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):

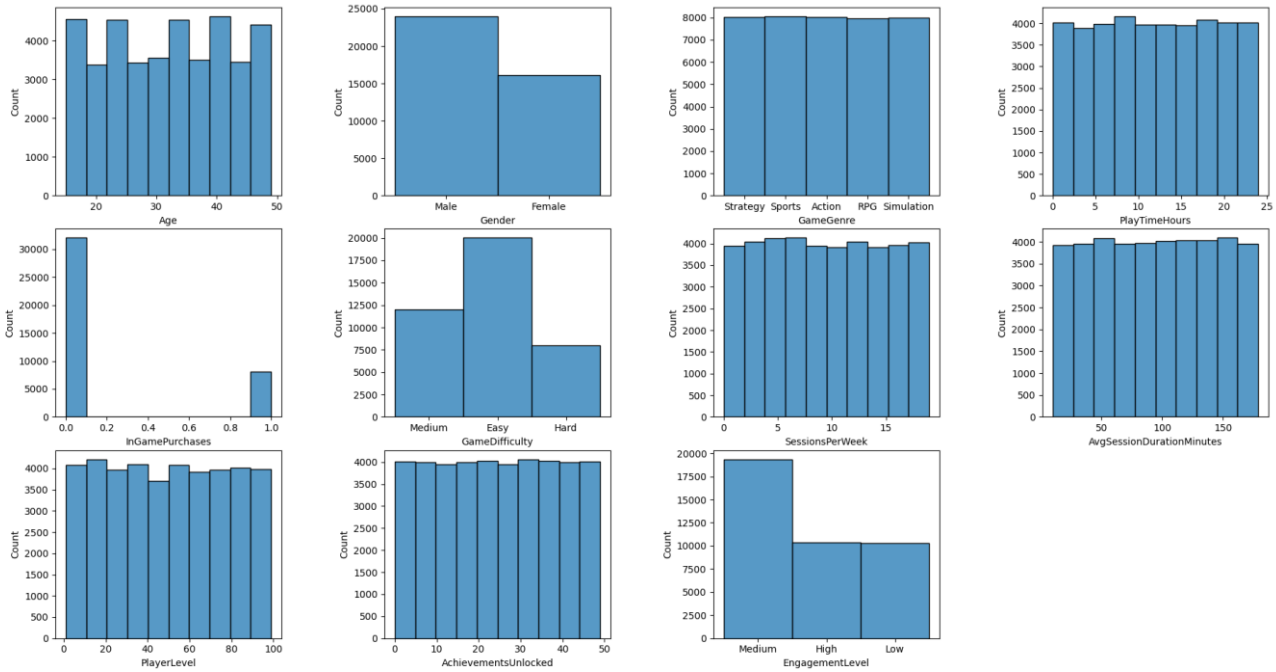

Рисунок 1 Гистограммы всех параметров

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.

	PlayerID	Age	PlayTimeHours	InGamePurchases	SessionsPerWeek	AvgSessionDurationMinutes	PlayerLevel	AchievementsUnlocked
count	40034.000000	40034.000000	40034.000000	40034.000000	40034.000000	40034.000000	40034.000000	40034.000000
mean	29016.500000	31.992531	12.024365	0.200854	9.471774	94.792252	49.655568	24.526477
std	11556.964675	10.043227	6.914638	0.400644	5.763667	49.011375	28.588379	14.430726
min	9000.000000	15.000000	0.000115	0.000000	0.000000	10.000000	1.000000	0.000000
25%	19008.250000	23.000000	6.067501	0.000000	4.000000	52.000000	25.000000	12.000000
50%	29016.500000	32.000000	12.008002	0.000000	9.000000	95.000000	49.000000	25.000000
75%	39024.750000	41.000000	17.963831	0.000000	14.000000	137.000000	74.000000	37.000000
max	49033.000000	49.000000	23.999592	1.000000	19.000000	179.000000	99.000000	49.000000

