А.В. Дождиков
к.полит.наук, независимый исследователь (Москва)

ПРОГНОЗИРОВАНИЕ РЕЗУЛЬТАТОВ КИНОПРОКАТА С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Аннотация. Предметом исследования являются результаты кинопроката российских национальных фильмов. Цель исследования — классификация проектов по принципу их успеха/неуспеха в прокате и прогнозирование характеристик проката. Задачи исследования — создание алгоритмов отбора (классификации) в инвестиционный портфель потенциально успешных проектов и прогнозирования (регрессии) прокатных характеристик: количество просмотров зрителей, окупаемость, зрительский рейтинг. Методика исследования основывается на применении ансамблевых моделей машинного обучения. База исследования — вся совокупность российских национальных фильмов в прокате с 2004 по апрель 2022 года (N=1469) и с мая 2022 г. по апрель 2023 г. (N=194). Основные результаты: достигнута точность (accuracy) в 0,95 и 0,89 для двух и четырёхклассовой классификации и высокие характеристики ROC_AUC = 0,97 для двухклассовой модели и 0,94–0,98 для четырёхклассовой. Более сложные метамодели (суперансамбли) могут достигать точности в 0,97–0,98 для двухклассовой классификации и 0,96 для четырёхклассовой. Сложные регрессионные метамодели прогнозируют абсолютные значения окупаемости, сборов, просмотров с коэффициентом детерминации (R2) в пределах 0,97–0,98 с использованием синтетических данных. В результате появилась возможность формирования инвестиционных портфелей кинопроектов с годовой исторической доходностью до 139%. Область применения — обеспечение отбора фильмов для инвестиционных «портфелей кинопроектов» государственных (Министерство культуры, Фонд кино) и частных инвесторов. Модели машинного обучения могут быть адаптированы к условиям глобального и иностранных рынков за счёт увеличения количества признаков, расширения арсенала методов машинного обучения, включая анализ текстов, изображений, видео и пользовательских данных социальных сетей.

Ключевые слова: *национальный кинематограф, инвестиции, финансирование, машинное обучение, прогнозирование, классификация, регрессия, портфельное инвестирование, рейтинг кинофильма, количество просмотров кинофильма, окупаемость кинофильма, успех фильма в прокате.*

JEL: G11, G17, Z11, C38, C53, C65, C45

УДК: 338.984

DOI: 10.52342/2587-7666VTE_2023_4_93_114

ДЛЯ ЦИТИРОВАНИЯ: *Дождиков А.В.* Прогнозирование результатов кинопроката с помощью машинного обучения // Вопросы теоретической экономики. 2023. № 4. С. 93–114. DOI: 10.52342/2587-7666VTE_2023_4_93_114.

FOR CITATION: *Dozhdikov A.* Prediction of the Results of Movie Release Using Machine Learning // Voprosy teoreticheskoy ekonomiki. 2023. No. 4. Pp. 93–114. DOI: 10.52342/2587-7666VTE_2023_4_93_114.

Введение[1]

Киноиндустрия является высокодоходной и одновременно высокорисковой сферой предпринимательской и инвестиционной деятельности. Исторически, с 2004 г. российский кинематограф убыточен, похожие суждения мы отметим у других исследователей данной тематики [*Аракелян*, 2016]. Окупается (собирает два своих бюджета) только 11,53% игровых фильмов в прокате. Средняя окупаемость (с учётом кассовых фильмов) не превышает 1,1, а медианная — всего 0,28[2].

Гипотеза исследования: с помощью алгоритмов машинного обучения до начала съёмок, самого затратного этапа, можно определить будет проект успешным или нет, а также насколько он будет успешным, основываясь на исторических данных совокупности проектов. Регрессионные модели могут определить величину сборов, количество просмотров зрителей и зрительский рейтинг.

Цель исследования: проведение эксперимента на выборке российских фильмов с 2004 г. с использованием обученных на тестовой выборке классификационных и регрессионных моделей машинного обучения по определению успешности фильма в прокате на тестовой и контрольной выборках.

Задачи исследования:

- изучение современного опыта аналитики данных в сфере кино с использованием машинного обучения;
- подготовка тренинговых и тестовых данных;
- обоснование выбора и обучение моделей машинного обучения;
- двух- и четырёхклассовая классификация кинопроектов, оценка точности и метрик моделей машинного обучения;
- использование регрессионных моделей для определения абсолютных величин сборов, просмотров и зрительского рейтинга с использованием выборки и синтетических данных;
- определение направлений совершенствования аналитики данных кинопроката.

Основной тезис исследования связан с предположением о том, что кинотеатральная аудитория достаточно консервативна, жанровые предпочтения и вкусы меняются очень медленно даже в современных условиях, а уход иностранных дистрибьютеров киноконтента ещё не повлиял на производственные условия (средний производственный цикл современного кинопроекта — 2 года, следовательно, в 2023 г. в прокат выходят фильмы, запланированные еще в «конкурентный» период).

Также не изменились критерии окупаемости и факторы, влияющие на итоговые результаты фильма в прокате. Тезисы о «социальном импакте» [*Печегина*, 2023] и измерении успешности фильма в прокате через его социальное воздействие [*Смекалин*, 2023] нивелируются простым фактом: фильм с высокими оценками от критиков и от ограничен-

---

[1] Автор выражает благодарность коллективу Высшей школы цифровой культуры ИТМО (Высшая школа цифровой культуры ИТМО https://dc.itmo.ru/) Михайловой Е.Г., Графеевой Е.Г., Егоровой О.Б., Бойцеву А.А., Романову А.А. за полученные знания; оргкомитету медиа холдинга ПЛАС, директору по развитию бизнеса Гризову К.А. за возможность представления практических результатов исследования и отраслевом финансовом журнале «ПЛАС» (*Дождиков А.В.* (2023). Могут ли инвестиции в кинематограф быть прибыльнее рынка акций и стабильнее сектора облигаций? PLUSworld (дата публикации: 19.07.2023) https://plusworld.ru/journal/2023/plus-7-2023/mogut-li-investitsii-v-kinematograf-byt-pribylnee-rynka-aktsiy-i-stabilnee-sektora-obligatsiy/ (дата обращения: 20.07.2023)) и выступления на Международном форуме «Финтех, банки и ритейл» 21–22 июня в секции «Большие данные. Стратегический капитал 21 века?» Программа 3-го Международного ПЛАС-Форума «ФИНТЕХ БАНКИ и РИТЕЙЛ (дата публикации: 22.06.2023). https://uz.plus-forum.com/uploads/uzforum-program-full-2023.pdf (дата обращения: 20.07.2023).

[2] *Дождиков А.В.* (2023). Успех «Чебурашки» можно повторить с помощью ИИ. И не только в киноиндустрии! PLUSworld. Дата публикации: 21.03.2023. https://plusworld.ru/journal/2023/plus-3-2023/uspekh-cheburashki-mozhno-povtorit-s-pomoshchyu-ii-i-ne-tolko-v-kinoindustrii/ (дата обращения: 06.07.2023).

ного пула кинозрителей фактически оказывает минимальное воздействие на общество, не выходя за рамки своей узкой аудитории. Отсутствие объективных количественных критериев оценки означает напрасно потраченные средства.

Инструменты машинного обучения позволят выявлять успешные (и отдельно — потенциально высокодоходные) проекты для проката на этапе «препродакшен», когда инвестиции в проект минимальны. Дополнительно можно определить количество просмотров каждого фильма, величину кассовых сборов, зрительский рейтинг. Объединение нескольких отобранных проектов в инвестиционный портфель нивелирует волатильность результатов каждого отдельного проекта и может поспособствовать росту инвестиционной привлекательности всей киноотрасли.

Обзор литературы

Использование алгоритмов машинного обучения и нейросетей для анализа кинопроката и кассовых сборов является достаточно распространенной практикой. Один из объектов изучения — настроения целевой аудитории [*Mohan Raj et al.*, 2017] в социальных сетях и их влияние на успех или неудачу фильма. «Большие данные» для прогнозирования успеха фильмов на основе активности онлайн-пользователей [*Mestyán et al*, 2013] используются, как минимум, с 2013 г. Алгоритмы машинного обучения применяются для анализа настроений [*Tripathi et al*, 2023] в социальных сетях, также осуществляется анализ мнений в обзорах и комментариях на специализированных киноресурсах [*Gupta et al*, 2021]. С помощью количественных методов прогнозируются награды киноакадемий, например «Оскар» [*Krauss et al*, 2008]. Современные аналитики предсказывают рейтинг IMDb [*Sivakumar et al*, 2021]. Количественной оценке может быть подвергнута эмоциональная реакция зрителей с помощью инструментальных диагностических средств, включая ЭКГ, МРТ [*Christoforou et al*, 2017].

