ELECTRICAL POWER ENGINEERING AND TRANSPORT AUTOMATION / ЭЛЕКТР ЭНЕРГЕТИКАСЫ ЖӘНЕ КӨЛІКТІ АВТОМАТТАНДЫРУ / ЭЛЕКТРОЭНЕРГЕТИКА И АВТОМАТИЗАЦИЯ ТРАНСПОРТА

Industrial Transport of Kazakhstan
ISSN 1814-5787 (print)
ISSN 3006-0273 (online)
Vol. 21. Is. 1. Number 81 (2024). Pp. 7–19
Journal homepage: https://prom.mtgu.edu.kz
https://doi.org/10.58420/ptk/2024.81.01.001

ANALYSIS OF MOBILE APPLICATION USER PREFERENCES BASED ON MACHINE LEARNING METHODS

D. Amrina*
International University of Transport and Humanities, Almaty, Kazakhstan.
E-mail: amrina.dana@mtgu.edu.kz

Dana Amrina — master student, International University of Transport and Humanities, Almaty, Kazakhstan.
E-mail: amrina.dana@mtgu.edu.kz, https:/orcid.org/0009-0000-8263-1962.

Abstract. This study examined classical machine learning methods and neural networks for analyzing user preferences in mobile applications. The target variable and preference criterion was the average app rating. A dataset from the open source Kaggle was used, followed by data cleaning and preprocessing. A comparative analysis was conducted on three classical machine learning methods (linear regression, random forest, XGBoost) and three neural network models (ANN, CNN, RNN) to predict users' average app ratings based on seven features. As the dataset was relatively small and of simple structure, some neural network models could not fully realize their potential. The XGBoost model demonstrated the best performance, highlighting its usefulness for this type of data. The CNN model performed slightly worse, as it is designed to capture significant patterns in complex datasets. The most important features for predicting user ratings were identified, including types, installations, genres, categories, and others. In future work on decision-making tasks aimed at improving user engagement, this study can assist in selecting appropriate models and input features to focus on when designing or enhancing an application.

Keywords: rating, application, machine learning, neural networks, prediction, data analysis

For citation: D. Amrina. Analysis of mobile application user preferences based on machine learning methods//Industrial Transport of Kazakhstan. 2024. Vol. 21. No. 81. Pp. 7–19. (In Russ.). https://doi.org/10.58420/ptk/2024.81.01.001

Conflict of interest: The authors declare that there is no conflict of interest.

МОБИЛЬДІ ҚОСЫМШАЛАР ПАЙДАЛАНУШЫЛАРЫНЫҢ ҚАЛАУЛАРЫН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ НЕГІЗІНДЕ ТАЛДАУ

Д. Амрина
Халықаралық көліктік-гуманитарлық университет, Алматы, Қазақстан.
E-mail: amrina.dana@mtgu.edu.kz

Дана Амрина — магистрант, Халықаралық көліктік-гуманитарлық университет, Алматы, Қазақстан.
E-mail: amrina.dana@mtgu.edu.kz, https:/orcid.org/0009-0000-8263-1962.

Аннотация. Осы жұмыста мобильді қосымшалардың пайдаланушылардың қалауларын талдау үшін классикалық машиналық оқыту және нейрондық желілер әдістері қарастырылды. Мақсатты көрсеткіш және қалаулар критерийі ретінде қосымшаның орташа рейтингі алынған. Ашық дереккөзі Kaggle-дан алынған датасет қолданылып, деректерді тазалау және алдын ала өңдеу жүргізілді. Орташа пайдаланушы баға көрсеткішін болжау үшін 7 параметр негізінде 3 классикалық машиналық оқыту әдісі (linear regression, random forest, XGBoost) және 3 нейрондық желі моделі (ANN, CNN, RNN) салыстырмалы талдаудан өткізілді. Датасет салыстырмалы түрде шағын және қарапайым құрылымға ие болғандықтан, кейбір нейрондық желі модельдері өз потенциалын толық ашпаған. Ең жақсы нәтижені XGBoost моделі көрсетті, бұл осы деректер түрінде аталған модельдің тиімділігін дәлелдейді. CNN моделі сәл төмен нәтижелер көрсетті, себебі ол күрделі деректердегі маңызды байланыстарды анықтауға арналған. Пайдаланушы рейтингісін болжауда ең маңызды сипаттамалар анықталды: типтер, орнатулар, жанрлар, категориялар және басқа факторлар. Болашақта пайдаланушылардың қатысуын арттыру саласындағы шешім қабылдау тапсырмаларымен жұмыс істегенде, бұл жұмыс тиісті модель мен кіріс сипаттамаларын анықтауға және қосымшаны жасау немесе жетілдіру кезінде назар аударуға көмектеседі.

Түйін сөздер: рейтинг, қосымша, машиналық оқыту, нейрондық желілер, болжау, деректерді талдау

Дәйексөздер үшін: Д. Амрина. Мобильді қосымшалар пайдаланушыларының қалауларын машиналық оқыту әдістері негізінде талдау//Қазақстан өндіріс көлігі. 2024. Том. 21. № 81. 7–19 бет. (Орыс тіл.). https://doi.org/10.58420/ptk/2024.81.01.001

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

АНАЛИЗ ПРЕДПОЧТЕНИЯ ПОЛЬЗОВАТЕЛЕЙ МОБИЛЬНЫХ ПРИЛОЖЕНИЙ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Д. Амрина
Международный транспортно-гуманитарный университет, Алматы, Казахстан.
E-mail: amrina.dana@mtgu.edu.kz

Д. Амрина — магистрант, Международный транспортно-гуманитарный университет, Алматы, Казахстан
E-mail: amrina.dana@mtgu.edu.kz, https:/orcid.org/0009-0000-8263-1962.

