**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра САПР**

Курсовая РАБОТА

**по дисциплине «Автоматизация схемотехнического проектирования»**

Тема: Анализ экспериментальных данных

|  |  |  |
| --- | --- | --- |
| Студент гр. 1301 |  | Семейкин С.А. |
| Преподаватель |  | Боброва Ю.О. |

Санкт-Петербург

2025

**ЗАДАНИЕ**

**на курсовую работу**

|  |  |  |
| --- | --- | --- |
| Студент Семейкин С.А. | | |
| Группа 1301 | | |
| Тема работы:Анализ экспериментальных данных | | |
| Исходные данные:  Дан набор данных Online Gaming Behavior Dataset. Необходимо проанализировать его в соответствии с заданием курсовой работы. | | |
| Содержание пояснительной записки:  «Введение», «Анализ данных», «Обучение классификатора», «Заключение». | | |
| Предполагаемый объем пояснительной записки:  Не менее 10 страниц. | | |
| Дата выдачи задания: 01.03.2024 | | |
| Дата сдачи реферата: 10.03.2024 | | |
| Дата защиты реферата: 10.03.2024 | | |
| Студент |  | Семейкин С.А. |
| Преподаватель |  | Боброва Ю.О. |

**Аннотация**

В рамках курсовой работы был произведен анализ экспериментальных данных из датасета о вовлеченности геймеров в компьютерные игры. Основная цель работы – выявить закономерность и параметры, которые влияют на вовлеченность человека в игру. С использованием выбранного языка программирования Python были построены графики, рассчитаны метрики, а также проведено обучение модели классификатора для определения вовлеченности на тестовых данных.

**Summary**

In this course work the experimental data from the dataset on gamers' involvement in computer games was analyzed. The main purpose of the work is to identify patterns and parameters that affect human involvement in the game. Using the chosen Python programming language, graphs were built, metrics were calculated, and a classifier model was trained to determine engagement based on test data.

**Анализ данных**

Исходный датасет содержал следующие параметры:

* PlayerID: Уникальный идентификатор для каждого игрока.
* Возраст
* Пол
* Местоположение: Географическое местоположение игрока.
* Жанр игры
* Количество игровых часов: Среднее количество часов, затрачиваемых на игру за сеанс
* Внутриигровые покупки: Указывает, совершает ли игрок внутриигровые покупки (0 = Нет, 1 = Да)
* Сложность игры: уровень сложности игры
* Количество игровых сессий в неделю
* Средняя продолжительность игровой сессии в минутах
* Уровень игрока: текущий уровень игрока в игре
* Разблокированные достижения
* Степень вовлеченности: "Высокий", "Средний", "Низкий"

Параметр PlayerID был сразу исключен из данных, как незначимый, ведь является лишь идентификатором игрока, никак не влияющим на итоговую степень вовлеченности.

Параметр Местоположения был проигнорирован, так как в данной работе целью является определение параметров, непосредственно связанных с игрой. Рассмотреть географическое расположение игроков и влияние этого фактора на вовлеченность можно в ходе дополнительных исследований.

Всего записей в датасете – 40034 строки.

В первую очередь было произведено преобразование категориальных данных в численные, по средству кодирования их в численные значения.

Соответствующие значения представлены далее:

**Gender**

{'Male': 0, 'Female': 1}

**GameGenre**

{'Strategy': 0, 'Sports': 1, 'Action': 2, 'RPG': 3, 'Simulation': 4}

**GameDifficulty**

{'Medium': 0, 'Easy': 1, 'Hard': 2}

**EngagementLevel**

{'Medium': 0, 'High': 1, 'Low': 2}

Затем были построены гистограммы всех параметров (Рисунок 1):

