**Разработка рекомендательной системы фильмов на основе нейросети и кластеризации**

**М. А. Субоч**

*Аннотация: Статья посвящена разработке рекомендательной системы на основе нейросетевой модели и алгоритма кластеризации KMeans. Основной целью является улучшение персонализации выбора фильмов на платформе для просмотра контента. В статье рассмотрены методы сбора данных, процесс обучения моделей и обоснование выбора используемых алгоритмов.*

*Введение.* В эпоху цифровизации и массового потребления контента системы персонализированных рекомендаций стали неотъемлемой частью многих онлайн-сервисов. Особенно это актуально для платформ, предлагающих широкий ассортимент фильмов, где пользователи сталкиваются с проблемой выбора из огромного каталога. Для пользователей становится сложной задачей быстро найти фильмы, которые соответствуют их интересам и предпочтениям. Это может привести к усталости от выбора (choice fatigue), что отрицательно сказывается на их вовлеченности и удовлетворении от сервиса.

Для решения этой задачи была разработана рекомендательная система, использующая нейронные сети и кластеризацию методом KMeans. Нейросетевые модели позволяют выявить скрытые закономерности в предпочтениях пользователей, а кластеризация помогает быстро формировать группы фильмов с похожими характеристиками. Совместное использование этих методов обеспечивает высокую точность рекомендаций и помогает пользователю быстро находить интересный контент.

*Актуальность исследования* подтверждается тем фактом, что системы рекомендаций позволяют существенно повысить уровень пользовательского опыта, увеличив вовлеченность и удержание аудитории. Применение нейросетей в связке с алгоритмами кластеризации обеспечивает более точный и разнообразный подбор фильмов, что особенно важно в условиях высокой конкуренции между платформами контент-сервисов.

*Цель исследования* – спроектировать и реализовать рекомендательную систему для сайта с фильмами, используя обученную нейронную сеть для формирования эмбеддингов фильмов и алгоритм KMeans для создания модели на основе алгоритма кластеризации данных.

*Объектом исс*ледования выступает рекомендательные системы на основе машинного обучения.

*Предметом исследования* данной работы являются методы построения рекомендательных систем на основе нейросетей и алгоритмов кластеризации данных.

*Постановка задачи***.** Для реализации рекомендательной системы необходимо решить следующие задачи:

1. Разработка архитектуры системы: построение структуры взаимодействия между нейросетевой моделью, алгоритмом кластеризации и базой данных.
2. Создание и обучение нейросетевой модели направленной на формирование эмбеддингов фильмов на основе их жанровых признаков.
3. Реализация алгоритма кластеризации на основе KMeans.
4. Формирование базы данных для хранения информации о пользователях, их предпочтениях и фильмах.
5. Разработка механизма персонализации рекомендаций на основе эмбеддингов пользователя и результатов кластеризации с применением модели XGBoost.
6. Веб-интерфейс на основе EF создаёт удобный способ получения персонализированных рекомендаций.
7. Завершающим этапом является тестирование системы, которое включает оценку качества модели с помощью метрик.

В современных системах рекомендаций широко применяются алгоритмы машинного обучения, включая нейросетевые подходы и методы кластеризации. Нейросетевые модели позволяют эффективно выделять сложные зависимости между предпочтениями пользователей и характеристиками фильмов, а алгоритмы кластеризации успешно группируют элементы по сходным признакам, что ускоряет процесс выдачи рекомендаций.

Современные платформы потокового видео предлагают обширный каталог фильмов, что приводит к необходимости эффективных систем рекомендаций. Существует несколько подходов к реализации таких систем: фильтрация на основе контента, коллаборативная фильтрация и гибридные методы. Каждый подход имеет свои плюсы и минусы.