Аннотация. В данной работе были рассмотрены методы классического машинного обучения и нейронных сетей для анализа предпочтений пользователей мобильных приложений. В качестве целевого значения и критерия предпочтения был взят средний рейтинг приложения. Был использован датасет из открытого источника Kaggle, далее была проведена очистка, предобработка данных. Проведен сравнительный анализ 3 классических методов машинного обучения, среди которых linear regression, random forest, XGBoost, а также 3 модели нейронных сетей, из которых ANN, CNN, RNN, для прогнозирования

средних оценок пользователей приложения исходя из 7 признаков. Так как датасет относительно небольшой с простой структурой данных, некоторые модели нейронных сетей не смогли полностью раскрыть весь свой потенциал. Наилучшую производительность показала модель XGBoost, что показывает пользу данной модели в таком виде данных. Немного хуже показала модель CNN, так как она предназначена для выявления важных зависимостей в сложных данных. Были выявлены наиболее важные признаки, которые больше всего помогли в прогнозировании значения рейтинга пользователей, среди которых типы, установки, жанры, категории, и другие. В будущем, работая с задачей принятий решений в области улучшения вовлеченности пользователей, эта работа поможет в определении подходящей модели и входных признаков, на которые стоит обратить внимание при создании или улучшении приложения.

Ключевые слова: рейтинг, приложение, машинное обучение, нейронные сети, прогнозирование, анализ данных.

Для цитирования: Д. Амрина. Анализ предпочтения пользователей мобильных приложений на основе методов машинного обучения // Помышленный транспорт Казахстана. 2024. Т. 21. No. 81. Стр. 7–19. (На рус.). https://doi.org/10.58420/ptk/2024.81.01.001

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Современные цифровые системы активно используют методы глубокого обучения для анализа поведения пользователей и прогнозирования их действий (Bell, 2022: 207–216). В последние годы искусственный интеллект становится ключевым инструментом для оптимизации и автоматизации процессов в различных сферах, включая цифровые платформы и мобильные приложения (Sharifani, 2023: 12). Машинное обучение (ML) является одной из областей искусственного интеллекта и позволяет создавать системы, способные самостоятельно обучаться на основе данных без необходимости явного программирования (Sarker, 2021: 160).

Нейроны, как биологическая модель, адаптируют свои связи для предсказания будущей активности и минимизации ошибок, что позволяет оптимизировать обработку информации и повышать эффективность работы системы (Luczak, 2022: 62–72). Подобные принципы используются в ML для анализа поведения пользователя, что помогает выявлять эффективные и неэффективные элементы интерфейса, повышать персонализацию и удовлетворенность пользователей (Dina, 2021: 100462; Sharifani, 2022: 5).

Объектом исследования в данной работе являются цифровые приложения, доступные в Google Play Store, а предметом — механизмы мотивации пользователей через анализ факторов, влияющих на их оценки мобильных приложений (Google Play Store Apps, 2025). Проблемная ситуация заключается в том, что, несмотря на активное применение методов ML для прогнозирования рейтингов приложений, большинство исследований не дают полного понимания взаимосвязей между характеристиками приложений и поведением пользователей (Suleman, 2019: 57).

Актуальность темы определяется как теоретическим, так и практическим интересом к изучению цифровой вовлеченности пользователей. Прогнозирование рейтингов приложений позволяет не только оценить степень удовлетворенности пользователей, но и выявить потенциальные закономерности, которые могут быть использованы для персонализации интерфейсов и улучшения цифрового опыта (Yu, 2022: 26; Taye, 2023: 91).

Целью исследования является анализ предпочтений пользователей мобильных приложений и прогнозирование их рейтинга как показателя вовлеченности. Для достижения цели были поставлены следующие задачи:

- Провести анализ характеристик приложений, таких как категория, тип, контент-рейтинг, количество установок, цена, размер и жанр.

- Разработать модели машинного обучения и глубокого обучения для прогнозирования рейтингов приложений.

- Провести сравнительный анализ различных моделей и выявить наиболее значимые признаки, влияющие на рейтинг.

- Сделать выводы о возможностях персонализации интерфейсов и повышении вовлеченности пользователей на основе анализа данных.

Методы исследования включают использование библиотек TensorFlow и scikit-learn, реализацию нейронных сетей (ANN, CNN, RNN/LSTM), а также классических моделей машинного обучения (Random Forest, XGBoost, линейная регрессия) с оценкой качества моделей через метрики MAE, RMSE и R^2.

Гипотеза исследования заключается в том, что с помощью моделей глубокого обучения возможно выявить зависимость между характеристиками приложений и рейтингами пользователей, а также определить ключевые факторы, влияющие на вовлеченность пользователей, несмотря на ограниченность и неравномерность данных (Somers & Black, 2025: 056036).

Значение исследования заключается в возможности использования полученных результатов для оптимизации интерфейсов мобильных приложений, разработки персонализированных цифровых решений и повышения удовлетворенности пользователей.

Материалы и методы

В качестве материалов исследования использован открытый набор данных Google Play Store Apps, размещенный на платформе Kaggle (Google Play Store Apps, 2025). Датасет содержит информацию о мобильных приложениях, включая 13 признаков: категорию приложения, тип, контент-рейтинг, количество установок, цену, размер, жанры, дату последнего обновления, версию приложения и ряд других характеристик. После очистки данных от пропусков и аномальных значений в исследование было включено 9366 примеров. Выборка отражает различные категории приложений и разнообразие пользовательских предпочтений, что обеспечивает возможность анализа факторов, влияющих на оценку и вовлеченность пользователей.

Количественная характеристика материала представлена распределением данных по 7 ключевым признакам, использованным в модели регрессии: категория, тип, контент-рейтинг, установки, цена, размер, жанры. Качественный анализ включал проверку достоверности данных, идентификацию выбросов и несоответствий в категории и жанрах, что позволило повысить надежность дальнейших моделей прогнозирования.