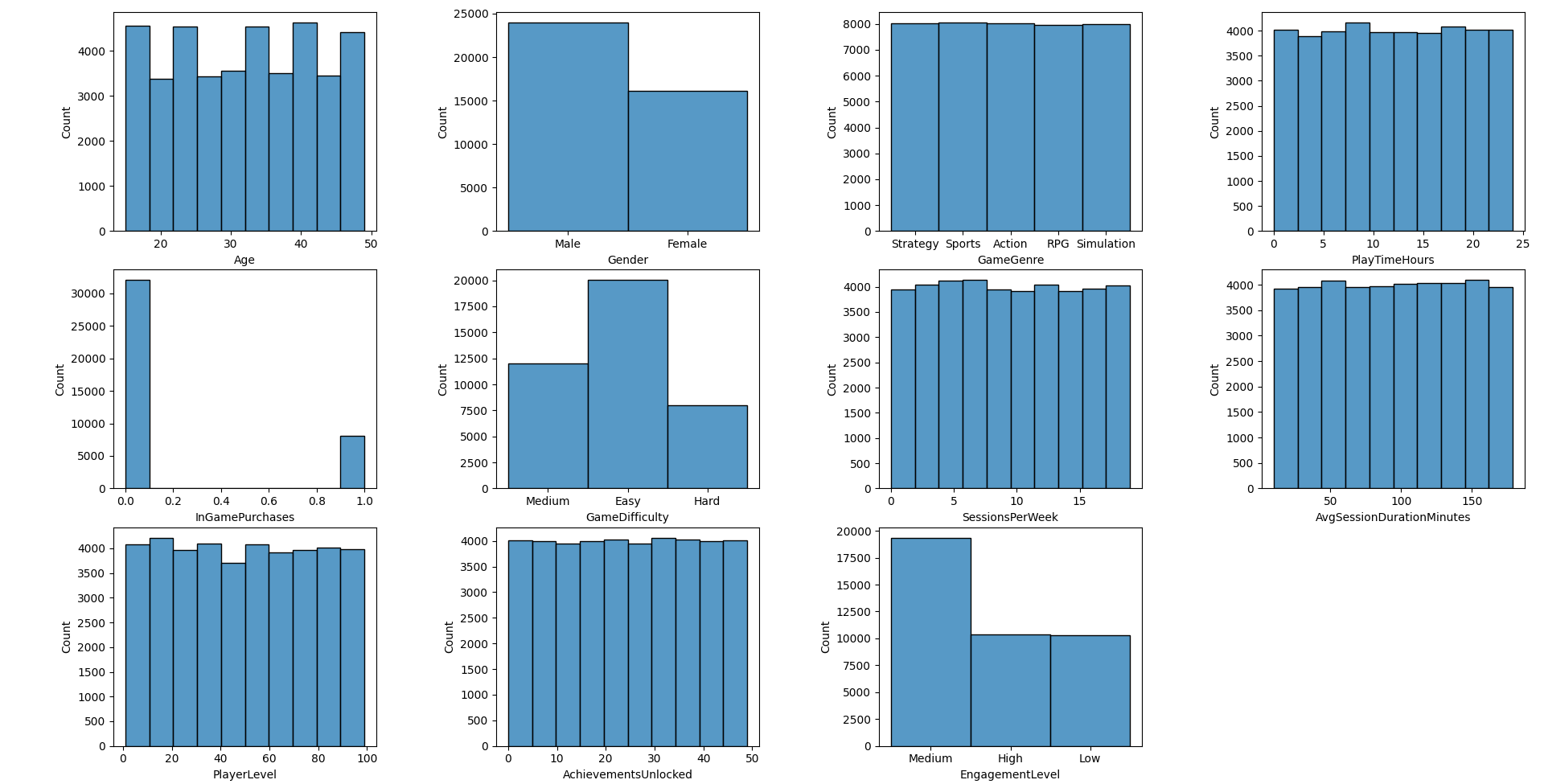


Рисунок 1 Гистограммы всех параметров

Как видно из последней гистограммы в исходных данных есть дисбаланс классов, так как «средняя» вовлеченность присутствует примерно в 2 раза чаще чем низкая или высокая, что может повлиять на результаты работа модели классификатора.