Рисунок 2 Анализ численных данных

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

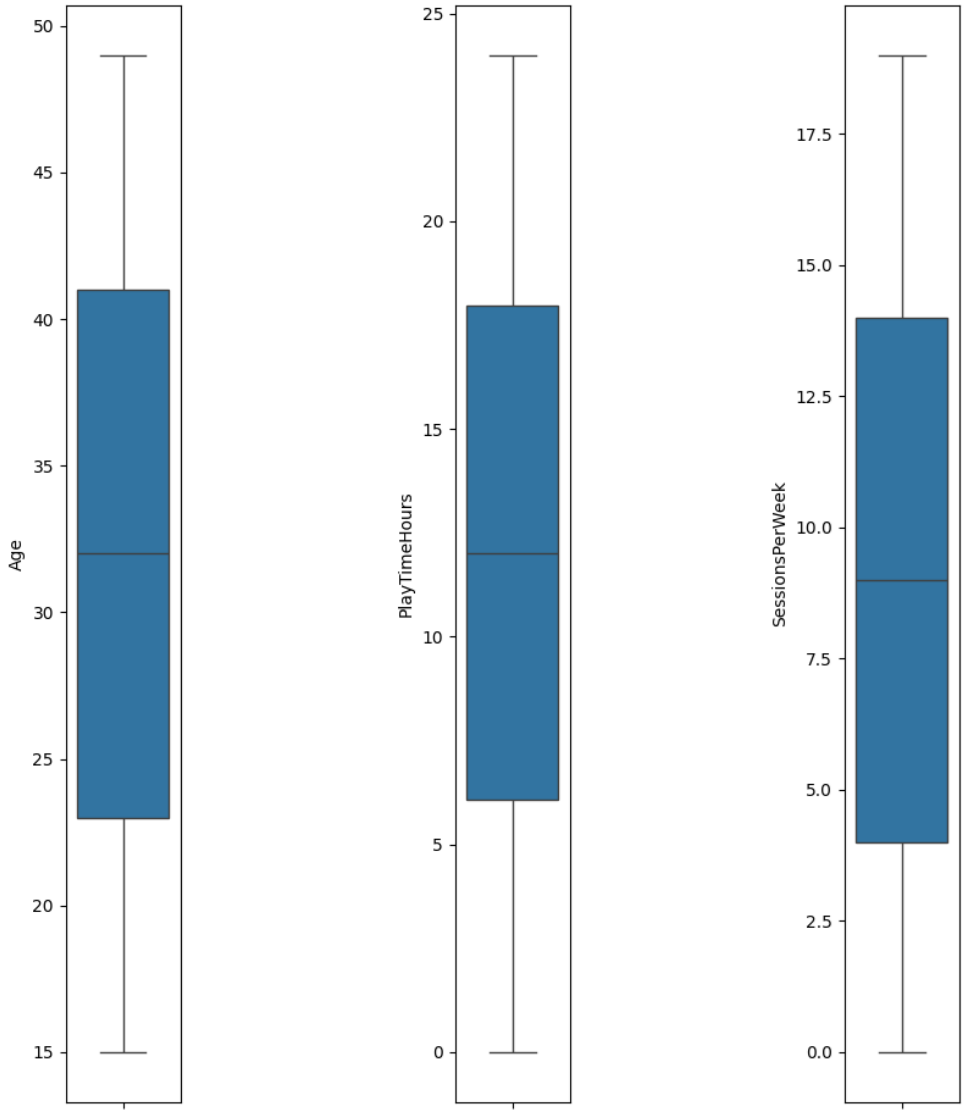

Рисунок 3 Коробки с усами для параметров