012]. For example, in the gaming industry, customer churn has become more essential and needs to be analyzed in all its bearings.

Online games have been around for a while now and they are still growing in popularity. The number of players is increasing, and the industry is still booming. But with growth come challenges, one of which is churn analysis. The main goal of churn analysis is to find out what causes customers to stop playing the game and then fix it before it becomes a problem. This can be done by looking

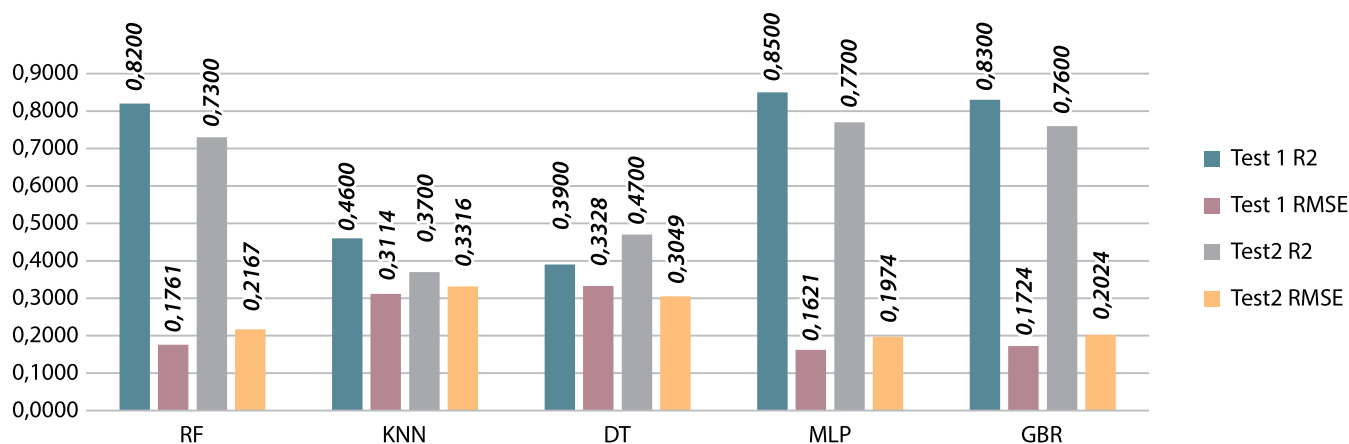


Fig. 7. Regression  $R^2$  and RMSE Error Rates on Test 1 and Test 2  
Рис. 7. Регрессия  $R^2$  и частоты ошибок RMSE для теста 1 и теста 2

Table 10 – Parameters for regression algorithms  
Таблица 10 – Параметры алгоритмов для задачи регрессии

| Algorithm | Best parameters  |
|-----------|--|
| RF        | n_estimators = 100, max_depth = 2, min_samples_split = 2, max_features = 1.0   |
| KNN       | n_neighbors = 3, weights = uniform, algorithm = 'auto', leaf_size = 30   |
| DT        | Splitter = 'best', max_depth = 2, min_samples_split = 2, max_features = 'auto'   |
| MLP       | hidden_layer_sizes = 77, activation = 'relu', solver = 'adam', alpha = 0.0001, learning_rate = 0.01, shuffle = True, early_stopping = 'True' |
| GBR       | loss = 'squared_error', learning_rate = 0.01, n_estimators = 100, criterion = 'friedman_mse', max_depth = 3, alpha = 0.9                     |

at player behavior and seeing what they do before they leave the game. The main objective of churn management is to decrease player dropout rates. One way to achieve this goal is by providing more content to players, which will increase their motivation and reduce the chances that they will leave the game. Machine learning offers an alternative approach and solution for game churn analysis [Drachen, El-Nasr, Canossa, 2013].

Churn management is usually a part of the game design and it is one of the most important game metrics for success. Companies can also reduce the dropout churn rate through a variety of actions that can be planned as a result of loss analysis in online games. Some of these include:

- Increasing advertising will increase players' motivation and recognition.
- It provides free content to players with a limited time frame for access. This gives players something to work towards, like a prize or reward.
- Providing constant rewards and updates in order to keep the game interesting.
- In line with customer analysis, in-game metrics can be analyzed, and changes can be made to the game-related structures (User Interface-UI, User Experience-UX, gameplay, and mechanics, etc.).