Цель и гипотеза исследования. Цель исследования — анализ предпочтений пользователей мобильных приложений и прогнозирование рейтинга приложений как показателя вовлеченности пользователей. Гипотеза исследования заключается в том, что использование моделей глубокого обучения позволяет выявить зависимость между характеристиками приложений и пользовательским рейтингом, а также определить ключевые признаки, влияющие на вовлеченность пользователей (Somers & Black, 2025: 056036).

Вопросы исследования:

- Какие характеристики приложений оказывают наибольшее влияние на пользовательский рейтинг?

- Насколько эффективно модели машинного и глубокого обучения прогнозируют рейтинги приложений?

- Какие методы позволяют улучшить точность предсказания в условиях неравномерного распределения данных?

- Как выявленные закономерности могут быть использованы для персонализации интерфейсов приложений и повышения удовлетворенности пользователей?

Этапы исследования:

- Подготовка данных: очистка, нормализация и кодирование категориальных

признаков.

- Разделение выборки: 80 % данных использованы для обучения моделей, 20 % — для тестирования.

- Разработка моделей: реализованы нейронные сети (ANN, CNN, LSTM) и классические модели машинного обучения (Random Forest, XGBoost, линейная регрессия).

- Обучение и настройка моделей: оптимизация гиперпараметров, использование оптимизатора Adam и функции потерь MSE для нейронных сетей.

- Оценка качества моделей: расчет метрик MAE, RMSE и R^2 для сравнения точности предсказаний.

- Анализ важности признаков: выявление наиболее значимых характеристик приложений для прогнозирования рейтинга.

- Интерпретация результатов: выявление закономерностей, оценка влияния редких и неравномерно распределенных данных на точность моделей.

Методы машинного обучения:

- Линейная регрессия — используется для визуализации зависимостей и первичного анализа данных.

- Random Forest Regressor — баланс между качеством предсказаний и скоростью обучения, выявление важности признаков.

- XGBoost — градиентный бустинг для табличных данных, показавший наилучшую точность прогнозирования.

Методы глубокого обучения:

- ANN (Artificial Neural Network) — полносвязная нейронная сеть для выявления сложных нелинейных зависимостей.

- CNN (Convolutional Neural Network) — использование сверток для выделения локальных особенностей признаков.

- LSTM (Long Short-Term Memory, RNN) — анализ последовательностей признаков для прогнозирования рейтинга.

- Предварительная обработка данных: нормализация числовых признаков, one-hot кодирование категориальных признаков, удаление пропусков и аномалий.

- Оценка моделей: метрики RMSE (Root Mean Square Error), MAE (Mean Absolute Error) и коэффициент детерминации R^2, позволяющие объективно оценить точность прогнозов и объясняемость моделей.

Применение данных методов позволяет выявить закономерности в оценках приложений, определить значимые признаки, влияющие на рейтинг, и предложить рекомендации для повышения вовлеченности пользователей на цифровых платформах. Новизна исследования заключается в сравнительном анализе традиционных и глубоких моделей на одном наборе данных и выявлении специфики прогнозирования рейтинга с учетом редких и неравномерно распределенных категориальных признаков.

Для анализа данных применяются методы машинного обучения с использованием библиотек TensorFlow и scikit-learn. Реализованы модели: нейронная сеть с архитектурой ANN (artificial neural network), CNN (convolutional neural network), RNN (recurrent neural network), а также для сравнения использованы модель Random Forest Regressor, XGBoost, и линейная регрессия. Для работы с данными была проведена очистка и нормализация данны. Также было произведено разделение данных на обучающую для тренировки нейронной сети и тестовую для произведения прогнозов. Качество моделей оценивалось с помощью метрик MAE, RMSE и R^2.

Датасет использовался из Kaggle, под названием Google Play Store Apps (Google Play Store Apps, 2025). После очистки данных от пустых значений, вышло 9366 примеров с 13 признаками. Далее, датасет делился на 80% обучающиеся данные, и 20% на тестовые данные.

В данной работе у нас решается проблема регрессии, так как будет предсказываться

средний рейтинг приложения, поставленный пользователями в пределах от 1 до 5, по 7 отобранным признакам (категория, тип, контент-рейтинг, установки, цена, размер, жанры), которые будут полезны для нашей задачи регрессии. Для нейронных сетей использовался оптимизатор “adam”, а критерий потери – “mse”.

Ниже приведены архитектуры 3-х разных нейронных сетей, а после них классические модели машинного обучения.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	512
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 16)	528
dense_3 (Dense)	(None, 1)	17

Рис. 1. Архитектура ANN

В качестве простой модели нейронной сети был взят ANN, где используется полносвязная нейронная сеть (MLP) (Рис. 1). Общее количество параметров модели составляет 3137.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 6, 32)	96
conv1d_1 (Conv1D)	(None, 5, 16)	1,040
flatten (Flatten)	(None, 80)	0
dense_4 (Dense)	(None, 16)	1,296
dense_5 (Dense)	(None, 1)	17

Рис. 2. Архитектура CNN

В качестве простой модели был взят CNN с количеством параметров 2449. Данная нейронная сеть использует свертки для поиска локальных признаков. (Рис. 2).

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 32)	4,352
dense_6 (Dense)	(None, 16)	528
dense_7 (Dense)	(None, 1)	17

Рис. 3. Архитектура RNN

В качестве RNN был взял LSTM, где модель предсказывает рейтинг приложения по 7 признакам, интерпретируемым как последовательность длиной 7 (Рис. 3). Общее количество параметров - 4897.

В качестве классических методов машинного обучения были приведены линейная регрессия, Random Forest, XGBoost.