Age, PlayTimeHours, SessionsPerWeek

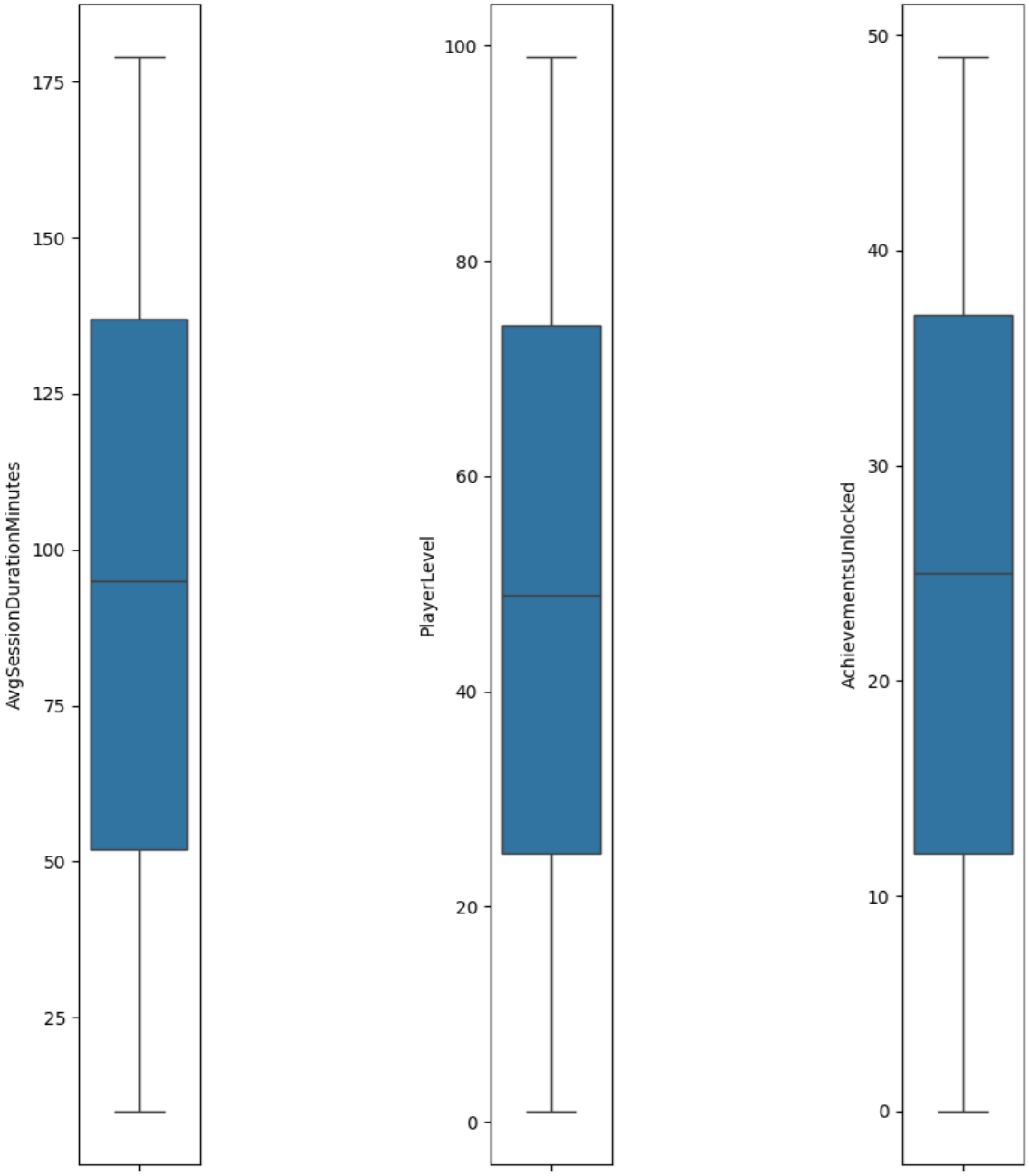

Рисунок 4 Коробки с усами для параметров AvgSessionDurationMinutes, PlayerLevel, AchievementsUnlocked