Разработка алгоритмов прогнозирования успеха фильмов в прокате ведётся как в Голливуде и Европе [*Murschetz*, 2020], так и в Болливуде [*Meenakshi et al*, 2018]. Необходимо отметить растущую школу аналитики данных в Индии. Абсолютным лидером по дата-аналитике в сфере кино по числу публикаций является Китай: эксперты с помощью «машинного зрения» изучают визуальные данные — афиши и постеры фильмов и прогнозируют на их основе успех проекта в прокате. Интеллектуальный анализ данных кинопроизводства применяется на материалах Нигерии [*Olubukola et al*, 2021] и в Шри-Ланке [*Sivakumar et al,* 2021].

В России в научной сфере имеется крайне малое количество публикаций о прогнозировании результатов кинопроката с помощью алгоритмов машинного обучения и нейросетей. В частности, отметим проект ВШЭ «Методика нейросетевого прогнозирования кассовых сборов кинофильмов» [*Ясницкий и др.*, 2017], основанный, к сожалению, только на показателях и метриках американского рынка. Другое российское исследование [*Князева, Иванова,* 2020] содержит подробный обзор мирового опыта аналитики кинопроката и базируется на модели с расчётами для Microsoft Excel [*Педяш,* 2013]и точностью прогнозирования в 0,76. Так же необходимо отметить работы по прогнозированию спроса на кинофильмы, проводимые в России в начале 2010-х гг. [*Ноакк и др.*, 2012; *Татарников,* 2012, *Татарников и др.*, 2012].

Прошедшее десятилетие подарило аналитикам более высокотехнологичные и точные методы исследования и прогнозирования, основанные на инструментах машинного обучения. Как правило, используются такие модели, как Random Forest, Support Vector Machine и Neural Network and Deep Neural Network. По мнению отдельных исследователей [*Souza,* 2021], Random Forest (и ансамблевые модели) показывают лучшие результаты. Так, например, на датасете Box Office Mojo с данными 1980 — 2019 гг. по 3 167 фильмам в исследовании достигнута точность в 0,97.

Источники данных

В качестве базы для исследования использована вся совокупность российских национальных фильмов (Приложение 1), вышедших в прокат с 2004 г. по апрель 2022 г. (N=1469) и отдельно — с мая 2022 г. по апрель 2023 г. (N=194). Данные собраны из открытых источников[3]. Проведена перекрёстная проверка (в ручном режиме). Не учитывались проекты, не вышедшие в прокат или вышедшие сразу на телевидении и киноплатформах, короткометражные проекты и альманахи (сборники).

Реализация эксперимента проводилась в оболочке Anaconda и Jupiter Notebook с помощью библиотек и модулей Python — numpy, pandas, matplotlib, seaborn, sklearn. ensemble, sklearn.model_selection, sklearn.metrics и других.

Если разработанная ранее нами модель классификатора успеха/неуспеха фильма в прокате использовала всего 8 параметров, новая модель содержит расширенный набор данных (25 признаков), включающий в себя неделю и месяц выхода фильма в прокат; количество экранов проката (экранных копий); бюджет фильма; возрастной рейтинг; длительность; данные основного жанра (средние сборы, средний рейтинг, средние просмотры, окупаемость или соотношение сборы/бюджет); аналогичные исторические данные по второму жанру, аналогичные исторические данные по режиссёру фильма и двум сценаристам, всего 25 столбцов данных. По 30% российских фильмов бюджет не указан продюсерами (как правило, данные фильмы относятся к категории провальных). В пропусках использовались медианные значения с использованием замены NaN по методу df.feature.fillna(df.feature. median(), inplace = True). Данный подход также способствовал меньшей дискриминации режиссёров и сценаристов-новичков с отсутствующей историей проката. Вместо нулевых или минимальных значений они получали медианное.

Изначально предполагалось использование ансамблевой модели машинного обучения, основанной на GradientBoostingClassifier и GradientBoostingRegressor в силу её меньшей требовательности к ресурсам.

Boosting или последовательное обучение слабых моделей, как правило, использующих деревья решений, улучшает точность прогнозирования за счёт преобразования слабых классификаторов в единую сильную модель обучения. В плане осуществления регрессии метод уязвим в отношении аномалий и выбросов данных, поэтому желательна их стандартизация. С помощью метода quantile отбрасывались данные со значением прогнозируемых величин, не входящими в интервал от 0,05 до 0,95. Из sklearn.preprocessing, был импортиован стандартный метод StandardScaler.

Далее основная выборка была разделена на тренинговую и тестовую с помощью метода train_test_split из sklearn.model_selection. Отдельно использовалась контрольная выборка из фильмов с мая 2022 г. по апрель 2023 г.

Обоснование использования методов машинного обучения

В отличие, например, от AdaBoost, метод GradientBoosting выдает сразу точные результаты, а не исправляет ошибки. По этой причине метод GradientBoosting даёт более точные результаты на новых данных (контрольная выборка новых фильмов). XGBoost разработан для многоядерной параллельной обработки прямо при работе с большими данными. И поэтому он избыточен для применяемого ограниченного набора данных.

---

[3] Российские сайты и ресурсы, посвящённые кинотематике: film.ru, kinopoisk.ru, kinometro.ru, kinobusiness.com.

Bagging с параллельным обучением и последующим агрегированием результатов направлен на уменьшение разброса (дисперсии) в данных. Он больше подойдет для решения задач регрессии, связанных с прогнозированием точных значений — для последующей прикладной версии модели.

Методы Blending ограниченно применимы в данном случае по причине малой выборки. Ни базовые алгоритмы, ни метаалгоритм не используют всего объёма данных обучения (каждый — только свой кусочек). Это непрактично для небольших выборок (полторы — две тысячи записей), но, в принципе, данный метод можно использовать при работе с индивидуальными оценками потребителей контента, контентом, генерируемым пользователями. Возможный вариант — использование синтетических данных, сгенерированных на базе основного датасета.

Stacking — модель, несмотря на возможность объединения разных базовых моделей, как правило, сложна в настройке гиперпараметров и требует более высокой квалификации исследователя, также подход отличается непредсказуемостью результатов.

Voting также предполагает использование нескольких, не похожих между собой, моделей машинного обучения и их объединения. Применение метода целесообразно в связи с объединением нескольких источников данных по кинопроизводству, причём не только исторических данных проката, но и данных социальных сетей, результатов анализа текста, визуального ряда и других данных кинопроектов. Поэтому отметим данный метод как перспективный для дальнейших исследований.

В рамках эксперимента по классификации сравним несколько ансамблевых моделей, отметим, что наиболее подходящими являются AdaBoost и GradientBoosting по совокупности метрик — см. Приложение 2.

Можно существенно улучшить AdaBoost до точности 0,9786 и ROC_AUC: 0,9966, как показано в Приложении 3, однако такие показатели достигнуты на синтетическом наборе данных — исходный датасет увеличен в 10 раз с помощью метода resample из модуля sklearn.utils. Особняком в исследованиях будет стоять проблема «переобучения» моделей, основанных на слишком слабых классификаторах, и при наличии сильно «зашумлённых» данных.

Второй вариант повышения эффективности — это создание «суперансамбля» моделей, состоящего, например, из AdaBoostClassifier, BaggingClassifier, ExtraTreesClassifier, GradientBoostingClassifier, RandomForestClassifier, HistGradientBoostingClassifier (точность от 0,93 до 0,95) с подобранными гиперпараметрами, с последующим дообучением метамодели на результатах базовых моделей через RandomForestClassifier или другую модель с точностью в 0,9683 для двухклассовой и 0,9569 для четырехклассовой классификации — см. Приложение 4.

Третий вариант — это одновременное использование «метамодели» и синтетических данных для получения более точных прогнозов и метрик машинного обучения. Для прогнозирования абсолютных значений по окупаемости, величине сборов, количеству просмотров зрителей и зрительскому рейтингу «Кинопоиска» используется ансамблевый алгоритм регрессии VotingRegressor, объединяющий результаты пяти моделей машинного обучения — RandomForestRegressor, ExtraTreesRegressor, AdaBoostRegressor, GradientBoostingRegressor, HistGradientBoostingRegressor с работой на увеличенной в 10 раз выборке с синтетическими данными.

