МБОУ «Гимназия №64»
ГОАОУ «Центр поддержки одаренных детей «Стратегия»

# Большие данные, искусственный интеллект, финансовые технологии и машинное обучение
# Система предсказания рейтингов кинофильмов

Автор:
Афонин Захар Андреевич, 10 класс
Научный руководитель:
Мирошников Артём Игоревич,
методист Детского технопарка
«Кванториум»

Липецк, 2020

Большие данные, искусственные интеллект, финансовые технологии и машинное обучение

Система предсказания рейтингов кинофильмов

Афонин Захар Андреевич, 10 класс
Липецкая область, город Липецк,
МБОУ «Гимназия №64»
ГОАОУ «Центр поддержки одаренных детей «Стратегия»

В современном мире индустрия кино бурно развивается благодаря компьютерной графике. Часто сцены, нарисованные на компьютере, уже сложно отличить от натурных съёмок. В процессе производства фильмов и мультфильмов принимают участие актёры, режиссеры, продюсеры, дизайнеры, программисты, операторы. Так как киностудиям важно максимизировать прибыль, они делают всё, чтобы предсказать рейтинг, а значит, и кассовые сборы будущего фильма. Точное предсказание рейтинга позволяет выявить, каким должен быть фильм и когда наиболее выгодно его выпустить. Таким образом, система предсказания рейтингов кинофильмов будет полезна как киностудиям, так и владельцам сайтов-агрегаторов для улучшения впечатлений конечных потребителей.

**Цель работы:** построение предсказательной системы рейтинга ещё не вышедших кинофильмов на основе статистических данных.

**Задачи:**

1. Изучить различные источники информации, в которой отражена тема создания предсказательных моделей.
2. Проанализировать исходные данных и определить критерии, учитываемые при анализе существующего или гипотетического фильма.
3. Выявить параметры, наиболее сильно коррелирующие с численным рейтингом фильма.
4. На основе полученных данных создать и обучить предсказательную систему для определения рейтинга.
5. Создать пользовательский интерфейс для удобного взаимодействия с системой.

Существует несколько численных показателей успешности фильма. Наиболее часто используется рейтинговая система - общее значение рассчитывается на основе оценок пользователей или критиков. Рейтинги по версии Rotten Tomatoes, Metacritic и других сайтов-агрегаторов, собирающих оценки критиков, зачастую существенно

отличаются от рейтингов рядовых зрителей, пользователей таких сайтов как КиноПоиск, IMDb.com. Так как коммерческие киностудии стремятся максимизировать сборы, имеет смысл опираться именно на массовые оценки. Для этого был выбран набор данных, содержащий рейтинги с IMDb. Далее был проведён анализ базовых параметров, содержащихся в выборке.

## Результаты первичного анализа базовых параметров выборки

Большинство параметров были анализированы относительно рейтинга, так как этот параметр является искомым для будущей предсказательной системы. При анализе данных и создании модели был использован язык программирования Python 3 и его модули; для приведения выборки в глобальную совокупность применён модуль pandas, программная библиотека для обработки и анализа больших наборов данных. Выбор обоснован тем, что другие Python-модули для анализа больших данных, такие как SciPy, Scikit-Learn, используют типы данных pandas в качестве входных данных.

Распределение пользовательских оценок (Приложение 1) соответствует распределению с отрицательным коэффициентом асимметрии; среднее значение рейтингов приходится на оценку 6.38; медианное значение соответствует 6.8.

После приведения входных данных к условно исчисленным были рассчитаны коэффициенты корреляции параметров между собой и с рейтингом (Приложение 2) Год выпуска фильма обратно коррелирует с его рейтингом ($r$ = -0.14); данный параметр будет учитываться в модели.

Заметна сильная корреляция длительности фильма с рейтингом ($r$ = 0.35). Также интересны такие параметры, как популярность страниц режиссёра и актёров в соцсетях ($r$ = 0.19) и число лиц на постере ($r$ = -0.08) - данный параметр отрицательно коррелирует не только с рейтингом, но и с бюджетом, популярностью актёров. Эти параметры также учитываются в модели. Интересен тот факт, что бюджет практически не коррелирует с рейтингом фильма ($r$ = 0.03).

Некоторые параметры, такие как кассовые сборы, число рецензий и оценки критиков, заведомо исключены из модели, так как система направлена на предсказание рейтинга ещё не вышедших фильмов.

## Выбор и обучение предсказательной модели

Так как приходится учитывать большое количество численных и категориальных параметров, имеет смысл использовать методы машинного обучения. Для решения задач классификации используются несколько методов, среди которых наиболее часто применяются наивные байесовские классификаторы, методы опорных векторов (SVM), деревья решений (decision tree) и леса случайных деревьев (random forest).