While answering business questions regarding online gaming to predict game cherner for classification and survival time for regression, this study contributes to the game churn literature by classifying and regression estimation of player data in F2P and P2P periods.

Also, it shows churn analysis performed on the Blade and Soul dataset, different machine learning algorithms were applied, and it was observed that they performed well in both the final score and the accuracy metrics. Also, academic research on gaming is ongoing, and when results are compared broadly, it is evident that the XGBoost algorithm outperformed in classification and RF, GBR and MLP gives best results for regression problems. Also, adding some hyperparameters may help to enhance the model performance. On the other hand, although the LR and SVM algorithm works stably, it has a lower final score in the final problem than the other XGBoost algorithms. Our research also draws attention to the data sparsity step in the data pre-processing process for customer churn analysis in the game. It also suggests that not deleting as many columns as possible in the data preprocessing step will improve the data quality and the results' success. This step also helps algorithms to work more efficiently.

High accuracy ratios were achieved by applying different approaches to the data of different MMORPG games. However, it should not be neglected that players may leave a game for various reasons. This further complicates the estimates made in studies to predict game churn. Therefore, each dataset needs an ideocratic modeling and evaluation process. In addition, it should be noted that it is challenging to find data for game customer churn analysis studies. Therefore, more open data should be presented to researchers, and it will also prevent the sterility of academic studies in this field. ■

## References

- Adams E., Dormans J. (2012). *Game mechanics: Advanced game design*. New Riders.
- Ahn J., Hwang J., Kim D., Choi H., Kang S. (2020). A survey on churn analysis in various business domains. *IEEE Open Access*, vol. 8, pp. 816–839.
- Almana A.M., Aksoy M.S., Alzahrani R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, vol. 4, pp. 165–171.
- Alpaydin E. (2020). *Introduction to machine learning*. 4th ed. The MIT Press.
- Altman N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, vol. 46, no. 3, pp. 175–185.
- Beyer H.R., Holtzblatt K. (1995). Apprenticing with the customer. *Communications of the ACM*, vol. 38, no. 5, pp. 45–52.
- Borbora Z.H., Srivastava J. (2012). User behavior modelling approach for churn prediction in online games (pp. 51–60). *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*.
- Borbora Z.H., Srivastava J., Hsu K.W., Williams D. (2011). Churn prediction in MMORPGs using player motivation theories and an ensemble approach (pp. 157–164). *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*.
- Breiman L. (2001). Random forests. *Machine Learning*, vol. 45, no. 1, pp. 5–32.
- Buhl H.U., Röglinger M., Moser F., Heidemann J. (2013). Big data. *Business & Information Systems Engineering*, vol. 5, no. 2, pp. 65–69.
- 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. 255–265.
- Chen T., Guestrin C. (2016). XGBoost: A Scalable tree boosting system. In: *The 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Crawford C. (2003). *Chris Crawford on game design*. New Riders.