Линейная регрессия обычно используется для визуализации данных, и не всегда идеально подходит для задачи регрессии. Random Forest показывает баланс между скоростью обучения и качеством. XGBoost подходит для табличных данных.

На основе данных моделей мы проведем сравнительный анализ, выявим преимущества и недостатки, и выберем важные признаки для предсказаний моделей.

Результаты и обсуждение

ML зарекомендовало себя как устойчивое и развивающееся направление, однако его развитие не является полностью прямолинейным. Машинное обучение основано на выведении закономерностей из исторических наблюдений, что не гарантирует точного вывода из данных. Также, как описано в статье Barbierato E. и Gatti A. (Barbierato, 2024), можно дать прогнозы, но без объяснений причинно-следственных связей, стоящими за ними, что и снижает доверие со стороны специалистов. Тем не менее, постоянно разрабатываются подходы для компенсирования недостатков, что позволяет развивать точность и внедрять в практическом плане. Ради оценки того или иного подхода, необходимы исследования и выявления недостатков разных подходов, или же, наоборот, их достоинств.

Yu J. (Yu, 2022) провели обзор 43 исследований, в которых использовались модели CNN и RNN/LSTM в области умных домов. Они обнаружили, что наибольшее внимание занимают проблемы мониторинга активности, безопасности и управления энергией. Также обсуждалось, что CNN модели больше подходят при работе с изображениями и видео, в таких типах данных данная модель показывает наибольшую эффективность. А RNN/LSTM модель больше подходит под анализ данных, отличных от CNN. Авторы прибегают к использованию обеих моделей (CNN+LSTM). Это подтверждает эффективность комбинированного подхода при анализе данных разной природы. Данный вывод может помочь при определении лучших моделей для использования анализов под другие сферы, например, при анализе данных удовлетворенности приложением.

Тем не менее, обзор состоит только из 43 статей и ограничен временным рамками (2016–2020), но в последнее время технологии, визуальные модели, значительно изменились. Кроме того, авторы отмечают, что большинство исследований не рассматривают вопросы конфиденциальности и морали использования данных в умных домах. Помимо этих потенциальных недостатков машинного обучения для предсказывания результатов по датасету, отметили эти проблемы в статье Zhu M. (Zhu, 2022: 107–116) подчеркивая, что сбор точных и больших данных является трудным, но, тем не меннее, необходимым для точности результатов. Помимо этого, по их работе было выявлено, что алгоритмы работают только в узких случаев и не учитывают всех сложностей и тонкостей работы с вычислением качества воды. Рассматривая другую статью Sahu S.K. (Sahu, 2023: 1956) в попытке прогнозирования фондового рынка заключили, что помощью машинного и глубокого обучения возможно выявить закономерности только частично, но результаты всё равно остаются нестабильными. Чему причинами в данном случае выступают ограниченность данных и высокая изменчивость фондового рынка.

Suleman M. (Suleman, 2019: 57–61) проводили исследования также по найденному датасету в Kaggle. В их работе авторы провели исследования по предсказанию приложений в Google Play Store, с использованием алгоритмов машинного обучения фокусируясь на факторы влияния на рейтинги приложений с ипользованием регрессионной модели. Датасет состоит из 10000 приложений с разделением данных на 75 % для обучения и 25 % данных для прогнозирования и тестинга. В работе были применены алгоритмы MATLAB 2018 (Regression Learner app), включая Regression Trees (Fine, Medium, Coarse), SVM (Linear, Quadratic, Cubic, Gaussian), Ensemble methods и Gaussian Process Regression. В качестве оценки результатов авторы использовали RMSE, R-Squared, MSE, MAE. По их заключению Fine Tree точнее показывает результаты (RMSE 0.33, R-Squared 0.52). На основе ихисследований, можно заключить, что предсказать рейтинг с помощью машинного обучения возможно со средней долей вероятности.

Способность автоматически извлекать сложные представления из больших объемов данных отличает глубокое обучение от традиционного машинного обучения Taye M.M.

(Taye, 2023: 91). Это область искусственного интеллекта, которая быстро развивается, поэтому глубокое обучение лучше подходит для анализа найденного датасета и оно будет браться за основу в методологии. Так, нейронные сети уже показали хорошии результаты в исследовании Somers A. и Black B.J в медицинском применении (Somers, 2025: 056036), но с большим датасетом.

Исходя из матрицы корреляций (Рис. 4), видно, что сильных зависимостей между признаками нет, помимо зависимости между жанром и категорией, где корреляция имеет значение 0.78. Все признаки почти не коррелируют с рейтингом. Это говорит о том, что линейные модели и ANN будут слабыми.