Исходя из основных принципов машинного обучения, данные делятся в соотношении 8:2 - 80% тренировочной выборки, 20% тестовой. Также было принято решение разделить оценки по трём категориям: низкие (< 5.5), средние (5.5 - 7.5), высокие (> 7.5). Возможно разделение по четырём категориям: низкие (< 5.5), средние (5.5 - 7.0), высокие (7.0 - 8.5), очень высокие (> 8.5).

1. **Наивный байесовский классификатор**

   Метод не продемонстрировал высоких результатов, поскольку в его основе лежит допущение, что все параметры независимы друг от друга. Это не позволило достичь высокой точности: при классификации по трём категориям метод дал точность в 58,4%.

2. **Дерево решений**

   Достоинством данного метода является возможность использования как категориальных, так и численных параметров. Также деревья решений легко поддаются ручной корректировке, что способствует значительному увеличению точности при некоторых допущениях. Этот метод может использоваться на маломощных компьютерах и тонких клиентах, так как он не требует вычислительных мощностей в отличие от моделей, основанных на нейронных сетях или методах с необходимостью значительной обработки входных данных. Из недостатков можно отметить опасность переобучения модели, которой можно избежать путём редукции - отсечения крайних ветвей дерева. При использовании реализации модуля scikit-learn была достигнута точность 71,4%, что является значительно лучшим результатом, чем при использовании наивного байесовского классификатора.

3. **Метод случайных деревьев**

   Использование метода случайных деревьев заключается в создании большого комитета (ансамбля) решающих деревьев, каждое из которых обладает низкой

точностью классификации по отдельности. Сам случайный ансамбль (лес) является бэггингом над большим количеством деревьев, что позволяет получить точную классификацию на основе неточных деревьев. Таким образом, проблема переобучения не так актуальна для данного метода.

Метод случайных деревьев широко используется в сфере больших данных и машинного обучения. Существует множество реализаций данного метода, среди которых наибольшей применимостью обладают LightGBM, H2O, CatBoost и XGBoost. После тестирования различных реализаций был выбран модуль CatBoost от компании Яндекс, поскольку его реализация обладает наибольшей точностью среди вышеописанных. Также она наименее требовательна к вычислительным мощностям как при обучении, так и при работе с обученной моделью, что позволяет использовать её на персональных компьютерах без специализированных аппаратных ускорителей.

Благодаря упору на категориальные параметры реализация CatBoost показала наибольший результат - 84% при разбиении рейтингов на три группы и 70% при разбиении на четыре группы. (Приложение 3)

## Реализация пользовательского интерфейса

Для обеспечения максимальной совместимости как с операционными системами от Microsoft, так и открытыми решениями на базе Linux или BSD было принято решение отказаться от графического пользовательского интерфейса в пользу текстового. Это позволяет использовать систему как для предсказания одиночных рейтингов, так и для массового автоматизированного анализа. В дальнейшем планируется создать упрощённую версию программы с графическим интерфейсом, но без возможности тонкой настройки параметров.

## Заключение

В наши дни проблема предсказания рейтинга кинофильмов стоит особенно актуально. Применение методов машинного обучения позволяет достичь высоких результатов в данной области. Учитывая малое количество параметров, доступных для анализа по причине того, что фильм ещё не вышел, предсказательная система хорошо

справляется с задачей оценки рейтингов: точность предсказания равна 84% для разбиения на три категории и 70% для разбиения на четыре категории. Решение отличается от существующих систем тем, что оно не опирается на рейтинги критиков и других сайтов-агрегаторов. Все задачи, связанные с построением и обучением модели, были выполнены, цель проекта достигнута. В ходе работы был проведён анализ наиболее продуктивных режиссёров как по кассовым сборам, так и по рейтингам. (Приложение 4) Технические подробности хода работы изложены в файле Jupyter Notebook, находящемся в git-репозитории по адресу: https://github.com/AfoninZ/imdbest.