- Cortes C., Vapnik V. (1995). Support-vector networks. *Machine Learning*, vol. 20, pp. 273–297. <https://doi.org/10.1007/BF00994018>
- Çelik Ö., Osmanoğlu U. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, vol. 4, no. 1, pp. 30–38.
- Dahiya K., Bhatia S. (2015). Customer churn analysis in telecom industry (pp. 1–6). *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*.
- Davenport T.H., Barth P., Bean R. (2012). How “Big Data” is different. *MIT Sloan Management Review*, vol. 54, no. 1, pp. 22–24.
- Ding J., Gao D., Chen X. (2015). Alone in the game: Dynamic spread of churn behavior in a large social network a longitudinal study in MMORPG. *International Journal of Smart Home*, vol. 9, no. 3, pp. 35–44. DOI: 10.14257/ijsh.2015.9.3.04
- Drachen A., Seif El-Nasr M., Canossa A., (2013). *Game analytics – The basics*. In: M. Seif El-Nasr, A. Drachen, A. Canossa. (Eds.). *Game Analytics*. Springer, London.
- Drachen A., Green J., Gray C., Harik E., Lu P., Sifa R., Klabjan D. (2016). Guns and guardians: Comparative cluster analysis and behavioral profiling in destiny (pp. 1–8). *2016 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Duke R.D. (1980). A paradigm for game design. *Simulation & Games*, vol. 11, no. 3, pp. 364–377.
- Friedman J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, vol. 29, no. 1. DOI: 10.1214/aos/1013203451
- George G., Haas M.R., Pentland A. (2014). Big data and management. *Academy of management Journal*, vol. 57, no. 2, pp. 321–326.
- Géron A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (Second edition)*. O'Reilly Media, Inc.
- Glorot X., Bengio Y. (2010). Understanding the difficulty of training deep feedforward neural networks (pp. 249–256). In: Y.W. The, M. Titterton (Eds.). *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (PMLR)*. Chia Laguna Resort, Sardinia, Italy, vol. 9.
- Gold C., Tzuo T. (2020). *Fighting churn with data: The science and strategy of customer retention*. Manning Publishing.
- Hadiji F., Sifa R., Drachen A., Thureau C., Kersting K., Bauckhage C. (2014). Predicting player churn in the wild (pp. 1–8). *2014 IEEE Conference on Computational Intelligence and Games*.
- Han J., Kamber M., Pei J. (2012). *Data mining: Concepts and techniques* (3rd ed). Elsevier.
- Hastie T., Tibshirani R., Friedman J. (2009). *Elements of statistical learning*. 2<sup>nd</sup> ed. Springer.
- Ijsselstein W., Nap H.H., de Kort Y., Poels K. (2007). Digital game design for elderly users (pp. 17–22). *Proceedings of the 2007 Conference on Future Play*.
- KhakAbi S., Gholamian M.R., Namvar M. (2010). Data mining applications in customer churn management (pp. 220–225). *2010 International Conference on Intelligent Systems, Modelling and Simulation*.
- Kim S., Choi D., Lee E., Rhee W. (2017). Churn prediction of mobile and online casual games using play log data. *PLoS ONE*, vol. 12, no. 7. <https://doi.org/10.1371/journal.pone.0180735>
- Kummer L.B.M., Cesar Nievola J., Paraiso E.C. (2018). Applying commitment to churn and remaining players lifetime prediction (pp. 1–8). *2018 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Lazarov V., Capota M. (2007). Churn prediction business analytics course. *TUM Computer Science*, pp. 33–34.
- Lee E., Jang Y., Yoon D.M., Jeon J., Yang S., Lee S.K., Kim D.W., Chen P.P., Guitart A., Bertens P., Perianez A., Hadiji F., Muller M., Joo Y., Lee J., Hwang I., Kim K.J. (2019). Game data mining competition on churn prediction and survival analysis using commercial game log data. *IEEE Transactions on Games*, vol. 11, no. 3, pp. 215–226.
- Lejeune M.A.P.M. (2001). Measuring the impact of data mining on churn management. *Internet Research*, vol. 11, no. 5, pp. 375–387.
- Long X., Yin W., An L., Ni H., Huang L., Luo Q., Chen Y. (2012). Churn analysis of online social network users using data mining techniques (pp. 551–556). In: *International MultiConference of Engineers and Computer Scientists, IMECS 2012*, vol. 2195.
- Lovato N. (2015) 16 reasons why players are leaving your game. *GameAnalytics*. <https://gameanalytics.com/blog/16-reasons-players-leaving-game/>
- Markus M.L. (2015). New games, new rules, new scoreboards: The potential consequences of big data. *Journal of Information Technology*, vol. 30, no. 1, pp. 58–59.
- Meyer C., Schwager A. (2007). Understanding customer experience. *Harvard Business Review*, vol. 85, no. 2, pp. 116.
- Payne A. (2008). *The handbook of CRM: Achieving excellence in customer management*. Elsevier Butterworth-Heinemann.
- Rothmeier K., Pflanzl N., Hullmann J.A., Preuss M. (2021). Prediction of player churn and disengagement based on user activity data of a freemium online strategy game. *IEEE Transactions on Games*, vol. 13, no. 1, pp. 78–88. DOI: 10.1109/TG.2020.2992282
- Sagiroglu S., Sinanc D. (2013). Big data: A review (pp. 42–47). *2013 International Conference on Collaboration Technologies and Systems (CTS)*.
- Schell J. (2019). *The art of game design: A book of lenses (3rd ed.)*. Taylor & Francis.
- Schermann M., Hemsén H., Buchmüller C., Bitter T., Krcmar H., Markl V., Hoeren T. (2014). Big data. *Wirtschaftsinformatik*, vol. 56, no. 5, pp. 281–287.
- Shearer C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22.
- Singh H., Samalia H.V. (2014). A business intelligence perspective for churn management. *Procedia-Social and Behavioral Sciences*, vol. 109, pp. 51–56. DOI: 10.1016/j.sbspro.2013.12.420