Коллаборативная фильтрация – это метод, который рекомендует товары или услуги, анализируя поведение и предпочтения пользователей, исходя из предположения, что пользователям со схожими вкусами понравятся похожие товары. Например, если двум пользователям в прошлом нравились похожие фильмы, система будет рекомендовать фильмы, которые понравились одному из них, другому пользователю, предполагая, что они тоже им понравятся, исходя из их схожих предпочтений. [1]

Но у коллаборативной модели есть серьёзные недостатки:

* Неустойчивость к «холодному старту» – когда приходят новые пользователи или добавляются новые фильмы, модель не может сразу дать рекомендации, так как нет истории оценок.
* Низкая объясняемость – коллаборативная фильтрация просто опирается на матрицу взаимодействий, но не учитывает содержательные признаки (жанры, описания).
* Масштабируемость – при большом числе пользователей и фильмов коллаборативная модель становится тяжёлой по вычислениям, так как требует постоянного пересчёта схожестей.

Поэтому комбинация нейросетей и кластеризации показывает высокую эффективность. Нейросетевые модели позволяют формировать эмбеддинги фильмов и предпочтений пользователей, что упрощает процесс поиска схожих элементов. Эмбеддинг (от англ. embedding) — это процесс преобразования слов или текста в набор чисел – числовой вектор. Векторы можно сравнивать между собой, чтобы определить насколько два текста или слова похожи по смыслу. [2] С помощью эмбеддингов, можно рекомендовать даже новые фильмы без историй оценок. Эти эмбеддинги могут быть эффективно обработаны алгоритмом KMeans, который обеспечивает быструю группировку объектов по их характеристикам.

Также Кластеризация KMeans снижает вычислительную нагрузку: вместо поиска по всей базе система ищет рекомендации в пределах релевантного кластера.

*Обоснование выбора моделей.* Для реализации рекомендательной системы были выбраны следующие модели:

1. Нейросеть для генерации эмбеддингов: Этот выбор обусловлен способностью нейросетевых моделей выявлять сложные зависимости в данных для последующей кластеризации.

2. *KMeans-кластеризация*: Данный алгоритм выбран по нескольким причинам. Во-первых, KMeans отлично справляется с задачами группировки данных по сходным признакам, что позволяет формировать кластеры фильмов с похожими характеристиками. Это существенно ускоряет процесс подбора рекомендаций, так как поиск фильмов ведётся не по всей базе данных, а по выбранному кластеру. Во-вторых, KMeans эффективно работает с эмбеддингами, поскольку они обеспечивают равномерное распределение данных в многомерном пространстве, что улучшает качество кластеризации. Более того, KMeans обладает высокой скоростью работы при больших объёмах данных и не требует значительных вычислительных ресурсов.

3. Выбор модели *XGBoost*: Для предсказания пользовательских оценок фильмов рассматривались два подхода: линейная регрессия и градиентный бустинг XGBoost. Линейная регрессия показала хорошие результаты, однако при сравнении метрик качества моделей (MSE, MAE, MAPE, R²) XGBoost продемонстрировал значительно более высокую точность. Это связано с тем, что XGBoost способен учитывать сложные нелинейные зависимости между признаками и более устойчив к переобучению благодаря встроенной регуляризации. На гистограмме №1 представлено сравнение метрик качества алгоритмов Linear Regression и XGBoost:

Гистограмма №1 – Сравнение метрик качества алгоритмов Linear Regression и XGBoost.

Также коэффициент детерминации, который позволяет измерить какая доля вариативности зависимой переменной объясняется независимыми переменными, в алгоритме Linear Regression равен 98.61%, а в XGBOOST 99.78%.

Исходя из этого можно сделать заключение: среднеквадратичная ошибка (MSE) показывает среднее значение квадрата отклонений предсказаний от фактических значений — она особенно чувствительна к большим ошибкам, и у XGBoost этот показатель существенно ниже, чем у Linear Regression. Средняя абсолютная ошибка (MAE) измеряет среднее абсолютное отклонение предсказаний, показывая, насколько в среднем ошибается модель без учёта направления ошибки, здесь XGBoost также точнее. Средняя абсолютная процентная ошибка (MAPE) отражает средний процент отклонения предсказаний от реальных значений у XGBoost данный показатель равен 1.42%, в сравнении с 3.02% в алгоритме LinearRegression. Коэффициент детерминации (R2R^2R2) показывает долю вариации целевой переменной, объяснённой моделью, и почти 100% у XGBoost указывает на его высокую объясняющую способность. Таким образом, на основе сравнения моделей по метрикам качества был сделан выбор в пользу XGBoost как основной модели для предсказания оценок пользователей.