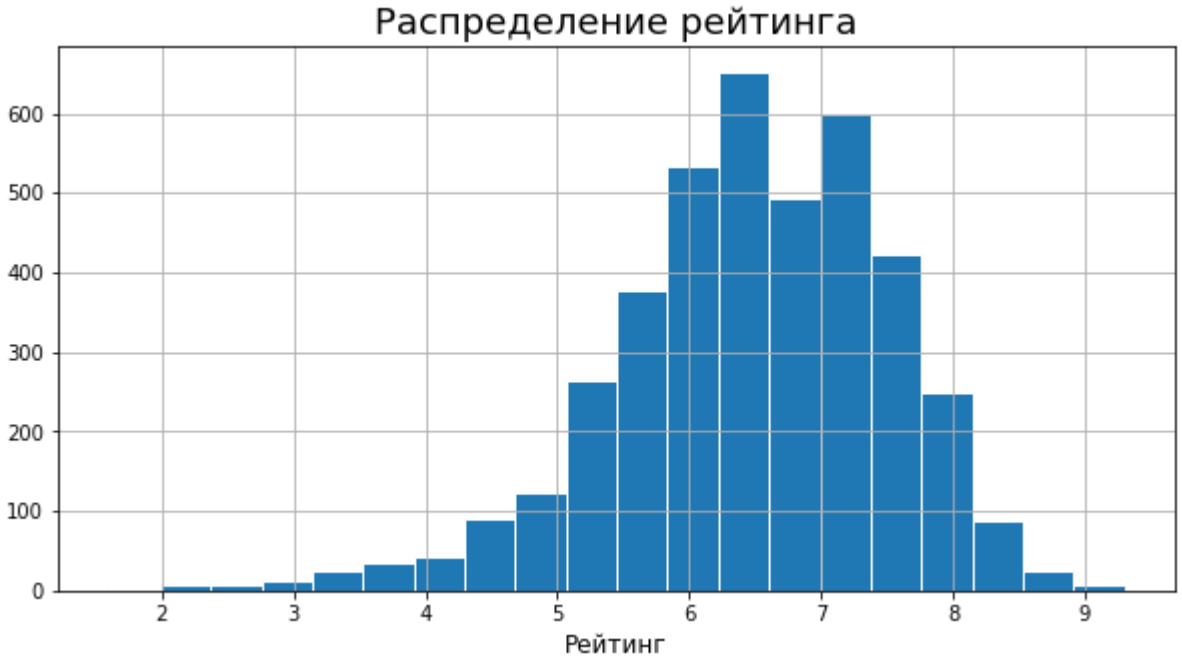
В рамках проекта были использованы следующие ресурсы, модули и реализации:

- Python 3.8 - в качестве основного языка программирования
- xlrd - для чтения набора данных из файла Excel-совместимого формата
- pandas, numpy - для обработки набора данных
- scikit-learn - для работы с категориальными параметрами
- catboost - основной модуль, для создания предсказательной модели
- matplotlib, seaborn - для визуализации полученных результатов