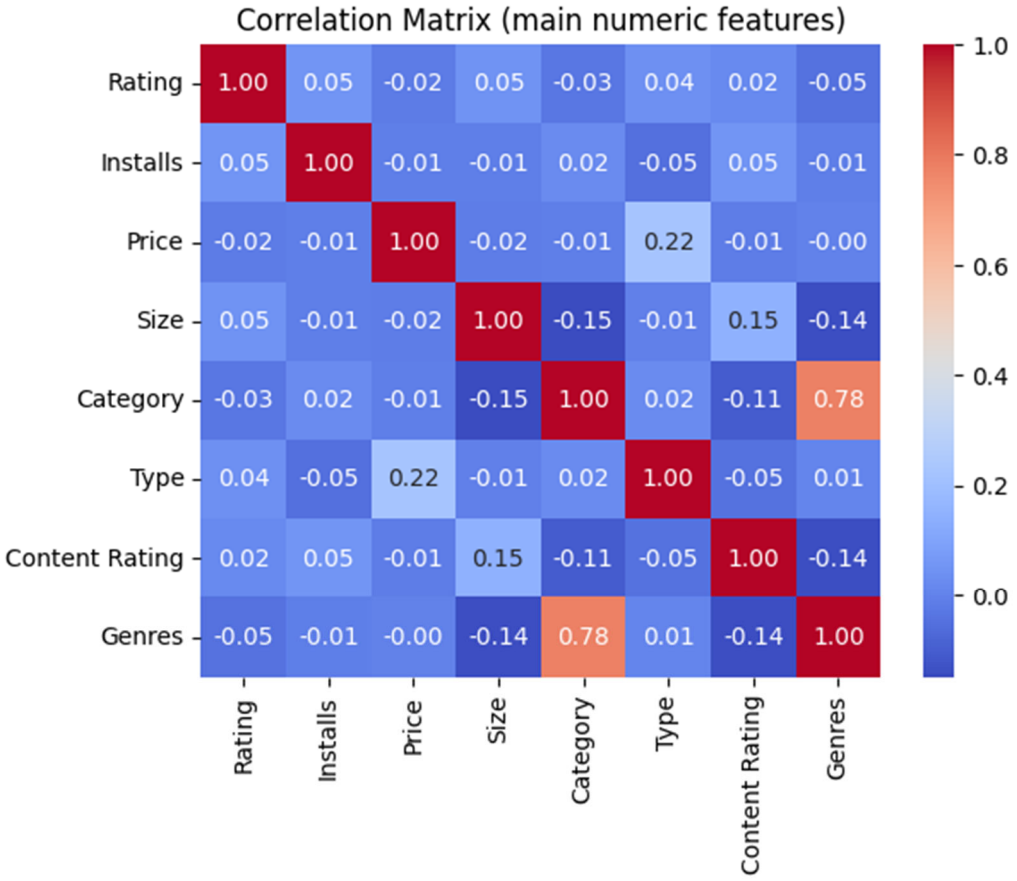

Рис. 4. Матрица корреляций признаков

Поэтому модели, которые хорошо справляются с нелинейными зависимостями, подойдут лучше всего.

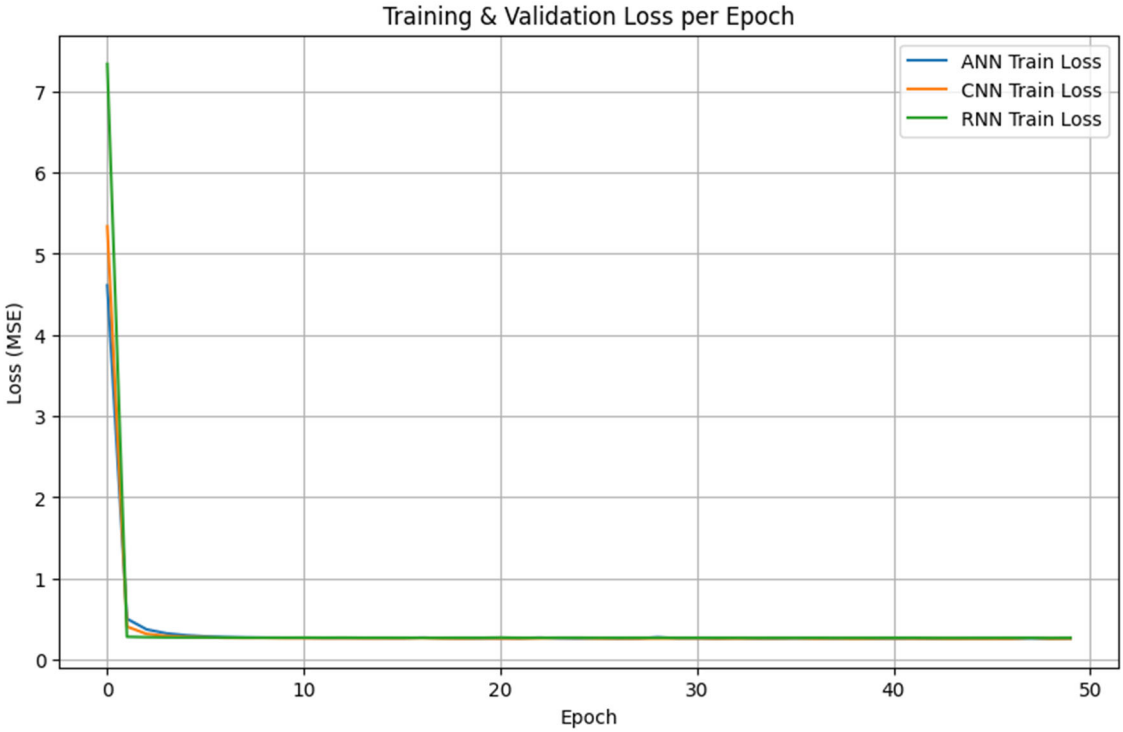

Рис. 5. История обучения ANN, CNN, RNN

Поскольку данные имели табличный и относительно простой характер, в отличие от изображений и временных рядов, все модели продемонстрировали схожие кривые обучения (Рисунок 5). Для тренировки хватило бы 5 эпох.