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

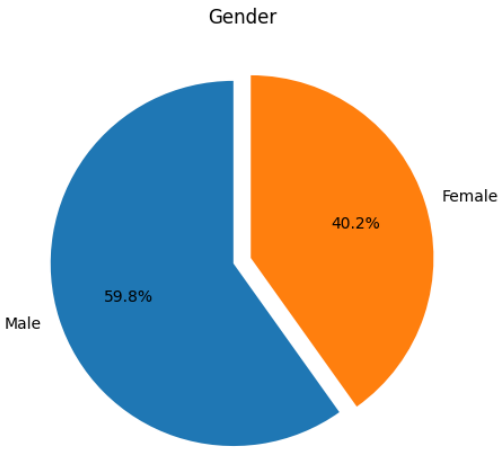

Рисунок 5 Круговая диаграмма распределения параметра Gender

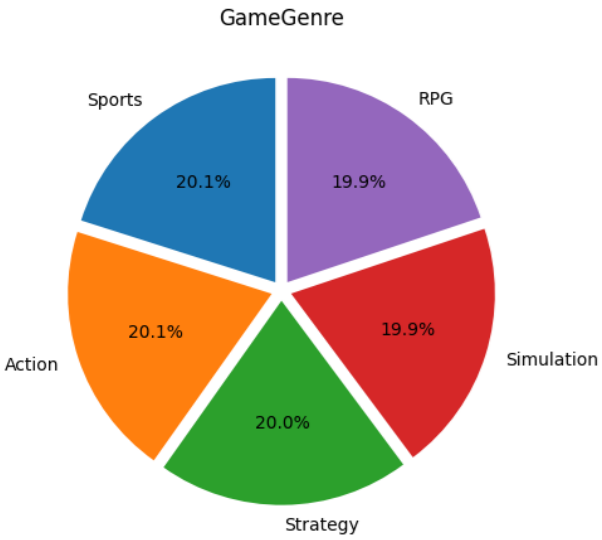

Рисунок 6 Круговая диаграмма распределения параметра GameGenre

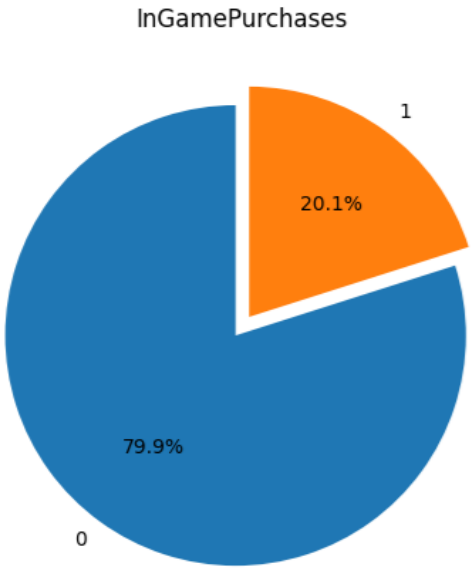

Рисунок 7 Круговая диаграмма распределения параметра InGamePurchases

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

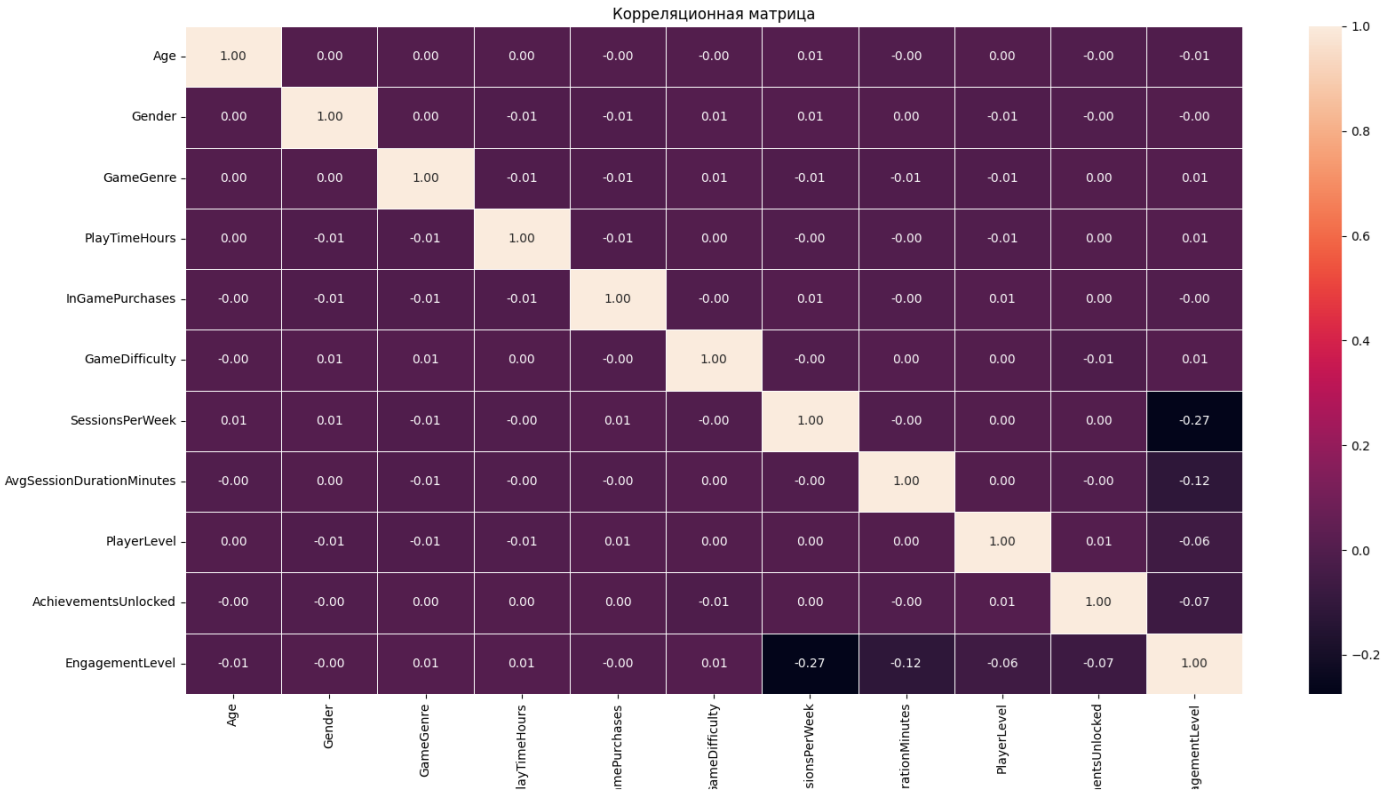

Рисунок 8 Корреляционная матрица

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

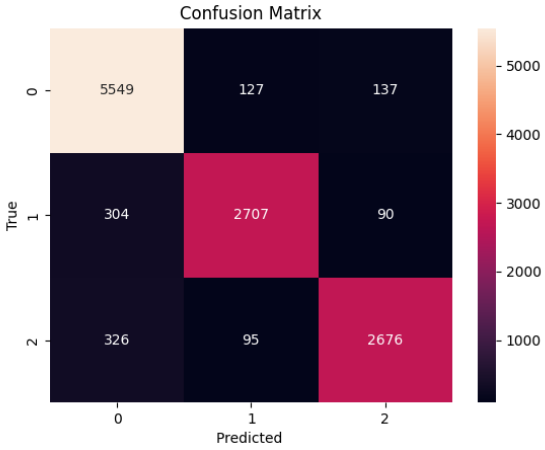

Рисунок 9 Confusion Matrix