В случае с классификацией для основной модели GradientBoosting с точностью в 0,95 отмечается большее удобство работы, поэтому применение моделей с точностью прогнозирования на тестовых данных в 0,97–0,98 представляет скорее теоретический интерес. Повышение точности предсказаний абсолютных величин связано с использованием синтетических данных и построением сложных метамоделей.

Результаты применения GradientBoostingClassifier для задач классификации кинопроекта

Ансамблевая модель прогнозирования, как и в случае с исследованиями коллег из Китая [*Wu*, 2021], задействовавших Gradient boosting decision tree (GBDT), должна давать лучший результат, чем «слабые модели» по отдельности.

Модель была обучена на тренинговой выборке из 1 028 фильмов и протестирована на тестовой выборке из 441 кинофильмов. Изначально достигнутая точность (accuracy) в 0,945 не является максимальной. Поэтому с помощью метода GridSearchCV был осуществлён подбор гиперпараметров ансамблевой модели (learning_rate: 0,02, max_depth 2, min_samples_leaf: 4, n_estimators: 130), точность улучшена до 0,952. Метрики качества модели представлены в табл. 1. Значение ROC_AUC[4] = 0,971 (рис. 1).

Таблица 1
Метрики эффективности двухклассовой модели GradientBoostingClassifier

	precision	recall	f1-score	support
0	0,96742	0,97970	0,97352	394
1	0,80952	0,72340	0,76404	47
accuracy			0,95238	441
macro avg	0,88847	0,85155	0,86878	441
weighted avg	0,95059	0,95238	0,95119	441

**Источник*: рассчитано автором в ходе эксперимента

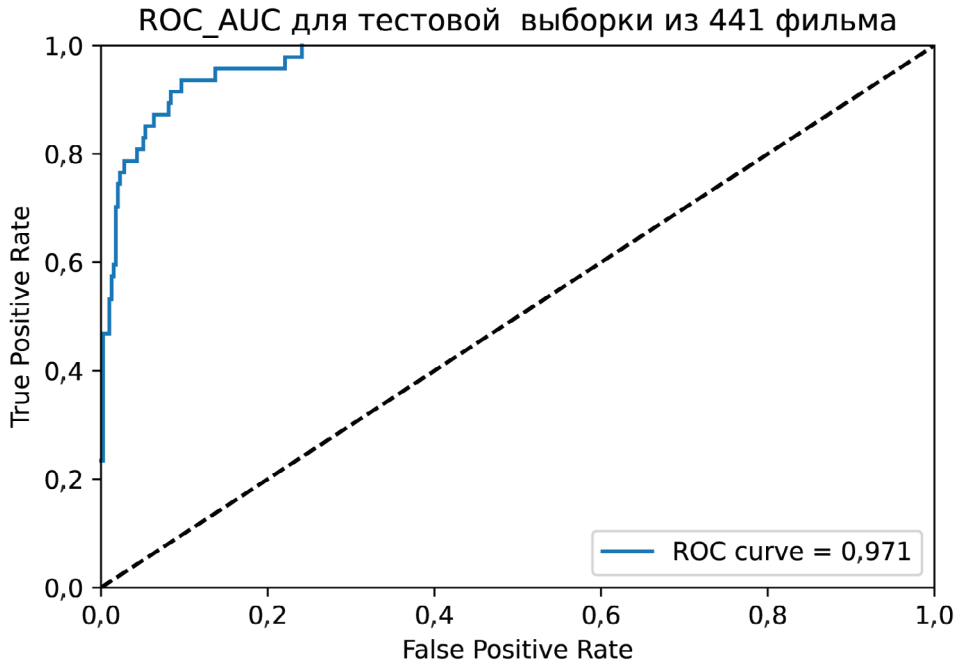

Рис. 1. ROC_AUC для тестовой выборки
Источник: рассчитано автором в ходе эксперимента.

[4] Площадь под кривой ROC показывает эффективность алгоритма, чем она ближе к единице, тем работа алгоритма совершеннее.

Дополнительно модель была протестирована на новой выборке (не вошедшей, соответственно, в тестовую выборку), представляющей новые данные кинопроката за последний период с мая 2022 г. по апрель 2023 г. При подтверждении заявленных характеристик модели можно будет признать справедливыми два следующих утверждения: модель классификатора эффективна на новых, неизвестных данных и может использоваться для определения типов новых проектов; с мая 2022 г. по апрель 2023 г. не произошло фундаментальных изменений в критериях окупаемости кинопроектов — несмотря на уход иностранных кинодистрибьютеров.

На дополнительной выборке модель показала похожую точность в 0,948 и ROC_AUC = 0,976 для выборки из 194 новейших фильмов российского проката (рис. 2).

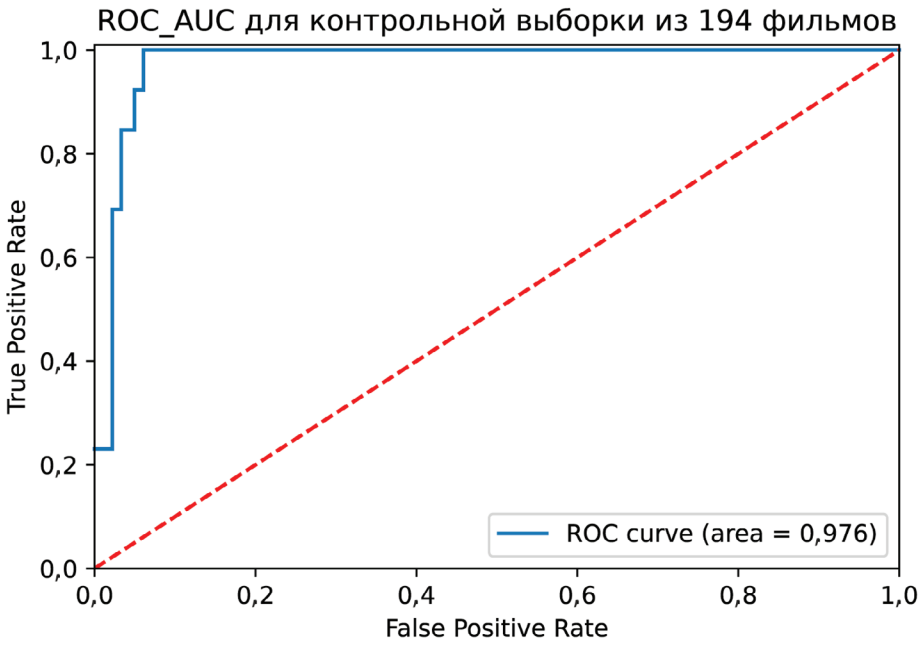

Рис. 2. ROC_AUC для контрольной выборки
Источник: рассчитано автором в ходе эксперимента.

Наши предположения получили подтверждение. Несмотря на уход иностранных дистрибьютеров с российского рынка, изменение критериев окупаемости кинопроектов ещё не произошло. Отчасти на данную ситуацию влияет средний производственный цикл большинства кинокартин в два года. В 2023 г. в прокат выходят картины, которые были приняты в производство до апреля 2022 г. То есть какие-то существенные изменения проявятся через несколько лет.

Имея исходный набор параметров фильма, мы можем с высокой степенью достоверности говорить о том, будет проект успешным или нет. В числе параметров: результаты режиссёра и сценаристов на прошлых работах (сборы, окупаемость, рейтинг «Кинопоиска», просмотры зрителей), жанровая принадлежность проекта, примерная длительность, возрастной рейтинг, предполагаемое количество экранов в прокате и ожидаемый период выхода и другие показатели. Следующее направление анализа — оценка того, насколько может быть успешным проект. Здесь мы воспользуемся той же ансамблевой моделью GradientBoostingClassifier, но только для варианта с четырёхклассовой классификацией: 0 — фильм провалился в прокате, собрал менее одного бюджета, 1 — фильм не окупился, собрал от одного до двух своих бюджетов, 2 — фильм окупился и принес прибыль, 3 — фильм принес своим создателям прибыль свыше 100%.