Для численных данных был произведен анализ, результаты которого представлены на рисунке 2.

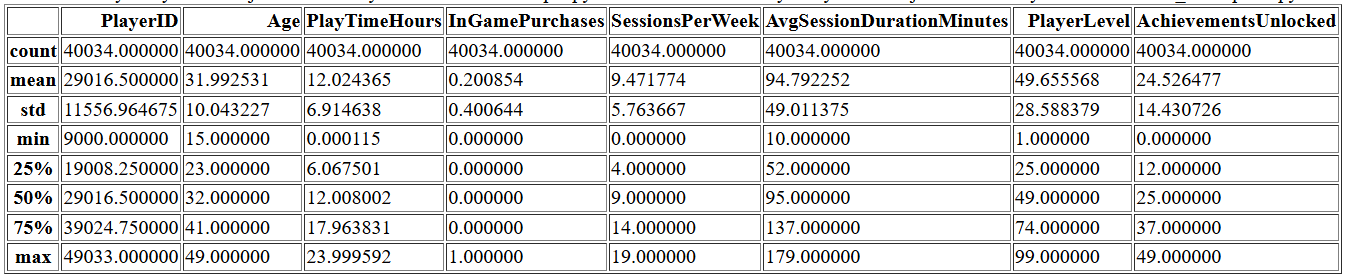


Рисунок 2 Анализ численных данных

Далее были построены «коробки с усами» для численных данных (рисунок 2 и 3):