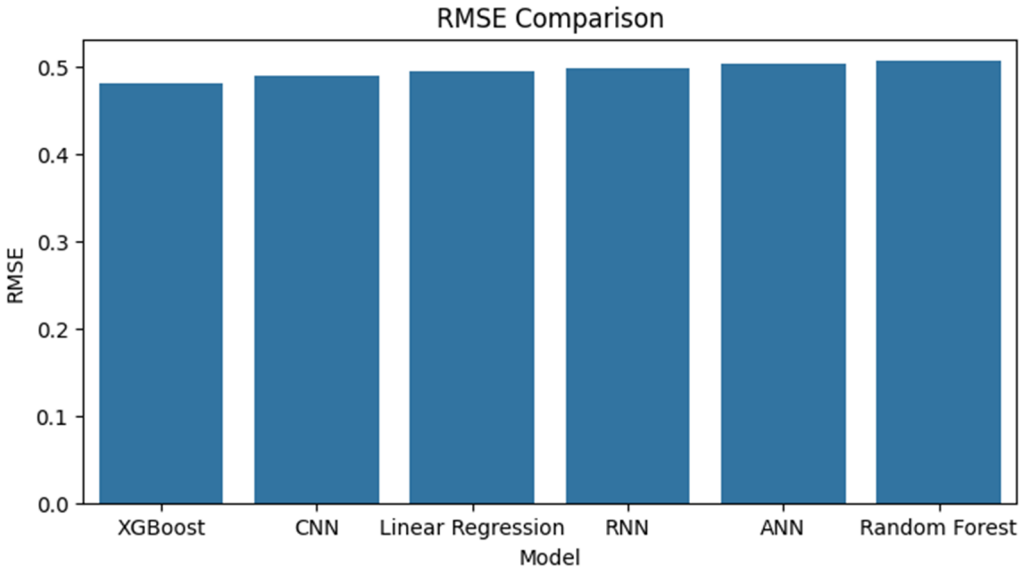

Рис. 6. Сравнение среднеквадратических ошибок моделей

Исходя из Рисунка 6, XGBoost показал наилучший, хотя и незначительно, результат по RMSE.

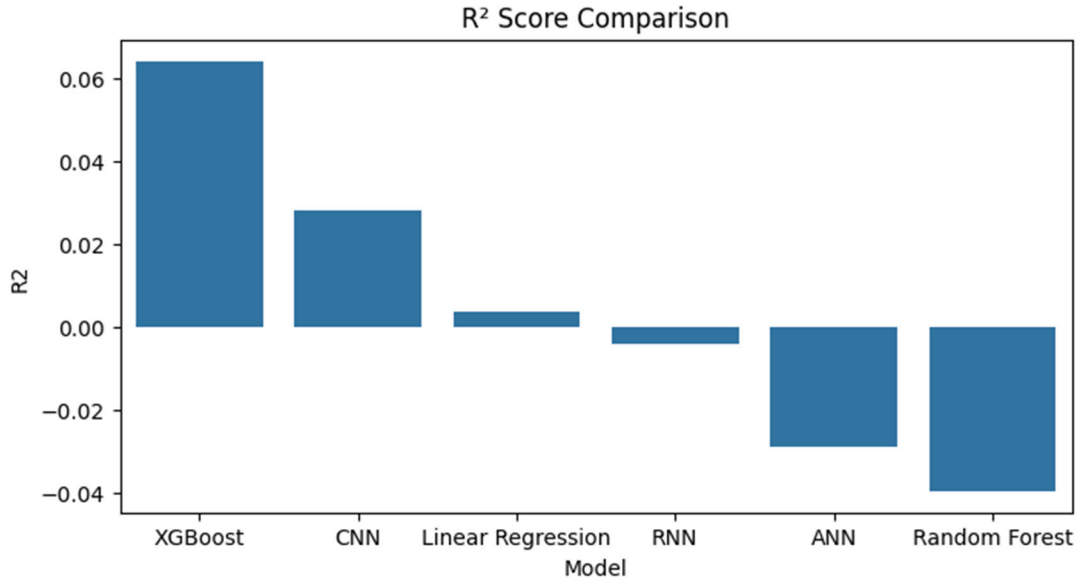

Рис. 7. Сравнение коэффициента детерминации

Исходя из Рисунка 7, XGBoost модель, которая лучше объясняет дисперсию, CNN - средний результат, тем временем как остальные модели хуже. Random Forest - хуже всего из-за переобучения.

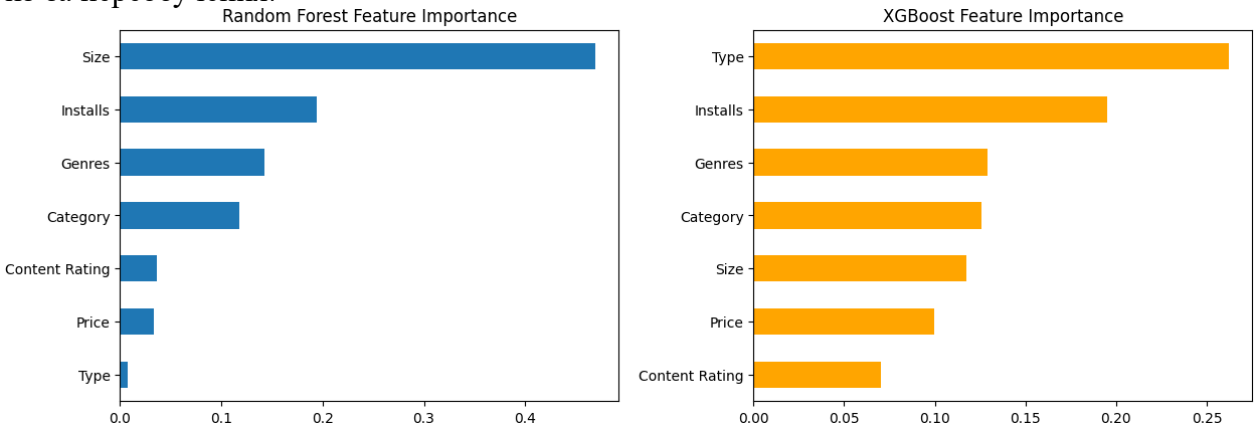

Рис. 8. Feature importance of Random Forest and XGBoost

Были проведены сравнения важностей особенностей Random Forest и XGBoost (Рисунок 8). Исходя из метрик RMSE и R2, нам стоит доверять XGBoost, что говорит о том, что такие признаки, как “type”, “installs”, “genres”, “category”, и другие, указанные в графике, больше всего подходят для использования в моделях машинного обучения.