	precision	recall	f1-score	support
0	0.90	0.95	0.92	5813
1	0.92	0.88	0.90	3101
2	0.92	0.86	0.89	3097
accuracy			0.91	12011
macro avg	0.91	0.90	0.90	12011
weighted avg	0.91	0.91	0.91	12011

roc_auc=0.9254858742300016

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

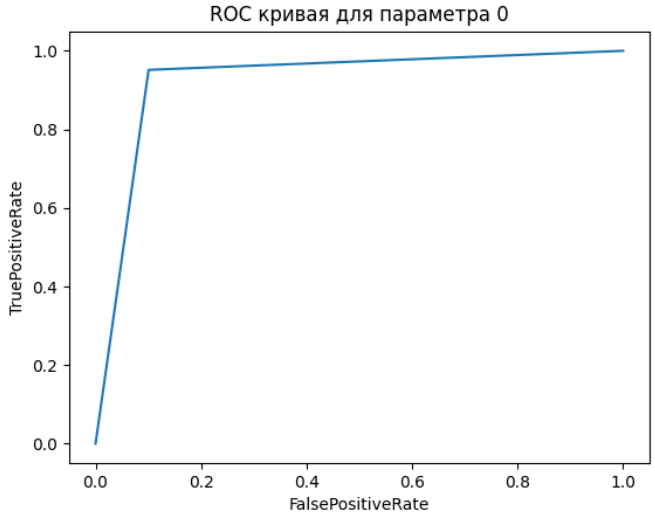

Рисунок 11 ROC кривая

Метрики для LogisticRegression

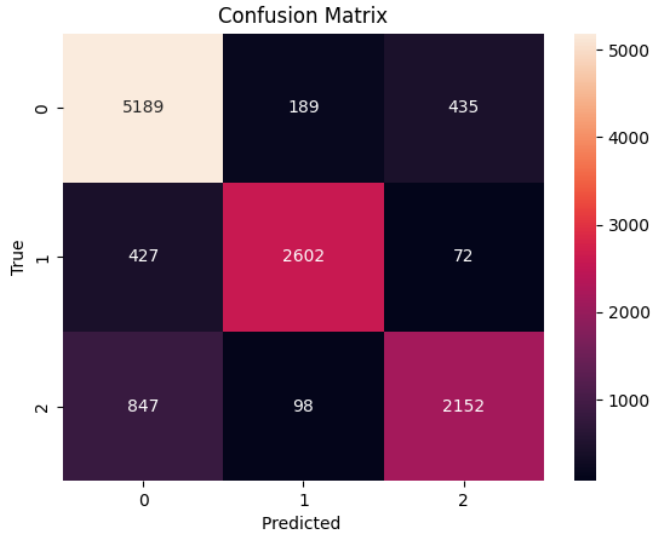

Рисунок 12 Confusion Matrix

	precision	recall	f1-score	support
0	0.80	0.89	0.85	5813
1	0.90	0.84	0.87	3101
2	0.81	0.69	0.75	3097
accuracy			0.83	12011
macro avg	0.84	0.81	0.82	12011
weighted avg	0.83	0.83	0.83	12011