Была использована похожая схема анализа: выборка разбита на тренинговую и тестовую в соотношении 70 к 30%. Точность модели четырехклассовой классификации составила

0,887. С использованием метода подбора гиперпараметров модели GridSearchCV (learning_rate: 0,1, max_depth: 5, n_estimators: 180) точность повышена до 0,893. Метрики эффективности модели представлены в табл. 2. Ещё раз повторимся, что более сложная в настройке по сравнению с GradientBoostingClassifier ансамблевая метамодель может давать более точные результаты — 0,968 для двухклассовой и 0,957 для четырёхклассовой классификации (Приложение 4).

Таблица 2
Метрики эффективности четырёхклассовой модели GradientBoostingClassifier

	precision	**recall**	**f1-score**	**support**
0	0,94340	0,97493	0,95890	359
1	0,63333	0,54286	0,58462	35
2	0,53846	0,33333	0,41176	21
3	0,66667	0,69231	0,67925	26
accuracy			0,89342	441
macro avg	0,69546	0,63586	0,65863	441
weighted avg	0,88319	0,89342	0,88666	441

**Источник:* рассчитано автором в ходе эксперимента.

Значения ROC_AUC для крайних классов (0 — абсолютный провал и 3 — абсолютный успех) достаточно высоки. Результаты для «промежуточных» классов ниже (рис. 3). Отметим, что при обучении на достаточном наборе исторических данных проката, модель машинного обучения будет эффективнее любого эксперта-человека из «Фонда кино» или частной организации. По результатам российского исследования 2019 г.[5] только 19 из 160 фильмов, поддержанных государством, окупились в прокате т.е. 12%, что близко средним показателям по рынку (11,5% — рассчитано автором).

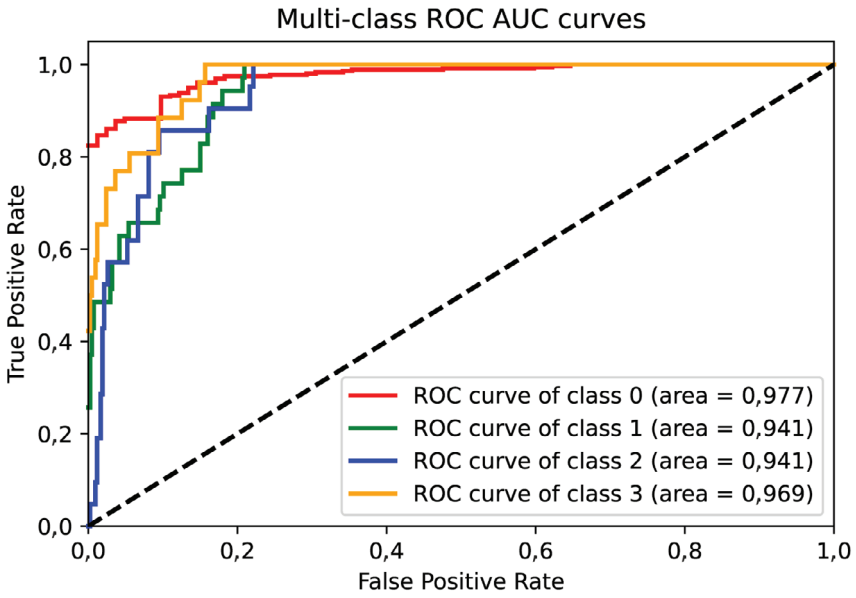

Рис. 3. ROC_AUC для тестовой выборки из 441 кинофильма
Источник: рассчитано автором в ходе эксперимента

[5] В 2022 году в прокате окупился лишь один поддержанный Фондом кино фильм. РБК. https://www.rbc.ru/technology_and_media/10/02/2023/63e38a649a7947a68eb0b59d (дата обращения 06.07.2023).

Для новой выборки на данных с мая 2022 г. по апрель 2023 г. точность предсказаний составила 0,887, что так же свидетельствует о неизменности факторов, влияющих на окупаемость кинофильма в прокате.

Результаты применения GradientBoostingRegressor и VotingRegressor для определения абсолютных показателей кинопроекта

С помощью регрессионных моделей машинного обучения можно определить, сколько человек просмотрит конкретный фильм, каковы будут его сборы в абсолютных значениях и, наконец, какой рейтинг «Кинопоиска» он может получить. Повышение точности возможно за счёт использования синтетических данных и увеличения объёма выборки.

В качестве первого примера используем ансамблевую модель GradientBoostingRegressor для оценки окупаемости кинопроектов (соотношение сборы/бюжет). В целях повышения точности исходный и контрольный датасеты были отмасштабированы, удалены «выбросы» и аномальные значения (крайне низкие и аномально высокие оценки). Для оценки модели мы воспользуемся коэффициентом детерминации (R2), чем ближе его значение к единице, тем лучше модель прогнозирует значения. На графиках представлена зависимость между фактическим и предсказанными значениями. До подбора гиперпараметров модели R2 составил 0,809, после — 0,825 для тестовой выборки (рис. 4). Для контрольной выборки R2 составил 0,813 в отношении кассовых сборов (рис. 5).

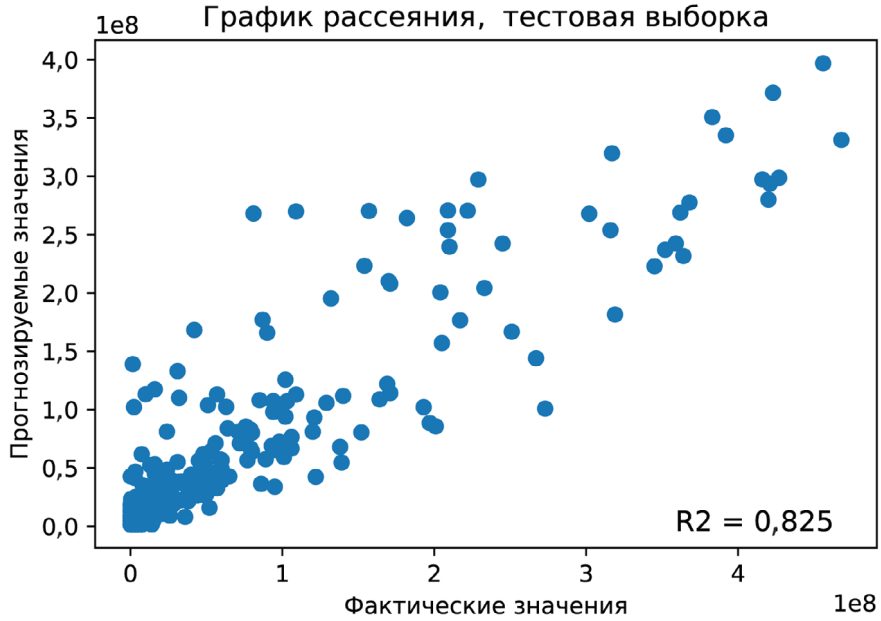

Рис. 4. График рассеяния для тестовой выборки из 441 кинофильма
Источник: рассчитано автором в ходе эксперимента

Точность исследований может быть повышена, в частности, за счёт анализа области от 1 до 1,0, поскольку она содержит наибольшее количество данных. Повторимся, что только 11,5% российских фильмов окупается в прокате. Отдельная область — это гиперуспешные, кассовые фильмы, собирающие в прокате свыше 4-х своих бюджетов. Таких проектов мало, статистические закономерности по ним выявлять сложнее всего, однако ранее рассмотренная модель четырёхклассового классификатора все же способна выделять данные проекты из общего массива.

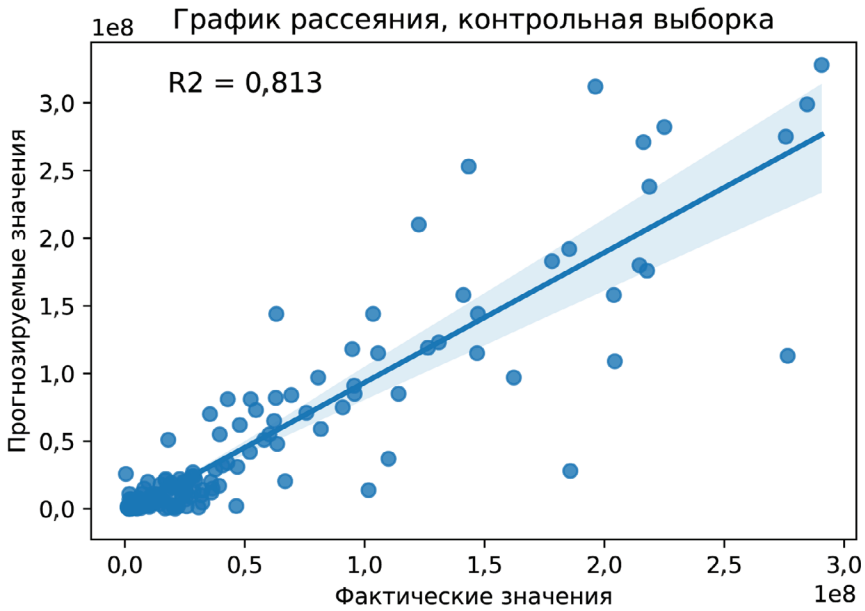

Рис. 5. График рассеяния для контрольной выборки из 194 кинофильмов
Источник: рассчитано автором в ходе эксперимента

Можно ли повысить точность прогнозирования? В качестве рабочего метода используем комбинацию «усложнённая метамодель» VotingRegressor (Приложение 5) с набором синтетических данных, сгенерированных на основе имеющегося датасета при помощи метода resample (рис. 6).