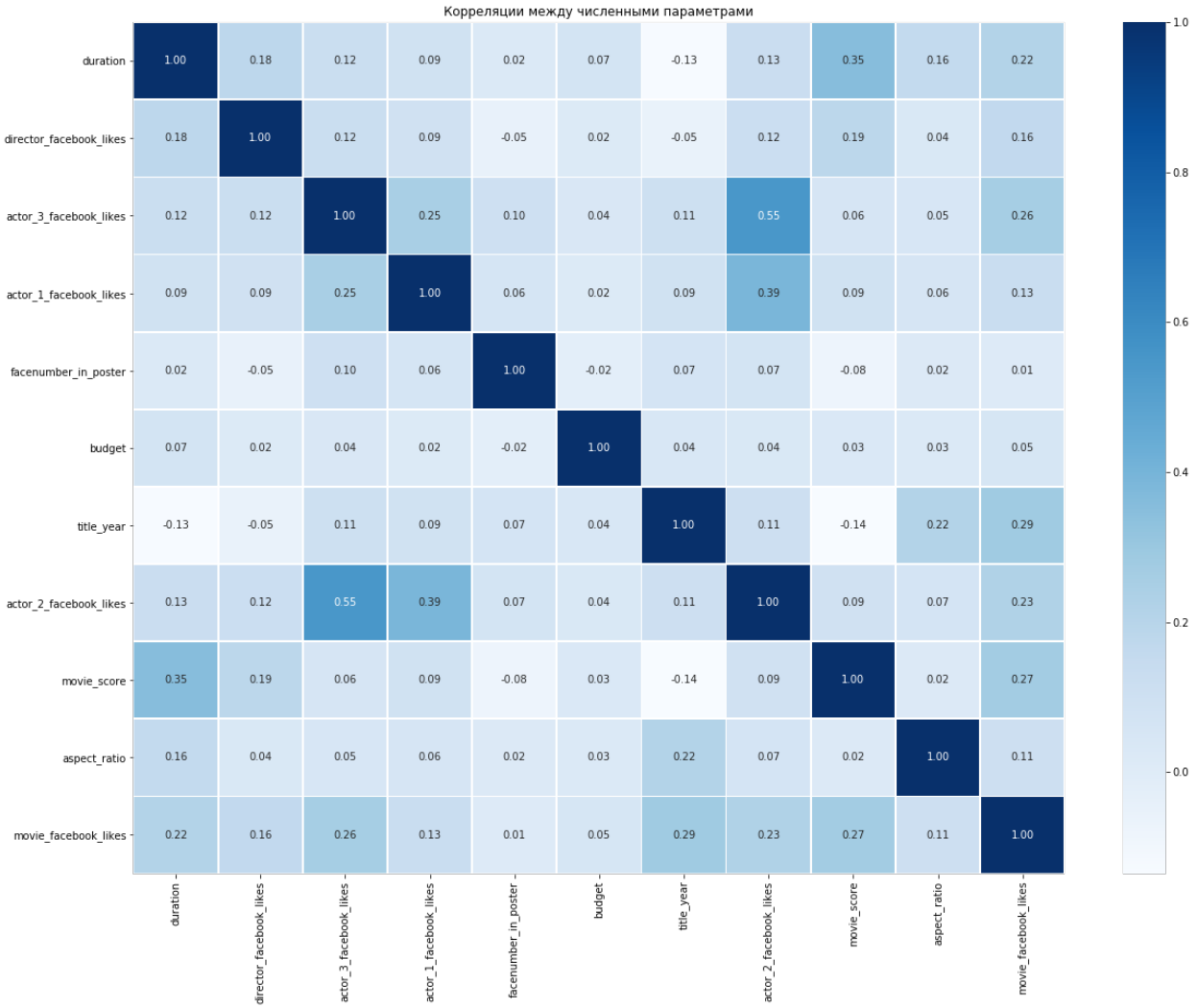
## Список ресурсов и источников

1. Жабский М. Кино в современном обществе: Функции - воздействие - востребованность / Министерство культуры Российской Федерации, НИИ киноискусства. М., 2002. С. 36.
2. Феномен массовости кино / Министерство культуры Российской Федерации, НИИ киноискусства; под общ. ред. М. И. Жабского. М., 2004. С. 67
3. Педро Д. Верховный алгоритм: как машинное обучение изменит наш мир; пер. с англ. В.Горохова; М.: Манн, Иванов и Фербер, 2016. 336 с.
4. Гудфеллоу Я., Бенджио И., Курвиль А. Глубокое обучение; пер. с англ. А.А. Слипкина. 2-е изд. М.: ДМК Пресс, 2018. 652 с.

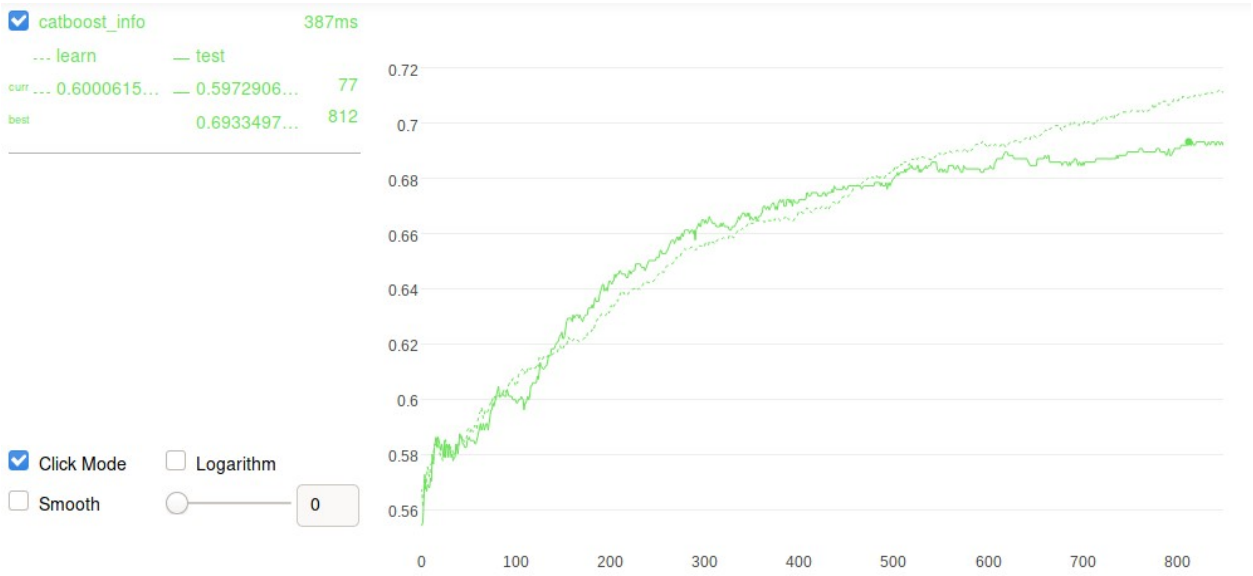
Приложение 1.




Пр
иложение 2.




Приложение 3.




Приложение 4.