- Sydow L. (2020). Mobile gaming: a \$100 billion industry that's only getting bigger. <https://www.data.ai/en/insights/market-data/mobile-gaming-a-100-billion-industry-thats-only-getting-bigger/>
- Tamassia M., Raffe W., Sifa R., Drachen A., Zambetta F., Hitchens M. (2016). Predicting player churn in destiny: A Hidden Markov models approach to predicting player departure in a major online game (pp. 1–8). *2016 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Vyas R., Prasad B.G.M., Vamshidhar H.K., Kumar S. (2018). Predicting inactiveness in telecom (prepaid) sector: A complex bigdata application (pp. 39–43). *2018 International Conference on Information Technology (ICIT)*.
- Wirth R., Hipp J. (1999). *CRISP-DM: Towards a standard process model for data mining*. Pp. 1–11.
- Woerner S.L., Wixom B.H. (2015). Big data: Extending the business strategy toolbox. *Journal of Information Technology*, vol. 30, no. 1, pp. 60–62.
- Wright R.E. (1995). Logistic regression. In: L.G. Grimm, P.R. Yarnold. (Eds.). *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association Press.
- Zhao S., Wu R., Tao J., Qu M., Zhao M., Fan C., Zhao H. (2022). *perCLTV: A general system for personalized customer lifetime value prediction in online games*. ACM Transactions on Information Systems.
- Zheng A., Chen L., Xie F., Tao J., Fan C., Zheng Z. (2020). Keep you from leaving: Churn prediction in online games (pp. 263–279). In: Y. Nah, B. Cui, S.-W. Lee, J.X. Yu, Y.-S. Moon, S.E. Whang. (Eds.). *Database Systems for Advanced Applications*.

### Источники

- Adams E., Dormans J. (2012). *Game mechanics: Advanced game design*. New Riders.
- Ahn J., Hwang J., Kim D., Choi H., Kang S. (2020). A survey on churn analysis in various business domains. *IEEE Open Access*, vol. 8, pp. 816–839.
- Almana A.M., Aksoy M.S., Alzahrani R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *International Journal of Engineering Research and Applications*, vol. 4, pp. 165–171.
- Alpaydin E. (2020). *Introduction to machine learning*. 4th ed. The MIT Press.
- Altman N.S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, vol. 46, no. 3, pp. 175–185.
- Beyer H.R., Holtzblatt K. (1995). Apprenticing with the customer. *Communications of the ACM*, vol. 38, no. 5, pp. 45–52.
- Borbora Z.H., Srivastava J. (2012). User behavior modelling approach for churn prediction in online games (pp. 51–60). *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*.
- Borbora Z.H., Srivastava J., Hsu K.W., Williams D. (2011). Churn prediction in MMORPGs using player motivation theories and an ensemble approach (pp. 157–164). *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*.
- Breiman L. (2001). Random forests. *Machine Learning*, vol. 45, no. 1, pp. 5–32.
- Buhl H.U., Röglinger M., Moser F., Heidemann J. (2013). Big data. *Business & Information Systems Engineering*, vol. 5, no. 2, pp. 65–69.
- 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. 255–265.
- Chen T., Guestrin C. (2016). XGBoost: A Scalable tree boosting system. In: *The 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Crawford C. (2003). *Chris Crawford on game design*. New Riders.
- Cortes C., Vapnik V. (1995). Support-vector networks. *Machine Learning*, vol. 20, pp. 273–297. <https://doi.org/10.1007/BF00994018>
- Çelik Ö., Osmanoglu U. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, vol. 4, no. 1, pp. 30–38.
- Dahiya K., Bhatia S. (2015). Customer churn analysis in telecom industry (pp. 1–6). *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*.
- Davenport T.H., Barth P., Bean R. (2012). How “Big Data” is different. *MIT Sloan Management Review*, vol. 54, no. 1, pp. 22–24.
- Ding J., Gao D., Chen X. (2015). Alone in the game: Dynamic spread of churn behavior in a large social network a longitudinal study in MMORPG. *International Journal of Smart Home*, vol. 9, no. 3, pp. 35–44. DOI: 10.14257/ijsh.2015.9.3.04
- Drachen A., Seif El-Nasr M., Canossa A., (2013). *Game analytics – The basics*. In: M. Seif El-Nasr, A. Drachen, A. Canossa. (Eds.). *Game Analytics*. Springer, London.
- Drachen A., Green J., Gray C., Harik E., Lu P., Sifa R., Klabjan D. (2016). Guns and guardians: Comparative cluster analysis and behavioral profiling in destiny (pp. 1–8). *2016 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Duke R.D. (1980). A paradigm for game design. *Simulation & Games*, vol. 11, no. 3, pp. 364–377.
- Friedman J. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, vol. 29, no. 1. DOI: 10.1214/aos/1013203451
- George G., Haas M.R., Pentland A. (2014). Big data and management. *Academy of management Journal*, vol. 57, no. 2, pp. 321–326.