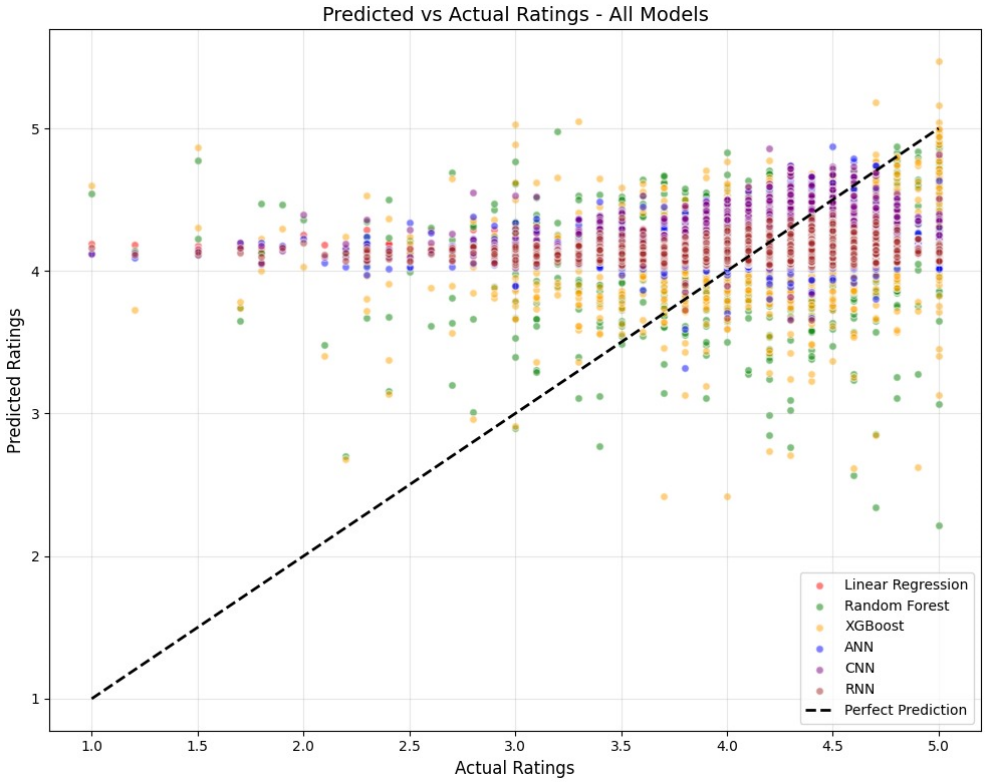

Рис. 9. Фактические и прогнозируемые результаты

Анализ по Рисунку 9 показал, что наличие редких и неравномерно распределённых данных в датасете негативно влияет на способность нейронной сети корректно обучаться. Модель выдаёт усреднённые предсказания и способна точно прогнозировать только рейтинги, находящиеся в диапазоне средних значений. Обнаружено, что для повышения точности требуется расширение выборки и балансировка категориальных данных.

Заключение

В ходе проведённого исследования был реализован комплекс моделей машинного и глубокого обучения для анализа предпочтений пользователей мобильных приложений и прогнозирования их рейтингов на основе данных Google Play Store. Используемый датасет включал 9366 примеров с 13 признаками, отражающими характеристики приложений, такие как категория, тип, контент-рейтинг, количество установок, цена, размер и жанры. Для обработки данных были проведены очистка, нормализация числовых признаков и one-hot кодирование категориальных признаков, что позволило повысить качество входных данных для моделей.

Реализованные модели включали классические алгоритмы машинного обучения — линейную регрессию, Random Forest и XGBoost, а также нейронные сети: ANN, CNN и LSTM. Выбор таких моделей был обусловлен целью сравнения эффективности традиционных и глубоких моделей при прогнозировании рейтингов, а также выявлением их преимуществ и ограничений в условиях работы с табличными данными и неравномерно распределёнными категориальными признаками.

Анализ матрицы корреляций показал, что большинство признаков не имеют сильной линейной зависимости с рейтингом, за исключением высокой корреляции между жанром и категорией приложения ($r = 0.78$). Это указывает на слабую предсказательную силу линейных моделей и необходимость применения алгоритмов, способных выявлять сложные нелинейные зависимости, таких как XGBoost или нейронные сети. Результаты обучения моделей показали, что при относительно простых табличных данных кривые обучения ANN, CNN и LSTM были схожи, что подтверждает возможность использования базовых архитектур нейронных сетей для предварительного анализа данных.

Сравнительный анализ моделей по метрикам RMSE и коэффициенту детерминации R^2 выявил, что наилучшие результаты показала модель XGBoost. Random Forest, несмотря на свою способность выявлять важность признаков, продемонстрировал худший результат из-за переобучения, а линейная регрессия и ANN показали ограниченную предсказательную силу. CNN продемонстрировала средние показатели, что объясняется отсутствием сложных локальных структур в табличных данных, где свёрточные операции не дают значительного преимущества.

Особое внимание было уделено анализу важности признаков (feature importance). По результатам XGBoost и Random Forest, наибольшее влияние на рейтинг оказали такие признаки, как тип приложения, количество установок, жанры и категория. Эти признаки могут служить основой для формирования стратегий персонализации интерфейса и улучшения пользовательского опыта. В частности, понимание того, какие категории или жанры наиболее востребованы, может помочь разработчикам оптимизировать контент и функциональные возможности приложений, повышая удовлетворенность пользователей.

Анализ фактических и прогнозируемых результатов выявил, что нейронные сети в условиях неравномерного распределения данных имеют тенденцию выдавать усреднённые предсказания. Модель способна точно прогнозировать лишь средние значения рейтингов, что ограничивает её применимость для предсказания экстремально высоких или низких оценок. Этот вывод подтверждает необходимость расширения объёма данных, балансировки категориальных признаков и разработки методов работы с редкими значениями для повышения точности глубоких моделей.