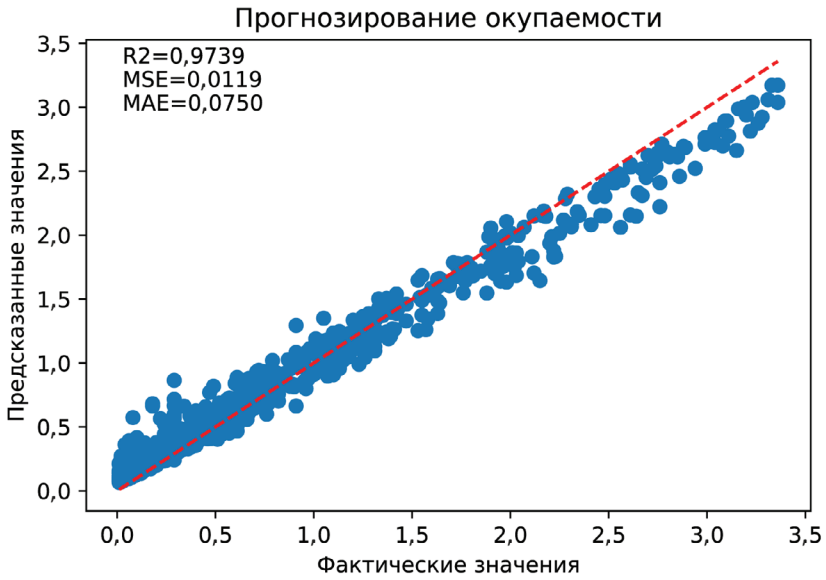

Рис. 6. График рассеяния, полученный на увеличенном с помощью синтетических данных датасете при помощи метамодели VotingRegressor
Источник: рассчитано автором в ходе эксперимента

Обученная на большом наборе данных метамодель VotingRegressor получила высокое значение коэффициента детерминации, свидетельствующее о прогностических возможностях. Другие метрики MSE (Mean Squared Error, средняя квадратическая ошибка), оценка среднего значения квадрата ошибок, различие между предсказанием и фактическим значением и MAE (Mean Absolute Error, средняя абсолютная ошибка), оценка того, насколько близки предсказания к фактическим значениям — так же свидетельствуют о точности модели.

Аналогичный прогноз с помощью регрессионной модели GradientBoostingRegressor можно провести в отношении рейтинга «Кинопоиска». Здесь распределение данных более равномерное и связь между предполагаемыми и фактическими значениями более наглядна (рис. 7).

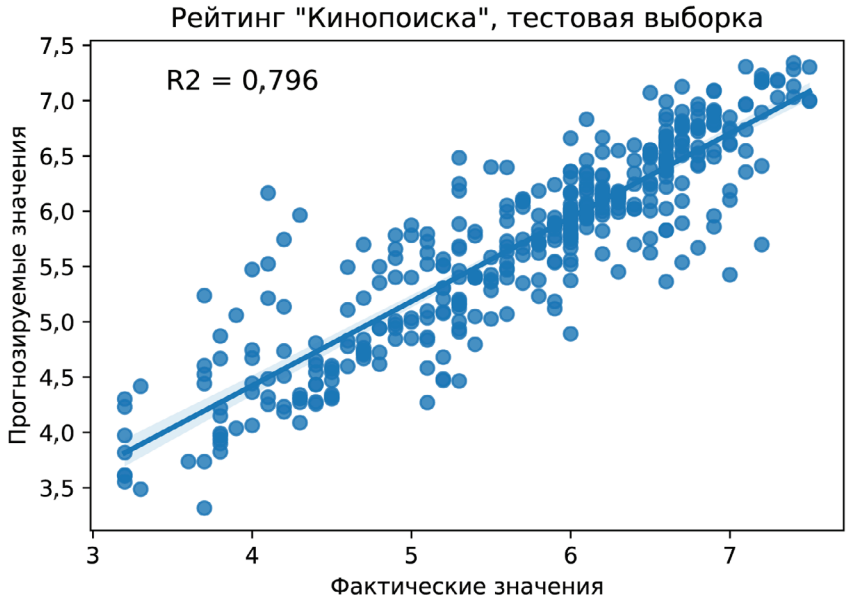

Рис. 7. Рейтинг «Кинопоиска», график рассеяния для тестовой выборки из 441 кинофильма
Источник: рассчитано автором в ходе эксперимента

С использованием синтетических данных и VotingRegressor можно получить более эффективную регрессионную метамодель с метриками R2, MSE, MAE для прогнозирования итогового зрительского рейтинга (рис. 8).

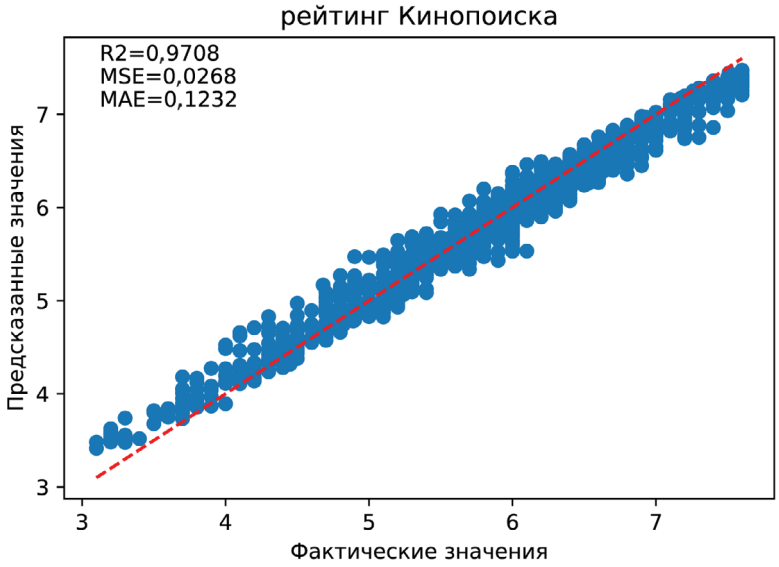

Рис. 8. Рейтинг «Кинопоиска», график рассеяния для тестовой выборки синтетических данных (увеличенная в 10 раз выборка)
Источник: рассчитано автором в ходе эксперимента

Не менее важный показатель для прогнозирования — количество просмотров. Для малоизвестных и малобюджетных фильмов, например, подпадающего под катего-

рию «феномен якутского кино»[6], не так важны сборы и рейтинг, как просмотры зрителей, поскольку малый бюджет окупается даже на ограниченном кинотеатральном прокате. И именно на «малом бюджете» опытные продюсеры тестируют и подбирают молодые режиссёрские дарования: если проект собирает полные залы на ограниченном прокате, без существенной рекламной поддержки — значит создан действительно стоящий продукт. На синтетических данных коэффициент детерминации составил 0,9748 (рис. 9).