*Сбор данных.* Датасет был собран из открытых источников, включающих следующие данные:

* Название фильма
* Жанр
* Пользовательская оценка фильма
* Время оценки фильма

Процесс очистки данных включал несколько этапов и видоизменялся в зависимости от вида модели.

В первую очередь были удалены дубликаты фильмов, чтобы избежать многократного учета одной и той же информации. Далее проводилась фильтрация по рейтингу: были исключены фильмы с крайне низкими рейтингами, чтобы уменьшить влияние аномальных данных. Также проводилась нормализация жанров: приведены все жанры к единому стандарту написания.

1.Обработка датасета для обучения нейронной сети создания эмбеддингов на основе жанров включала в себя следующую последовательность действий: из всего набора данных были выделены уникальные жанры, отсортированные в алфавитном порядке. Затем каждому жанру было присвоено уникальное числовое значение (ID), что позволило преобразовать строковые описания жанров в числовые последовательности.

Поскольку фильмы могут иметь различное количество жанров, последовательности были приведены к единой длине с помощью дополнения (padding). На выходе была получена матрица признаков X, эти данные и послужили основой для обучения нейронной сети. На рисунке 1 представлен результат, использования обученной нейронной сети для генерации эмбеддингов:



Рисунок 1 – Подготовленный датасет после использования нейронной сети для генерации эмбеддингов.

2. Полученный датасет использовался для обучения модели кластеризации Kmeans, а именно, на основе столбца «embedding» производился процесс кластеризации. На рисунке 2 представлен датасет после обработки моделью Kmeans:



Рисунок 2 – Подготовленный датасет после использования модели Kmeans.

3. Подготовка обучающего датасета для модели XGBboost включала в себя:

* Изменение столбца «время оценки фильма». Создание дополнительных столбцов: месяц, день, день недели, а также часть дня для более чёткого предсказания пользовательских оценок.
* Создание дополнительных признаков в виде метрик: средние оценки по фильму, количество оценок, фильма среднее отклонение оценки пользователя.
* Использование One-hot кодирования жанров для создания более глубокой предсказывающей модели.

Визуальная демонстрация подготовленного датасета для модели XGBoost представлена на рисунке 3.

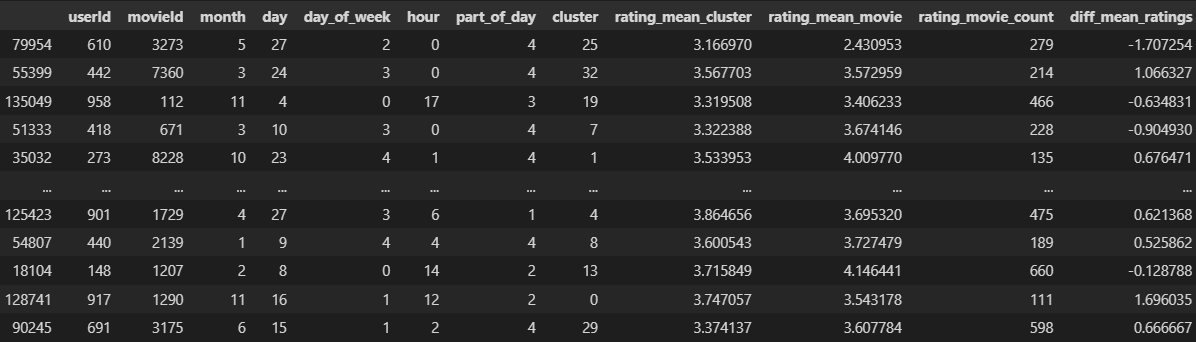


Рисунок 3 – Подготовленный датасет для модели XGBoost

*Обучение модели Kmeans*. После получения эмбеддингов фильмов с помощью нейросетевой модели следующим этапом стала кластеризация данных для формирования тематически схожих групп фильмов. Для этого использовался алгоритм KMeans из библиотеки scikit-learn. Перед обучением модель была протестирована на разных значениях количества кластеров (k) с целью определения оптимального числа групп. Для оценки качества кластеризации использовались следующие метрики:

* Silhouette Score — показывает, насколько хорошо объекты соответствуют своим кластерам и насколько плохо — другим.
* Davies–Bouldin Score — оценивает внутрикластерную плотность и межкластерное расстояние (чем меньше — тем лучше).