- Géron A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (Second edition)*. O'Reilly Media, Inc.
- Glorot X., Bengio Y. (2010). Understanding the difficulty of training deep feedforward neural networks (pp. 249–256). In: Y.W. The, M. Titterton (Eds.). *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (PMLR)*. Chia Laguna Resort, Sardinia, Italy, vol. 9.
- Gold C., Tzuo T. (2020). *Fighting churn with data: The science and strategy of customer retention*. Manning Publishing.
- Hadiji F., Sifa R., Drachen A., Thureau C., Kersting K., Bauckhage C. (2014). Predicting player churn in the wild (pp. 1–8). *2014 IEEE Conference on Computational Intelligence and Games*.
- Han J., Kamber M., Pei J. (2012). *Data mining: Concepts and techniques* (3rd ed). Elsevier.
- Hastie T., Tibshirani R., Friedman J. (2009). *Elements of statistical learning*. 2<sup>nd</sup> ed. Springer.
- Ijsselstein W., Nap H.H., de Kort Y., Poels K. (2007). Digital game design for elderly users (pp. 17–22). *Proceedings of the 2007 Conference on Future Play*.
- KhakAbi S., Gholamian M.R., Namvar M. (2010). Data mining applications in customer churn management (pp. 220–225). *2010 International Conference on Intelligent Systems, Modelling and Simulation*.
- Kim S., Choi D., Lee E., Rhee W. (2017). Churn prediction of mobile and online casual games using play log data. *PLoS ONE*, vol. 12, no. 7. <https://doi.org/10.1371/journal.pone.0180735>
- Kummer L.B.M., Cesar Nievola J., Paraiso E.C. (2018). Applying commitment to churn and remaining players lifetime prediction (pp. 1–8). *2018 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Lazarov V., Capota M. (2007). Churn prediction business analytics course. *TUM Computer Science*, pp. 33–34.
- Lee E., Jang Y., Yoon D.M., Jeon J., Yang S., Lee S.K., Kim D.W., Chen P.P., Guitart A., Bertens P., Perianez A., Hadiji F., Muller M., Joo Y., Lee J., Hwang I., Kim K.J. (2019). Game data mining competition on churn prediction and survival analysis using commercial game log data. *IEEE Transactions on Games*, vol. 11, no. 3, pp. 215–226.
- Lejeune M.A.P.M. (2001). Measuring the impact of data mining on churn management. *Internet Research*, vol. 11, no. 5, pp. 375–387.
- Long X., Yin W., An L., Ni H., Huang L., Luo Q., Chen Y. (2012). Churn analysis of online social network users using data mining techniques (pp. 551–556). In: *International MultiConference of Engineers and Computer Scientists, IMECS 2012*, vol. 2195.
- Lovato N. (2015) 16 reasons why players are leaving your game. GameAnalytics. <https://gameanalytics.com/blog/16-reasons-players-leaving-game/>
- Markus M.L. (2015). New games, new rules, new scoreboards: The potential consequences of big data. *Journal of Information Technology*, vol. 30, no. 1, pp. 58–59.
- Meyer C., Schwager A. (2007). Understanding customer experience. *Harvard Business Review*, vol. 85, no. 2, pp. 116.
- Payne A. (2008). *The handbook of CRM: Achieving excellence in customer management*. Elsevier Butterworth-Heinemann.
- Rothmeier K., Pflanzl N., Hullmann J.A., Preuss M. (2021). Prediction of player churn and disengagement based on user activity data of a freemium online strategy game. *IEEE Transactions on Games*, vol. 13, no. 1, pp. 78–88. DOI: 10.1109/TG.2020.2992282