roc_auc=0.8435521089221136

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

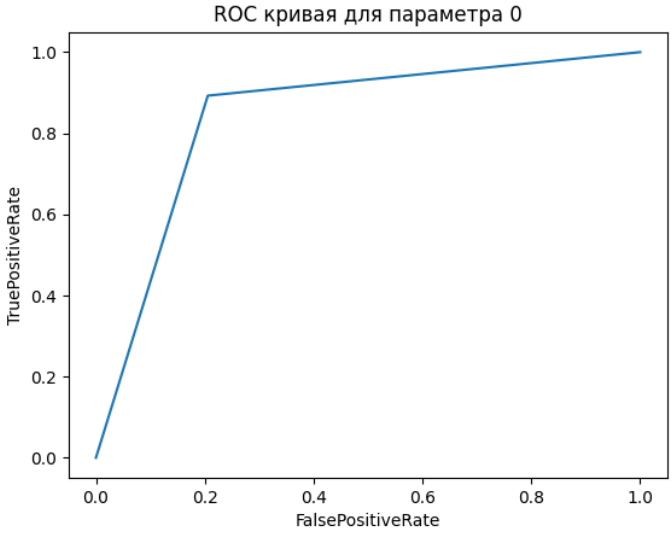

Рисунок 14 ROC кривая

Метрики для DecisionTree с параметром max_depth=11

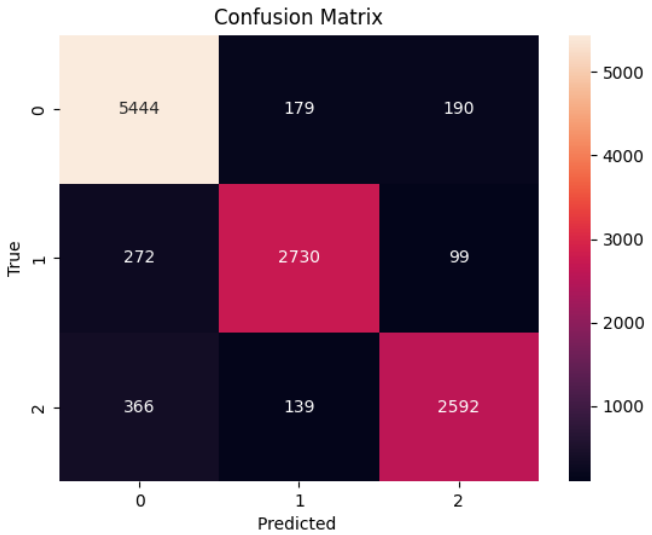

Рисунок 15 Confusion Matrix

	precision	recall	f1-score	support
0	0.90	0.94	0.92	5813
1	0.90	0.88	0.89	3101
2	0.90	0.84	0.87	3097
accuracy			0.90	12011
macro avg	0.90	0.88	0.89	12011
weighted avg	0.90	0.90	0.90	12011

roc_auc=0.9175293057193359

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

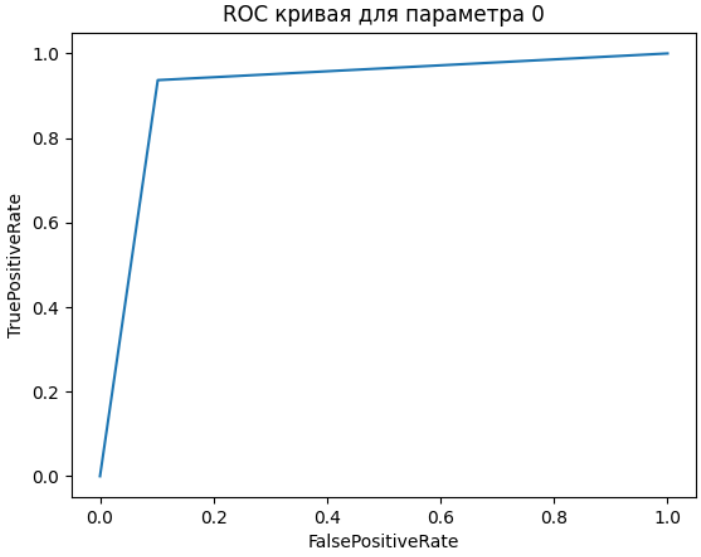

Рисунок 17 ROC кривая

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).