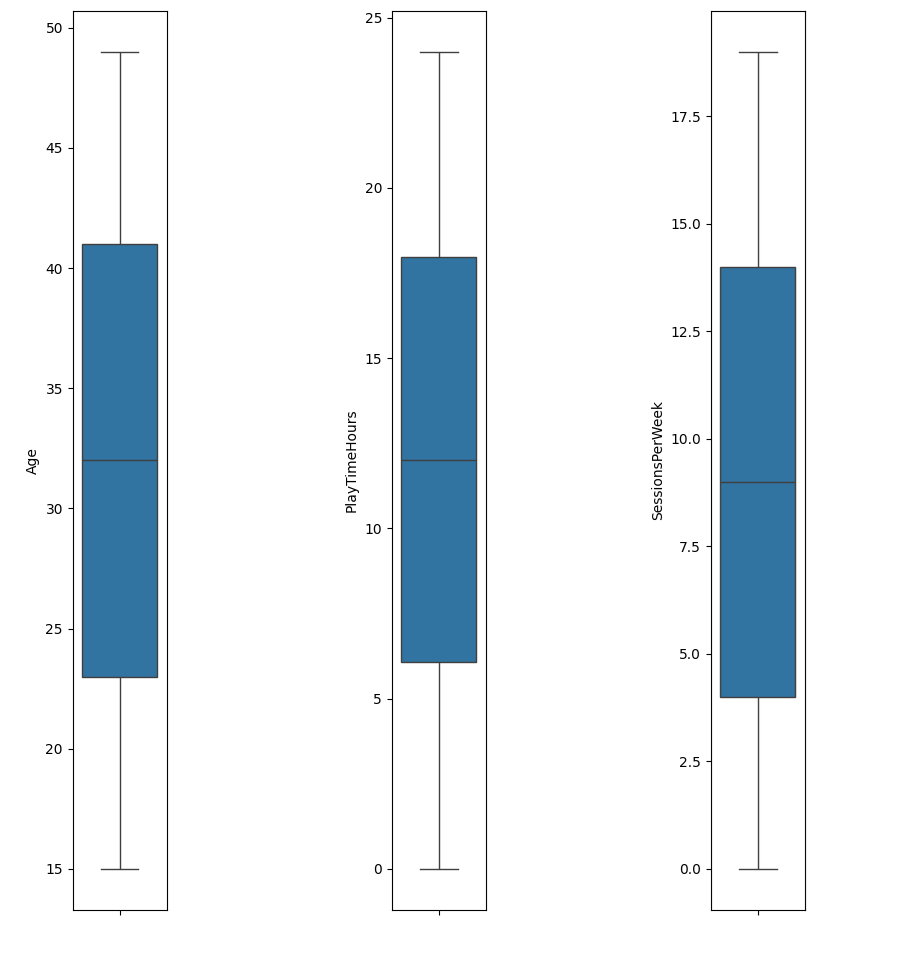


Рисунок 3 Коробки с усами для параметров Age, PlayTimeHours, SessionsPerWeek

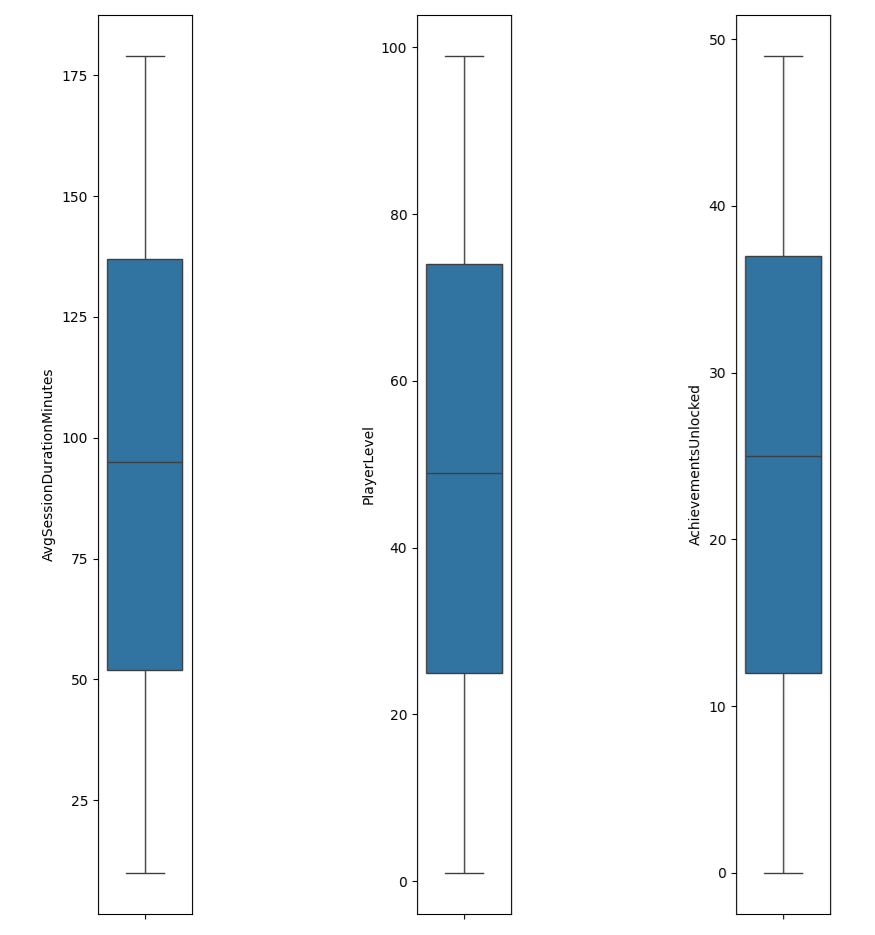


Рисунок 4 Коробки с усами для параметров AvgSessionDurationMinutes, PlayerLevel, AchievementsUnlocked

Для категориальных данных были построены круговые диаграммы (Рисунок 5-7):