В ходе исследования были построены графики зависимости метрик от числа кластеров. Графики представлены на рисунке 3 и 4.

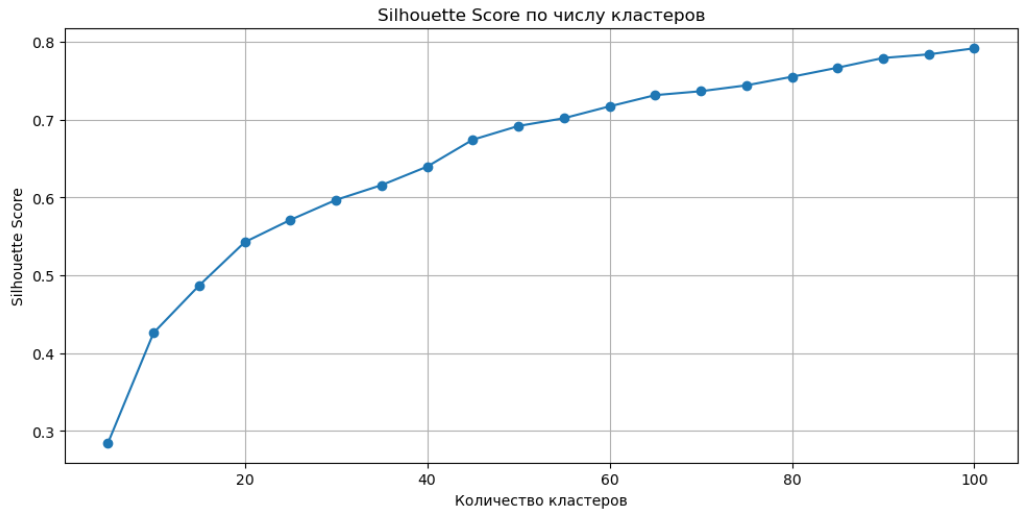


Рисунок 4 – Зависимость силуэтного коэффициента от количества кластеров.

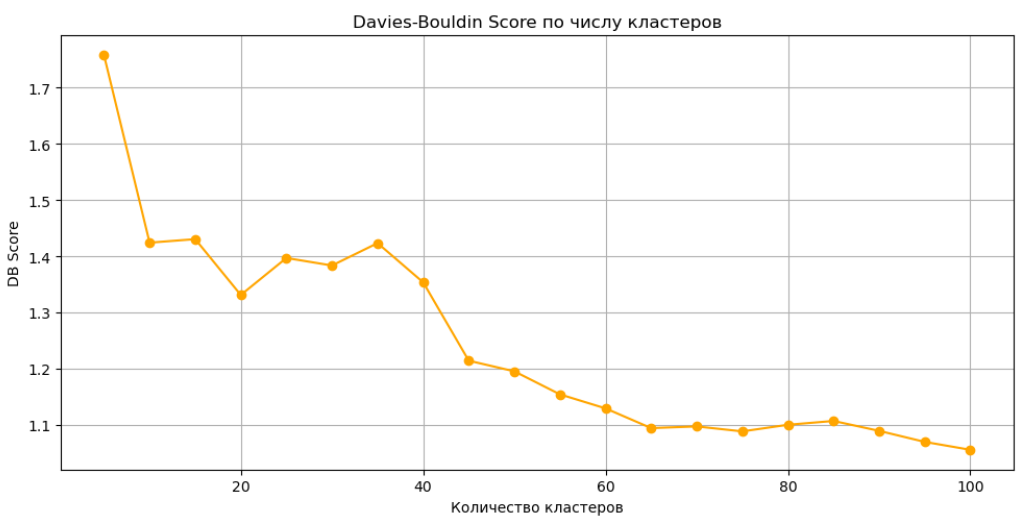


Рисунок 5 – Зависимость коэффициента Дэвиса от количества кластеров.