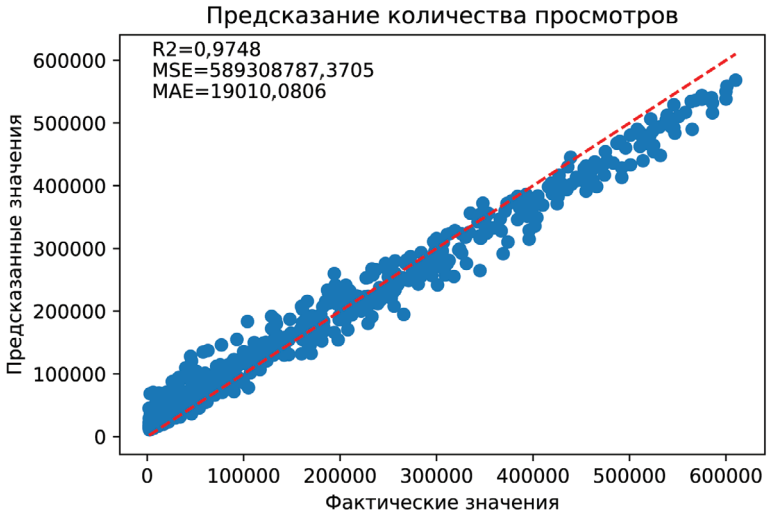

Рис. 9. Зрительские просмотры, график рассеяния, синтетические данные
Источник: рассчитано автором в ходе эксперимента.

Наибольший интерес для потенциальных инвесторов представляют данные по сборам фильмов, выраженным в абсолютных значениях. Для этого мы проанализируем все российские фильмы с помощью набора синтетических данных. Текущего набора данных в 1,6 с небольшим тысяч представляется недостаточным для точного обучения алгоритма. К тому же необходимо учесть, что 99% всех фильмов собирает 1,2 и менее млрд руб в прокате, отдельные «блокбастеры» являются скорее исключениями, чем правилом. Поэтому из выборки исследования исключим проекты со сборами свыше 1,2 млрд, в силу их достаточной редкости и остановимся на проектах, связанных с окупаемостью средних и небольших бюджетов (рис. 10).

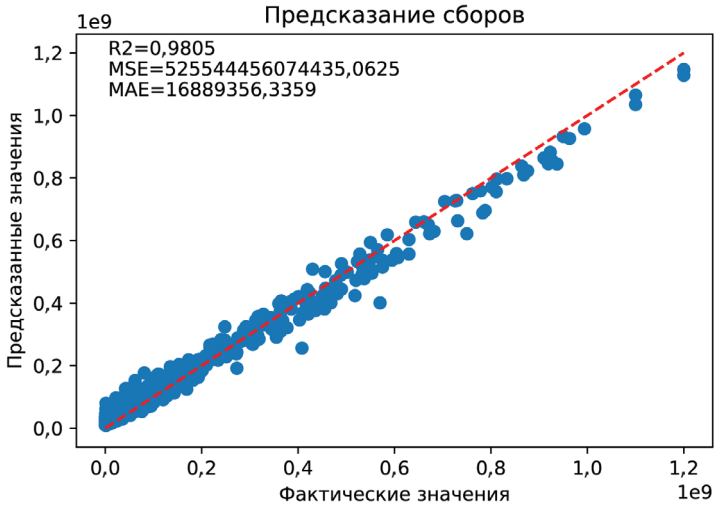

Рис. 10. Сборы фильма в прокате, график рассеяния, синтетические данные
Источник: рассчитано автором в ходе эксперимента.

[6] *Аброшина М.* (2021). Новый взгляд: как якутское кино завоевало мир и в чем его главная особенность (дата публикации: 10.06.2021). Лента. https://lenta.ru/articles/2021/10/06/yakutkino/ (дата обращения: 06.07.2023).

В результате получена достаточно точная модель прогнозирования на основе исторических данных, позволяющая, в том числе, оценивать предполагаемую величину сборов. Последнее является крайне необходимым для бюджетирования портфеля проектов и расчётов данных по возврату инвестиций.

У нас есть исторические данные по всем режиссёрам, работающим с 2004 г. Соответственно, зная величину предполагаемых сборов, мы можем оценить ожидаемый возврат инвестиций. Возьмём условное правило, что все режиссеры и их проекты получают одинаковую сумму инвестиций. Тогда отобранные с помощью данной методики проекты, которые будут сопоставлены с перечнем отобранных режиссёров, принесут прибыль, стремящуюся к средним результатам данных режиссеров (рис. 11). В приведённой таблице, если воспользоваться усреднёнными значениями между авторами, набравшими больше 1 по критерию «окупаемость» (box_budget_ch), мы получим среднее значение окупаемости в 3,31, что означает ((3,31/2-1), доход от инвестиций в размере 65%. Если мы остановимся только на первой семёрке самых успешных режиссёров, то получим возврат от инвестиций (4,78\2-1), или 139%. Однако достижение этого показателя будет достаточно рискованной практикой. Для инвестиционного портфеля целесообразнее отбирать порядка 15–20 режиссёров, реализующих 20–30 проектов ежегодно в течение минимум 5 лет. Только в этих условиях можно рассчитывать на существенную доходность и возврат инвестиций как с учётом средней длительности периода производства, так и с учётом возможной индивидуальной волатильности итоговых результатов по каждому проекту.

Режиссер	снято проектов	среднее кол-во экранов	средние сборы на проект	средние просмотры на проект	средний бюджет проекта	средние сборы на экран	средние просмотры на экран	среднее соотношение сборы/ бюджет
Жора Крыжовников	5	1 616	899 400 000	3 780 000	147 035 213	572 324	2 383	8,33
Клим Шипенко	7	914	645 428 571	2 457 643	156 418 750	359 201	1 395	5,72
Алексей Нужный	6	1 359	489 833 333	1 806 500	240 746 282	316 287	1 173	4,63
Тимур Бекмамбетов	8	996	646 912 500	4 028 625	207 097 668	812 618	5 884	4,24
Дмитрий Дьяченко	11	1 191	784 272 727	3 086 000	204 079 197	563 821	2 362	3,81
Виктор Шамиров	5	469	130 256 800	535 460	32 212 625	146 599	728	3,54
Егор Баранов	8	1 068	227 250 000	947 500	109 270 671	179 214	742	3,22
Роман Каримов	8	766	69 500 000	299 000	39 403 482	108 071	494	3,06
Марюс Вайсберг	12	1 213	336 333 333	1 496 500	119 268 489	303 905	1 415	3,04
Сарик Андреасян	15	1 070	231 866 667	1 058 133	165 582 204	229 303	1 073	2,46
Фёдор Бондарчук	6	1 131	887 500 000	4 133 333	701 020 636	888 854	5 237	1,71
Святослав Подгаевский	6	1 059	91 833 333	408 000	56 213 542	90 440	406	1,68
Артем Аксененко	6	862	120 500 000	537 833	84 866 725	134 793	601	1,50
Пётр Буслов	5	931	268 480 000	1 510 600	258 826 950	330 377	2 128	1,44
Карен Оганесян	10	726	90 240 000	434 600	91 451 456	112 786	608	1,27
Алексей Балабанов	6	118	31 733 333	225 167	54 948 800	224 709	1 529	0,87
Олег Асадулин	7	831	54 557 143	253 286	102 189 567	60 793	295	0,80
Павел Руминов	6	342	37 000 000	177 250	44 454 390	68 117	387	0,75
Денис Чернов	6	1 412	209 333 333	984 000	435 707 450	143 818	689	0,70
Николай Хомерики	7	749	99 320 429	409 338	317 694 742	97 969	404	0,67
Анна Матисон	6	488	23 211 667	91 500	68 246 093	46 506	200	0,66
Валерий Тодоровский	5	647	171 200 000	909 000	286 256 875	222 716	1 195	0,59
Руслан Бальтцер	5	241	37 915 600	320 200	85 224 183	157 354	1 418	0,56
Константин Буслов	5	1 022	68 435 600	287 727	233 362 363	61 707	272	0,55
Дмитрий Суворов	5	769	40 320 000	169 400	89 039 815	48 214	204	0,51
Павел Лунгин	6	390	52 833 333	287 167	212 740 640	158 416	960	0,49
Ренат Давлетьяров	7	987	69 257 143	311 857	178 029 725	60 832	284	0,41
Анна Меликян	7	279	17 494 000	83 383	63 292 820	60 049	377	0,34
Игорь Волошин	5	333	39 099 000	180 362	110 597 050	52 098	255	0,33
Алексей Учитель	7	608	109 785 714	488 857	430 643 520	175 863	1 115	0,32
Кирилл Серебренников	8	187	22 106 375	77 488	96 067 100	84 566	344	0,27

Рис. 11. Средние результаты в прокате российских режиссеров

Источник: рассчитано автором в ходе эксперимента, цитирование по источнику[7]

[7] *Дождиков А.В.* (2023). Успех «Чебурашки» можно повторить с помощью ИИ. И не только в киноиндустрии! PLUSworld (дата публикации: 21.03.2023). https://plusworld.ru/journal/2023/plus-3-2023/uspekh-cheburashki-mozhno-povtorit-s-pomoshchyu-ii-i-ne-tolko-v-kinoindustrii/ (дата обращения: 06.07.2023).