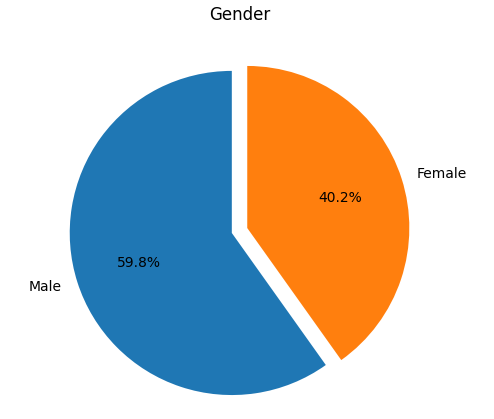


Рисунок 5 Круговая диаграмма распределения параметра Gender

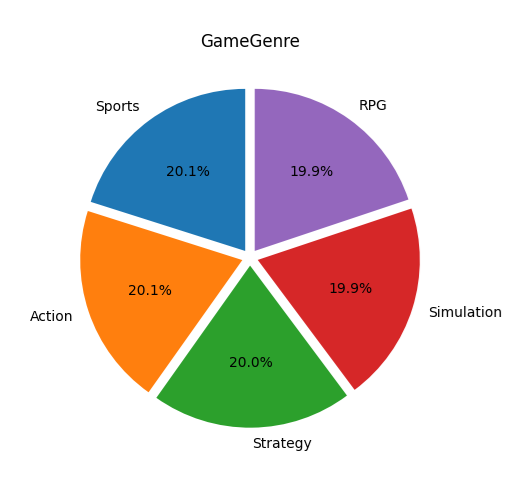


Рисунок 6 Круговая диаграмма распределения параметра GameGenre

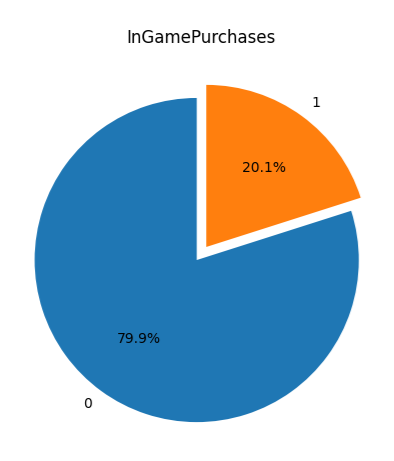


Рисунок 7 Круговая диаграмма распределения параметра InGamePurchases

Затем была построена корреляционная матрица всех параметров (Рисунок 8):