Сравнение результатов с предыдущими исследованиями (Suleman, 2019; Yu, 2022; Somers, 2025) показало, что использование алгоритмов глубокого обучения для прогнозирования рейтингов приложений является оправданным, особенно при анализе сложных и нелинейных зависимостей. Однако, как и в исследованиях других авторов, точность моделей ограничена характером данных и их распределением, а также отсутствием дополнительных факторов, которые могут влиять на пользовательские оценки, таких как пользовательские отзывы, время использования приложения и качество обновлений.

Таким образом, ключевые выводы исследования включают:

- Эффективность моделей: XGBoost показал наилучшие результаты при прогнозировании рейтингов приложений на табличных данных. Нейронные сети могут быть полезны, но их точность ограничена из-за малой выборки и неравномерности данных.
- Важность признаков: Наибольшее влияние на рейтинг оказали признаки типа приложения, количество установок, категория и жанры, что может быть использовано для персонализации интерфейсов и повышения вовлеченности пользователей.
- Ограничения: Основными ограничениями исследования являются небольшое количество данных для редких категорий, неравномерное распределение признаков и отсутствие дополнительных факторов, которые могли бы усилить точность моделей.

Практическая значимость: Результаты исследования могут быть использованы разработчиками мобильных приложений для оптимизации пользовательского интерфейса, определения приоритетных категорий и жанров, а также для разработки систем персонализированных рекомендаций. Прогнозирование рейтингов позволяет заранее оценить потенциальную популярность приложения и скорректировать стратегию его продвижения.

Направления для дальнейших исследований: Для повышения точности прогнозирования рекомендуется: расширить выборку данных, включив более редкие и новые приложения; учитывать дополнительные факторы, такие как отзывы пользователей, частота обновлений и поведение пользователей внутри приложения; использовать методы балансировки данных для категориальных признаков; исследовать возможности ансамблирования моделей глубокого и классического обучения.

В целом, проведённая работа демонстрирует, что использование моделей глубокого обучения и современных алгоритмов машинного обучения предоставляет возможности для более глубокого понимания поведения пользователей и прогнозирования их оценок мобильных приложений. Несмотря на ограничения, выявленные закономерности и важные признаки могут стать основой для разработки более персонализированных, эффективных и привлекательных цифровых продуктов, что способствует повышению вовлеченности пользователей и улучшению пользовательского опыта на мобильных платформах.

ЛИТЕРАТУРА

Bell, 2022 — Bell J. What Is Machine Learning? // Machine Learning and the City. — John Wiley & Sons, Ltd. — 2022. — Pp. 207–216. [Eng.]

Luczak, 2022 — Luczak A., McNaughton B.L., Kubo Y. Neurons learn by predicting future activity // Nat. Mach. Intell. — Nature Publishing Group, — 2022. — Vol. 4. — №1. — Pp. 62–72. [Eng.]

Sharifani, 2023 — Sharifani K., Amini M. Machine Learning and Deep Learning: A Review of Methods and Applications. // World Information Technology and Engineering Journal. — 2023. — Volume 10. — Issue 07. — Pp. 3897–3904. [Eng.]

Dina, 2021 — Dina A.S., Manivannan D. Intrusion detection based on Machine Learning techniques in computer networks // Internet Things. — 2021. — Vol. 16. — P. 100462. [Eng.]

Sharifani, 2022 — Sharifani K., Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model. // International Journal of Science and Information System Research. — 2022. — Volume 12. — Issue 9. — Pp. 20–44. [Eng.]

Sarker, 2021 — Sarker I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions // SN Comput. Sci. — 2021. — Vol. 2 . — №3. — P. 160. [Eng.]

Janiesch, 2021 — Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning // Electron. Mark. — 2021. — Vol. 31. — №3. — Pp. 685–695. [Eng.]

Barbierato, 2024 — Barbierato E., Gatti A. The Challenges of Machine Learning: A Critical Review // Electronics. — Multidisciplinary Digital Publishing Institute. — 2024. — Vol. 13. — №2. — Pp. 416. [Eng.]

Yu, 2022 — Yu J., de Antonio A., Villalba-Mora E. Deep Learning (CNN, RNN) Applications for Smart Homes: A Systematic Review // Computers. — Multidisciplinary Digital Publishing Institute. — 2022. — Vol. 11. — №2. — P. 26. [Eng.]

Zhu, 2022 — Zhu M., A review of the application of machine learning in water quality evaluation // Eco-Environ. Health. — 2022. — Vol. 1. — №2. — Pp. 107–116. [Eng.]

Sahu, 2023 — Sahu S.K., Mokhade A., Bokde N.D. An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges // Appl. Sci. — Multidisciplinary Digital Publishing Institute, — 2023. — Vol. 13. — №3. — P. 1956. [Eng.]

Suleman, 2019 — Suleman M., Malik A., Hussain S.S. Google play store app ranking prediction using machine learning algorithm // Proceedings of the International Conference on Data Science 2019, 7-9 February. — 2019. — Pp. 57–61. [Eng.]

Taye, 2023 — Taye M.M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions // Computers. — Multidisciplinary Digital Publishing Institute. — 2023. — Vol. 12. — №5. — P. 91. [Eng.]

Somers, 2025 — Somers A., Black B.J. Co-cultured sensory neuron classification using extracellular electrophysiology and machine learning approaches for enhancing analgesic screening // J. Neural Eng. — IOP Publishing. — 2025. — Vol. 22. — №5. — P. 056036. [Eng.]

Google Play Store Apps, 2025 — Google Play Store Apps [Electronic resource]. — URL: https://www.kaggle.com/datasets/lava18/google-play-store-apps (accessed: 20.08.2025). [Eng.]