Заключение

У разработанной методики оценки прокатных характеристик кинопроектов есть ограничения — отсутствие «исторических данных» для сценаристов и режиссёров-новичков, в тестовом датасете их показатели заменяются на медианные значения. Аналогичное ограничение возникнет и при добавлении в модель данных по актёрам-новичкам и другим участникам творческой группы.

Демографические и социально-экономические данные зрительской аудитории для более детального анализа в отношении интересов, выставляемых рейтингов пока недоступны для исследования и сопоставления. В перспективе это данные по просмотрам российских киноплатформ и обезличенные, деперсонифицированые пользовательские данные, информация от электронных систем по продаже билетов в классические кинотеатры (например, данные ЕИАС[8] и «Пушкинской карты»[9]).

Прогнозирование успеха/неуспеха фильма по производственным факторам — сценарий, режиссёрская экспликация, костюмы, реквизит, пластический грим и спецэффекты, локации и другие составляющие — требует использования алгоритмов другого уровня, включая «машинное зрение» и сверточные нейронные сети, а также языковые модели, методики векторизации и токенизации текста. В отношении производственных и финансовых документов и планов выполнения возможна оценка хода исполнения на большом количестве документов, составляющих внутреннюю документацию киностудий, получение которой требует доступа.

В базовую модель прогнозирования могут быть добавлены исторические данные по актёрам и другим участникам творческих групп. Для повышения точности анализа может быть сделан акцент на сложных ансамблевых моделях машинного обучения и нейросетях, увеличении выборки с помощью синтетических данных.

Дополнительно в перечень анализируемых данных можно включить элементы аудиовизуального ряда (постеры, трейлеры, рекламные материалы), текстовое описание (аннотации, логлайн, синопсис) и сценарий в «американском» формате, который очень хорошо поддается формализации. Данное направление потребует больших вычислительных ресурсов и применения алгоритмов работы с текстом и изображениями, превышающих возможности стандартного офисного оборудования.

Помимо представленного метода, основанного на исторических и текущих данных [*Chakraborty*, 2019], возможно изучение реакции человека на видеокартинку из трейлера фильма, зафиксированной с помощью средств МРТ и ЭЭГ, систем трекинга глаз. Здесь необходимо отметить гипотезу о возможной связи полученных нейрофизиологических показателей с когнитивными состояниями сфокусированного внимания, кодированием долговременной памяти и синхронизацией различных компонентов сети вознаграждений мозга [*Christoforou et al*, 2019].

Для коммерческой индустрии крайне важно раннее предсказание производительности фильма за счёт подхода, основанного на «глубоком обучении» с учётом анализа только сюжетной части проекта (текста), когда факторы стадий «продакшен» и «постпродакшен» неизвестны заранее. Для осуществления этой деятельности возможно использование «гибридных» методов, учитывающих как характеристики фильма, а также настроения, выраженные в обзорах фильма [*Tripathi, Tiwari., Saini, Kumari,* 2023]. Особенно перспективным считается использование мнений пользователей, высказанных в социальных сетях по

---

[8] Единая федеральная автоматизированная информационная система сведений о показах фильмов в кинозалах. Официальный ресурс подведомственной организации органа власти. https://ekinobilet.fond-kino.ru/ (дата обращения: 06.07.2023)

[9] Программа популяризации культурных мероприятий среди молодежи. Официальный ресурс: ФКУ Цифровая культура. https://пушка.рф/ (дата обращения: 06.07.2023)

поводу ожиданий от фильма, с учётом страновых особенностей и отличий рынков, например, Голливуда от Болливуда.

К перечню «гибридных» методов мы можем отнести: аналитику визуальных изображений — промопостеров [*Wi et al*, 2020]; анализ маркетинговой среды и прогнозов показателей кинопроката с помощью рекуррентных нейронных сетей [*Yu, Liu*, 2022]; анализ неформальной коммуникации и «сарафанного радио» в отношении информации о будущих фильмах [*Yun Kyung*, 2017].

При отборе потенциально успешных/неуспешных проектов модели машинного обучения заведомо опережают любого продюсера или киноведа по точности прогнозирования финансового успеха проекта в прокате.

На данных датасета, окупаемость российских кинопроектов равна 1,1, даже с учётом проектов, многократно собирающих свой бюджет и приносящих прибыль. Медианная окупаемость российских кинопроектов в 2004–2023 гг. составила всего 0,28. При этом показатели проектов, получивших безвозвратную государственную поддержку, хуже, чем данные по рынку, в отличие, например, от Китая [*Сокуренко, Маглинова*, 2021], где государственная поддержка способствует увеличению спроса и ёмкости рынка кинопроката.

Обученная модель может обрабатывать сотни и тысячи кинопроектов. Для предварительной финансовой оценки проектов масштабного конкурса наподобие «Метода»[10] или «Питчинга дебютантов»[11] предварительно обученной ансамблевой модели или нейросети потребуется несколько секунд. Государственные фонды и частные инвесторы могут использовать исторические данные и инструменты машинного обучения для определения перспективных инвестиционных направлений деятельности в сфере кинематографа.

Одно из возможных направлений развития «киберпродюссирования»: оптимальный подбор «гиперпараметров» фильма, принятого в производство для достижения максимального охвата целевой аудитории, по аналогии, например, с методами оптимизации модели машинного обучения GridSearchCV, RandomizedSearchCV и других. Проект любого фильма можно улучшить за счёт подбора или уточнения длительности, жанровой принадлежности, тайминга, возрастного рейтинга, наличия ключевых актёров, режиссёра, композитора и других важных членов творческой группы.

Прогнозирование развития российского кинематографа как важной части креативных индустрий связано с повышением количества успешных фильмов в прокате, переориентацией частных и государственных ресурсов на коммерчески успешные проекты, созданием эффективной программы государственного финансирования российского кинематографа [*Сокуренко, Маглинова*, 2021]. Если в современных условиях из 200 проектов прибыль приносят в среднем 23, то увеличение доли успешного кино выведет отрасль на уровень самоокупаемости и самофинансирования, откроет путь к развитию индустрии с помощью кредитных средств, а также эмиссии акций, облигаций и других финансовых инструментов, включая, например, краундфандинг [*Замбалаева* и др., 2019].

[10] Метод — Всероссийский образовательный проект. https://metod.one/movie (дата обращения: 06.07.2023).
[11] Всероссийский питчинг дебютантов. https://moviestart.ru/pitchingi/ (дата обращения: 06.07.2023).

ПРИЛОЖЕНИЯ

Приложение 1

Основной датасет

ID_kinopoisk	date	week	month	screens	budget	age_R	time	genre_box_buget	genre_avr_kinopoisk_R	genre_avr_box
70952	12.02.2004	7	2	100	46096480	12	115	0,51	6,18	66377566
77396	01.04.2004	14	4	117	57620600	0	98	0,94	5,23	181394667
79850	08.07.2004	28	7	315	121003260	16	115	0,94	5,23	181394667
253754	30.09.2004	40	9	47	46096480	12	101	0,51	6,18	66377566
252013	28.10.2004	44	10	188	57620600	6	90	1,19	6,07	148983326
85104	28.10.2004	44	10	93	86430900	0	80	0,86	5,34	141798727
81201	11.11.2004	46	11	35	57620600	16	100	1,88	5,36	129218872
81202	09.12.2004	50	12	300	201672100	12	105	0,86	5,34	141798727
81041	23.12.2004	52	12	200	115241200	12	79	1,88	5,36	129218872
208105	30.12.2004	53	12	14	43215450	16	94	1,88	5,36	129218872
104958	13.01.2005	3	1	130	14144600	12	105	1,88	5,36	129218872
85127	20.01.2005	4	1	73	70723000	16	85	1,88	5,36	129218872
254806	03.02.2005	6	2	22	28289200	18	112	1,88	5,36	129218872
40775	24.02.2005	9	2	367	113156800	16	130	0,86	5,34	141798727
81485	17.03.2005	12	3	320	99012200	16	132	0,51	6,18	66377566
95943	24.03.2005	13	3	137	42433800	16	93	1,88	5,36	129218872
81203	31.03.2005	14	3	270	50920560	16	110	0,79	5,96	52280641
94111	21.04.2005	17	4	352	113156800	16	127	0,79	5,96	52280641
418372	19.05.2005	21	5	30	70723000	18	80	1,88	5,36	129218872