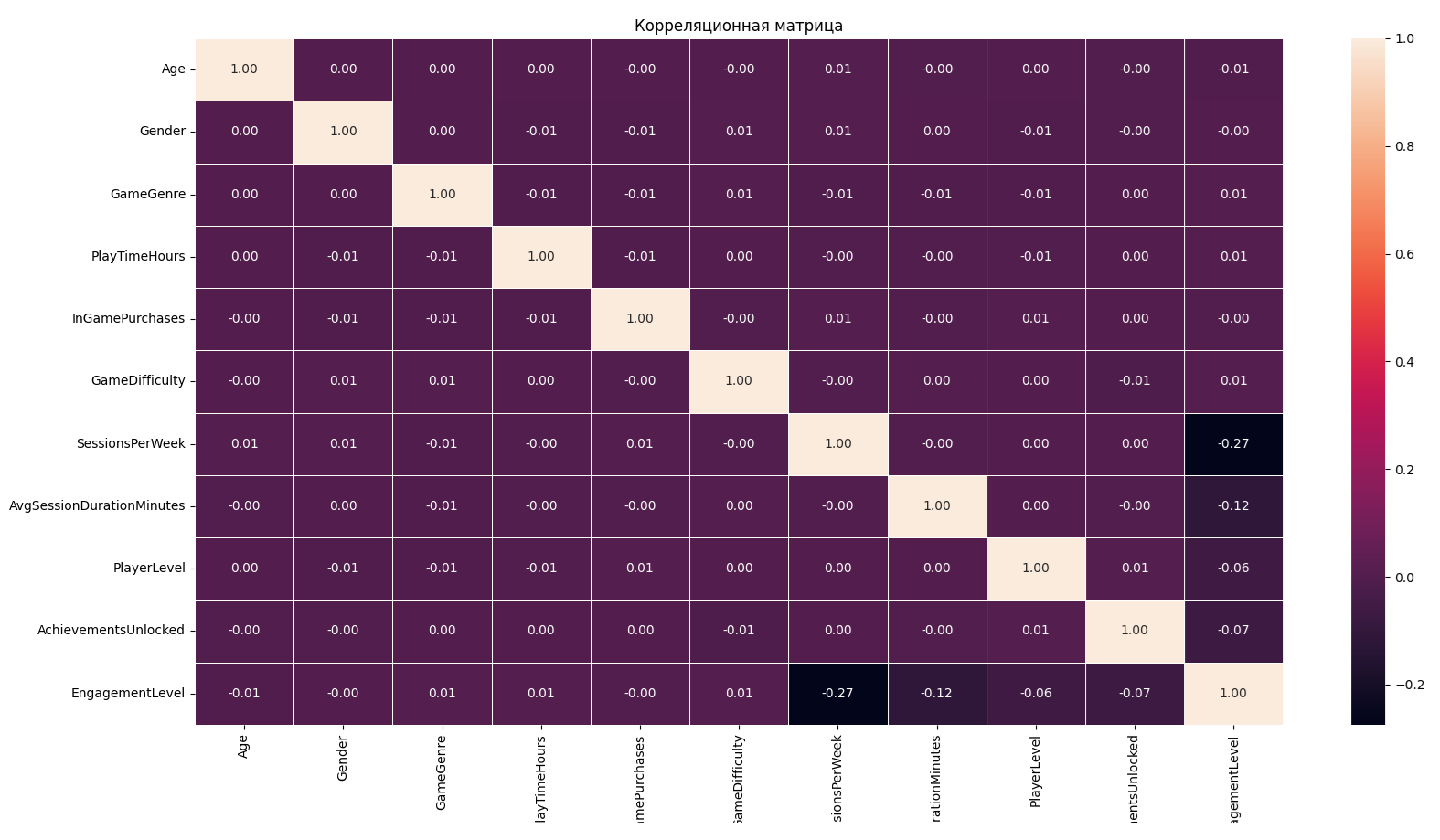


Рисунок 8 Корреляционная матрица

Исходя из нее можно сделать вывод о том, что наиболее значимыми для предсказания значения уровня вовлеченности являются параметры: SessionsPerWeek, AvgSessionDurationMinutes, PlayerLevel и AchievementsUnlocked.

Так же можно сказать о том, что между собой параметры практически не имеют корреляции.

**Обучение классификатора:**Для решения данной задачи были проверены три классификатора: RandomForestClassifier, LogisticRegression и DecisionTreeClassifier