Анализ графиков показал, что при увеличении числа кластеров Silhouette Score растёт, а Davies-Bouldin Score уменьшается. Однако для избежания переобучения и чрезмерной детализации было выбрано значение k=70 кластеров, соответствующее сглаженному участку кривых графиков, где прирост качества начинает замедляться.

После обучения модели кластеризации на k=70, в качестве критериев оценки качества кластеризации использовались следующие метрики:

* Silhouette Score — 0.7363. Этот показатель свидетельствует о чёткой разделённости кластеров: значения выше 0.7 указывают на качественную кластеризацию.
* Davies–Bouldin Score — 1.097. Значение близкое к 1 говорит о хорошем соотношении внутрикластерной компактности и межкластерного расстояния.

Визуальная интерпретация результатов кластеризации представлена на рисунке 5.

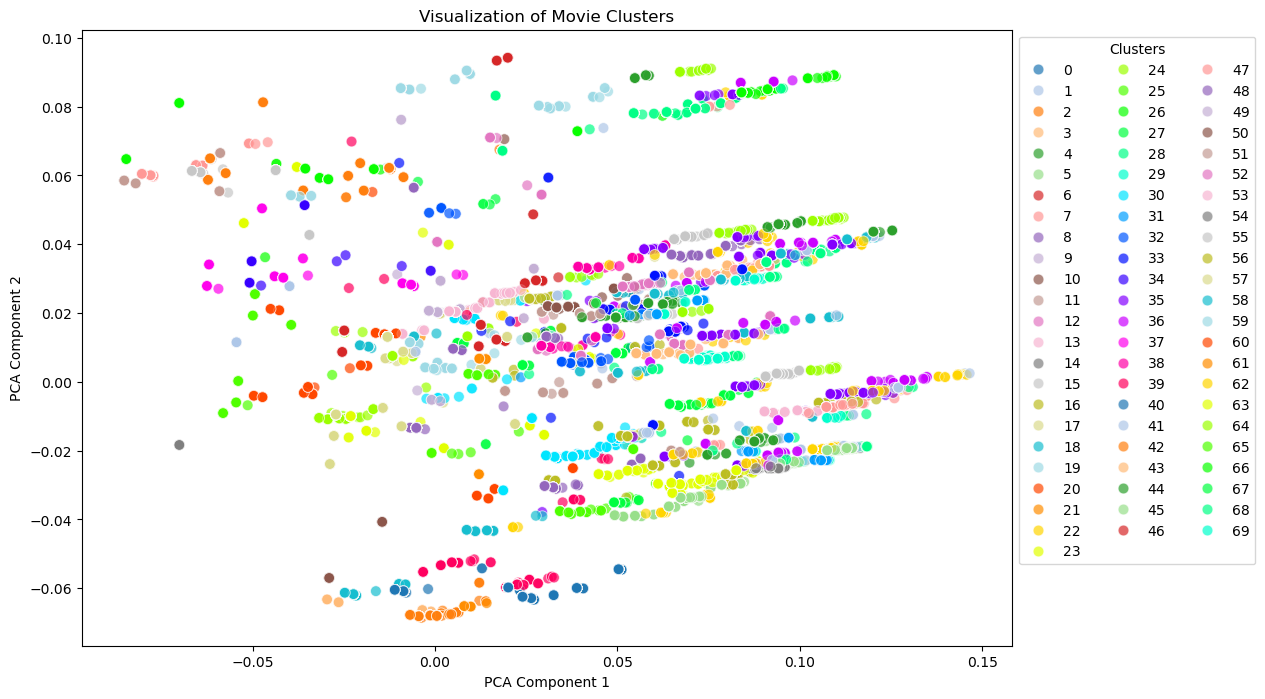


Рисунок 6 – Визуальное представление результатов кластеризации.

Финальным этапом разработки рекомендательной системы стало обучение модели XGBoost для предсказания оценок кластеров пользователями, для рекомендации фильмов из этих кластеров.