- Sagiroglu S., Sinanc D. (2013). Big data: A review (pp. 42–47). *2013 International Conference on Collaboration Technologies and Systems (CTS)*.
- Schell J. (2019). *The art of game design: A book of lenses (3rd ed.)*. Taylor & Francis.
- Schermann M., Hensen H., Buchmüller C., Bitter T., Krcmar H., Markl V., Hoeren T. (2014). Big data. *Wirtschaftsinformatik*, vol. 56, no. 5, pp. 281–287.
- Shearer C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, vol. 5, no. 4, pp. 13–22.
- Singh H., Samalia H.V. (2014). A business intelligence perspective for churn management. *Procedia-Social and Behavioral Sciences*, vol. 109, pp. 51–56. DOI: 10.1016/j.sbspro.2013.12.420
- Sydow L. (2020). Mobile gaming: a \$100 billion industry that's only getting bigger. <https://www.data.ai/en/insights/market-data/mobile-gaming-a-100-billion-industry-thats-only-getting-bigger/>
- Tamassia M., Raffe W., Sifa R., Drachen A., Zambetta F., Hitchens M. (2016). Predicting player churn in destiny: A Hidden Markov models approach to predicting player departure in a major online game (pp. 1–8). *2016 IEEE Conference on Computational Intelligence and Games (CIG)*.
- Vyas R., Prasad B.G.M., Vamshidhar H.K., Kumar S. (2018). Predicting inactiveness in telecom (prepaid) sector: A complex bigdata application (pp. 39–43). *2018 International Conference on Information Technology (ICIT)*.
- Wirth R., Hipp J. (1999). *CRISP-DM: Towards a standard process model for data mining*. Pp. 1–11.
- Woerner S.L., Wixom B.H. (2015). Big data: Extending the business strategy toolbox. *Journal of Information Technology*, vol. 30, no. 1, pp. 60–62.
- Wright R.E. (1995). Logistic regression. In: L.G. Grimm, P.R. Yarnold. (Eds.). *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association Press.
- Zhao S., Wu R., Tao J., Qu M., Zhao M., Fan C., Zhao H. (2022). *perCLTV: A general system for personalized customer lifetime value prediction in online games*. *ACM Transactions on Information Systems*.
- Zheng A., Chen L., Xie F., Tao J., Fan C., Zheng Z. (2020). Keep you from leaving: Churn prediction in online games (pp. 263–279). In: Y. Nah, B. Cui, S.-W. Lee, J.X. Yu, Y.-S. Moon, S.E. Whang. (Eds.). *Database Systems for Advanced Applications*.

## Information about the authors

**Kaan Arik**

Lecturer of Multidimensional Modeling and Animation Dept. **Sakarya Applied Science University**, Sakarya, Turkey. E-mail: kaanarik@subu.edu.tr

**Murat Gezer**

PhD (Electrical Electronics Engineering), Associate Professor of Informatics Dept. **Istanbul University**, Istanbul, Turkey. E-mail: murat.gezer@istanbul.edu.tr

**Seda Tolun Tayali**

PhD (Quantitative Methods), Professor Doctor of Quantitative Methods Dept. **Istanbul University**, Istanbul, Turkey. E-mail: stolun@istanbul.edu.tr

## Информация об авторах

**Арик Каан**

Преподаватель кафедры многомерного моделирования и анимации. **Университет прикладных наук Сакарья**, г. Сакарья, Турция. E-mail: kaanarik@subu.edu.tr

**Гезер Мурат**

PhD (Электротехника и Электроника), доцент кафедры информатики. **Стамбульский университет**, г. Стамбул, Турция. E-mail: murat.gezer@istanbul.edu.tr

**Толун Тайали Седа**

PhD (Количественные методы), профессор кафедры количественных методов. **Стамбульский университет**, г. Стамбул, Турция. E-mail: stolun@istanbul.edu.tr