**Метрики для RandomForest с параметром n\_estimators=20, max\_depth=13**

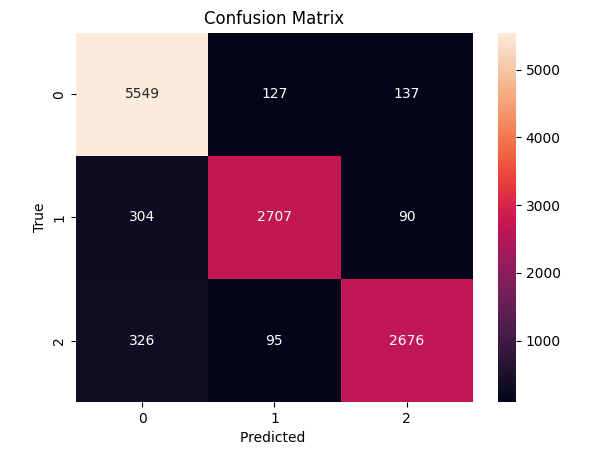


Рисунок 9 Confusion Matrix

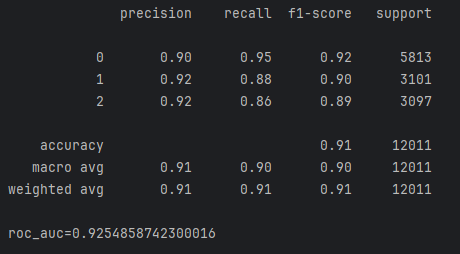


Рисунок 10 Precision, recall, f1, accuracy и roc-auc метрики

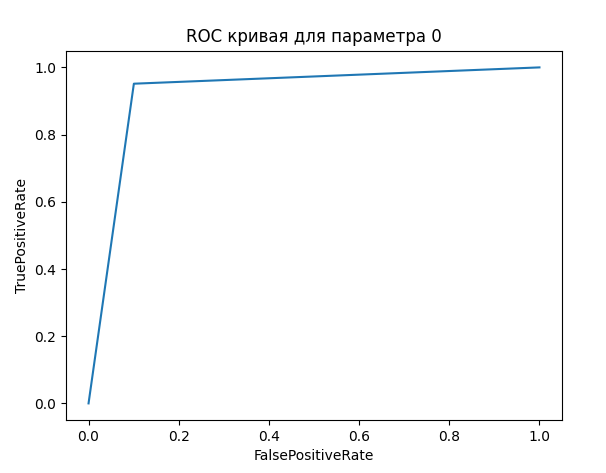


Рисунок 11 ROC кривая

**Метрики для LogisticRegression**

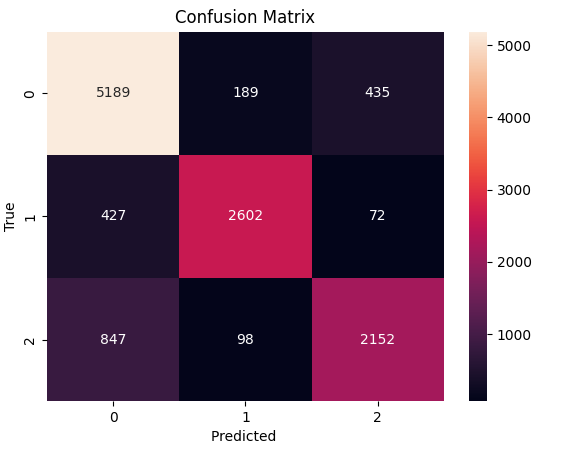


Рисунок 12 Confusion Matrix

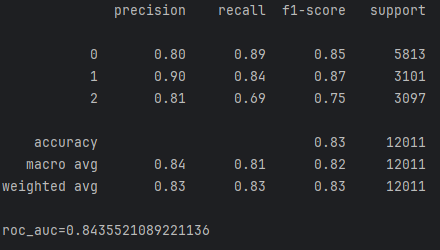


Рисунок 13Precision, recall, f1, accuracy и roc-auc метрики

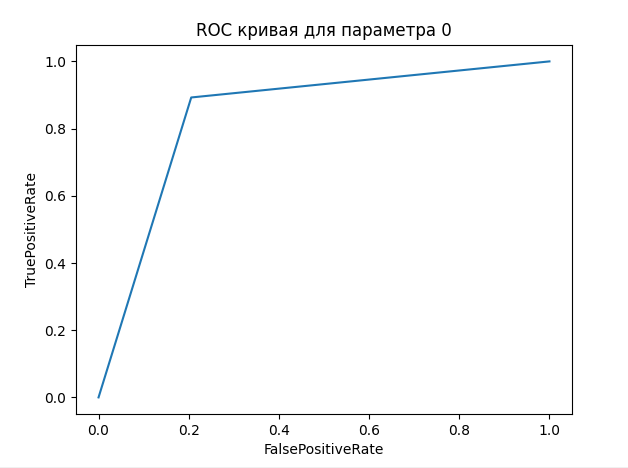


Рисунок 14 ROC кривая

**Метрики для DecisionTree с параметром max\_depth=11**

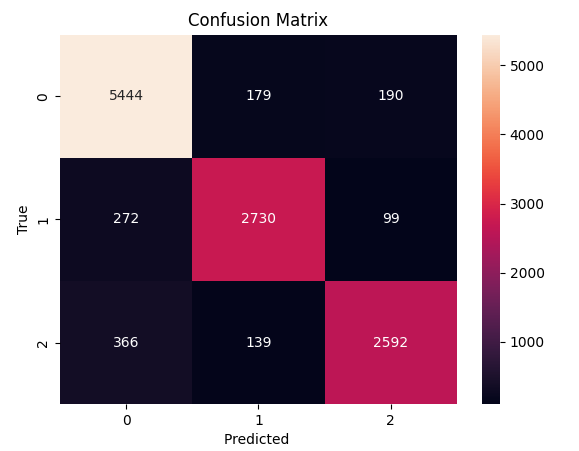


Рисунок 15 Confusion Matrix

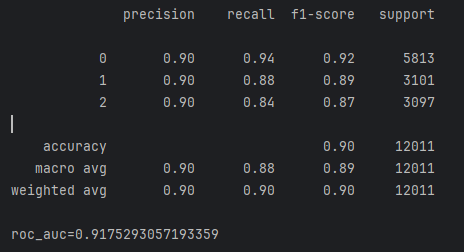


Рисунок 16 Precision, recall, f1, accuracy и roc-auc метрики

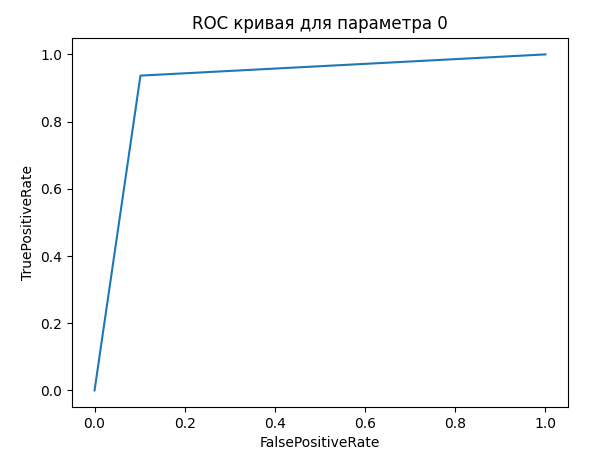


Рисунок 17 ROC кривая

**заключение**

В ходе выполнения задания для курсовой работы удалось успешно провести сбор экспериментальных данных и выполнить анализ данных, используя выбранный датасет Online Gaming Behavior Dataset.

По результатам обучения классификаторов лучшим из трех выбранных оказался классификатор на основе случайного леса – RandomForestClassifier. Мета-параметры (n\_estimators, max\_depth) которого были подобраны на основе метрик, путем перебора различных значений и оптимизации их для получения наилучшего сочетания результатов.