Для повышения точности предсказания пользовательских оценок фильмов, помимо базовых признаков, в обучающий датасет были добавлены дополнительные метрики, отражающие контекст взаимодействия пользователя с системой. Эти признаки позволяют модели учитывать, как индивидуальные предпочтения пользователя, так и общие тенденции по фильму. Ниже представлены ключевые из них:

1. Средняя оценка фильма

2. Количество оценок фильма

3. Отклонение оценки пользователя от его средней

4. Средняя оценка кластера

Так как, каждому фильму соответствует набор жанров. Для учёта этого фактора была применена процедура one-hot кодирования: каждый жанр получил отдельную бинарную переменную: 1, если фильм относится к этому жанру, и 0, если нет. Далее происходит поэлементное перемножение one-hot матрицы жанров на пользовательские оценки для получения среднего рейтинга пользователя для каждого жанра. Таким образом, создаётся профиль жанровых предпочтений пользователя, который затем используется в модели XGBoost как дополнительный признак.

Этап разделения данных на обучающую и тестовую выборки проводился следующим образом: 80% данных отводилось на обучающую выборку и 20% на тестовую.

Для проверки точности модели на тестовой выборке использовались следующие метрики:

1. MSE (Mean Squared Error) — 0.0024. Среднеквадратичное отклонение между предсказанием и реальными оценками.

2. MAE (Mean Absolute Error) — 0.0314. Среднее отклонение модели от реальной оценки — всего 0.031 балла на шкале от 0.5 до 5.0.

3. MAPE (Mean Absolute Percentage Error) — 1.14%. Средняя ошибка.

4. R² (Коэффициент детерминации) — 0.9978. Модель объясняет 99.78% вариативности пользовательских оценок.

Также был произведен анализ важности признаков, которыми обучалась модель. Важность признаков представлена на рисунке 7.

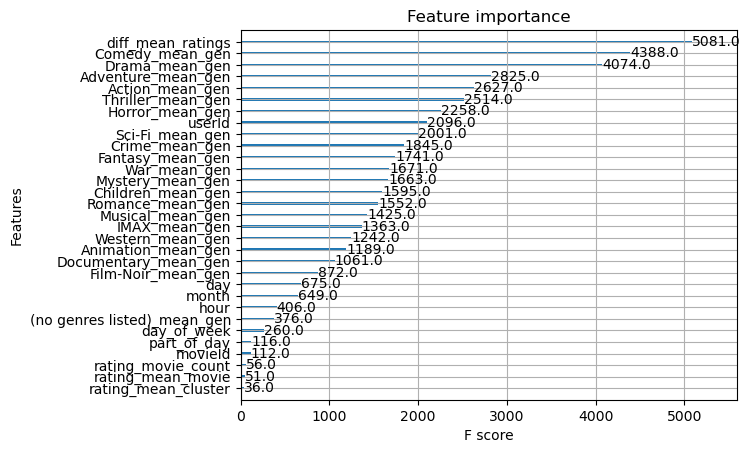


Рисунок 7 – Важность признаков для обучения модели XGBoost

На рисунке 7 показана важность признаков (Feature importance) в модели XGBoost, измеряемая по 𝐹-оценке, которая отражает, сколько раз признак использовался для разделения данных в деревьях модели и насколько сильно он улучшал качество предсказаний.

Самыми важными признаками оказались: diff\_mean\_ratings — разница средних оценок; Comedy\_mean\_gen, Drama\_mean\_gen, Adventure\_mean\_gen и т.д. — средние оценки по жанрам.

Данная модель позволяет учитывать широкий спектр признаков, включая как характеристики фильмов, так и поведенческие особенности пользователей. Это необходимо для повышения точности персонализированных рекомендаций.

**ЛИТЕРАТУРА**

1. Recommendation System in Python [Electronic resourse]. – Режим доступа: https://www.geeksforgeeks.org/recommendation-system-in-python/– Дата доступа: 06.05.2025.

2. Эмбеддинги для поиска [Электронный ресурс]. – Режим доступа: https://vc.ru/dev/605712-pishem-umnyi-poisk-po-kodu-s-open-ai. – Дата доступа: 03.05.2025.