Точность итогового классификатора составила около 91%, при этом для разных итоговых классов метрика F1 – гармоническое среднее precision и recall, находилась в промежутке от 90 до 92 процентов.

Так же площадь под ROC кривой была около 0.925, что является хорошим показателем для классификатора.

Таким образом, выполнение данного задания позволило провести анализ экспериментальных данных, выявить особенности распределений параметров.

На основе данной работы можно выдвинуть предположения о необходимости использования тех или иных механик и факторов во время разработки и аналитики компьютерных игр. К примеру, задания которые будут вынуждать игрока повышать количество сессий в неделю и продолжительность этих сессий могут позитивно сказаться на итоговой вовлеченности игрока.

**список использованных источников**

1. Predict Online Gaming Behavior Dataset // https://www.kaggle.com/ URL: https://www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset (дата обращения: 01.03.2025).
2. Джордан Морроу «Как вытащить из данных максимум». – «Альпина Паблишер», 2021. - 349 с.
3. Документация библиотеки scikit-learn URL: https://scikit-learn.org/stable/index.html (дата обращения: 01.03.2025).
4. Документация библиотеки seaborn URL: https://seaborn.pydata.org/ (дата обращения: 01.03.2025).
5. Документация библиотеки pandas URL: https://pandas.pydata.org/docs/user\_guide/index.html (дата обращения: 01.03.2025).