Рис. 12. Фрагмент основного датсета
Fig. 12. Fragment of the main dataset

Приложение 2

Сравнение метрик ансамблевых моделей

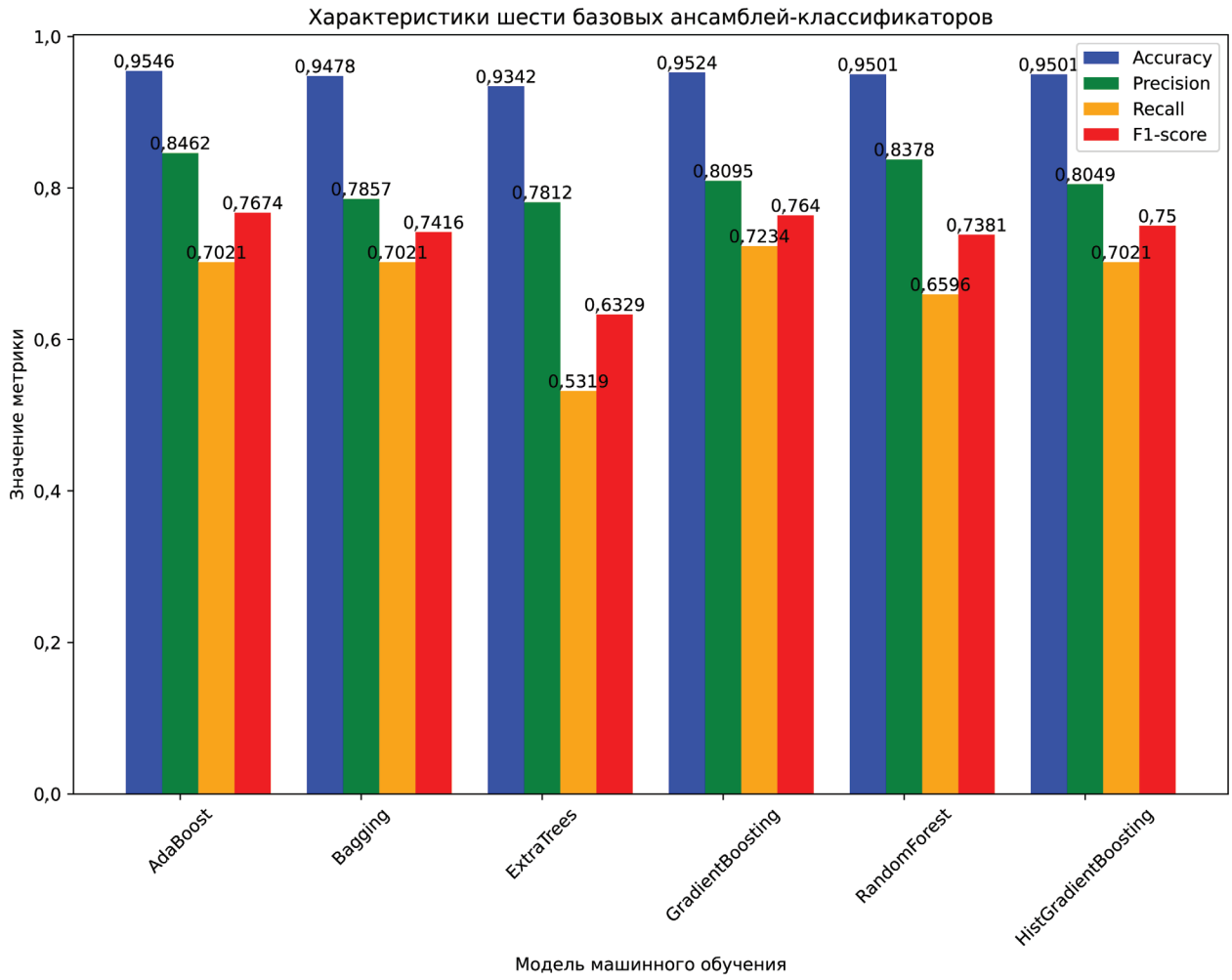

Рис. 13. Метрики моделей машиного обучения
Fig. 13. Metrics of Machine Learning Models

Приложение 3
Метрики AdaBoostClassifier на синтетическом (увеличенном в 10 раз) наборе данных

Табл. 3. Матрица ошибок
Tab. 3. Matrix of errors

Confusion Matrix:	Predicted Negative	Predicted Positive
Actual Negative	4417	34
Actual Positive	73	465

Табл. 4. Метрики двухклассовой классификации
Tab. 4. Two-class classification metrics

	precision	recall	f1-score	support
0	0,9837	0,9924	0,9880	4451
1	0,9319	0,8643	0,8968	538
accuracy			0,9786	4989
macro avg	0,9578	0,9283	0,9424	4989
weighted avg	0,9781	0,9786	0,9782	4989

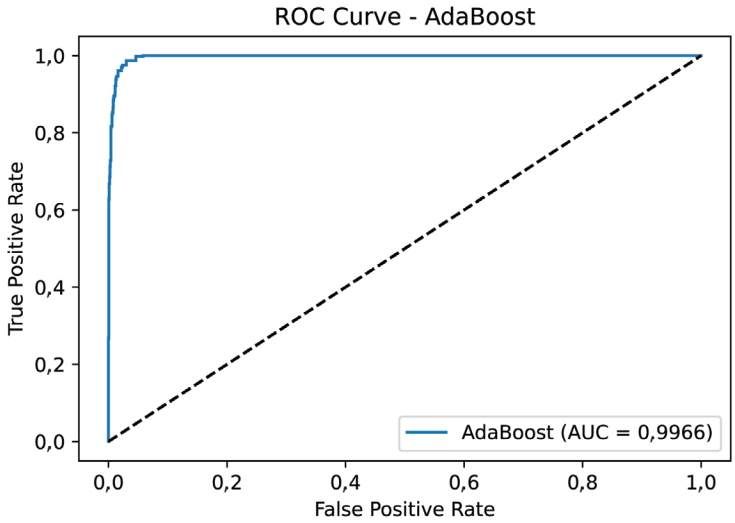

Рис. 14. Кривая ROC+AUC для алгоритма Adaboost с использованием синтетических данных
Fig. 14. ROC_AUC curve for Adaboost algorithm using synthetic data

Приложение 4
Показатели и метрики метамодели классификатора (суперсансамбля)

```
# Задание базовых моделей ансамбля
    base_models = [
        AdaBoostClassifier(),
        BaggingClassifier(),
        ExtraTreesClassifier(),
        GradientBoostingClassifier(),
        RandomForestClassifier(),
        HistGradientBoostingClassifier()
    ]
# Обучение базовых моделей и поиск параметров
    for i, model in enumerate(base_models):
        grid_search = GridSearchCV(model, params[i], scoring='accuracy', cv=5)
        grid_search.fit(X_train, y_train)
        best_model = grid_search.best_estimator_
        base_models[i] = best_model
        print(best_model)

# Подобранные гиапрпараметры моделей
    AdaBoostClassifier(n_estimators=10)
    BaggingClassifier(max_samples=0,6, n_estimators=35)
    ExtraTreesClassifier(max_depth=12, n_estimators=360)
    GradientBoostingClassifier(learning_rate=0.5, max_depth=7, n_estimators=200)
    RandomForestClassifier(max_depth=10, n_estimators=200)
    HistGradientBoostingClassifier(l2_regularization=0.1, max_depth=4)
```

Табл. 5. Метрики двухклассовой классификации
Tab. 5. Metrics of two-class classification

	precision	recall	f1-score	support
0	0,9774	0,9873	0,9823	394
1	0,8837	0,8085	0,8444	47
accuracy			0,9683	441
macro avg	0,9306	0,8979	0,9134	441
weighted avg	0,9674	0,9683	0,9676	441

Табл. 6. Метрики четырёхклассовой классификации
Tab. 6. Four-class classification metrics

	precision	recall	f1-score	support
0	0,9650	0,9972	0,9808	359
1	0,8621	0,7143	0,7813	35
2	0,9333	0,6667	0,7778	21
3	0,9615	0,9615	0,9615	26
accuracy			0,9569	441
macro avg	0,9305	0,8349	0,8753	441
weighted avg	0,9551	0,9569	0,9542	441

Приложение 5
Структура метамодели и гиперпараметры моделей, составляющих VotingRegressor

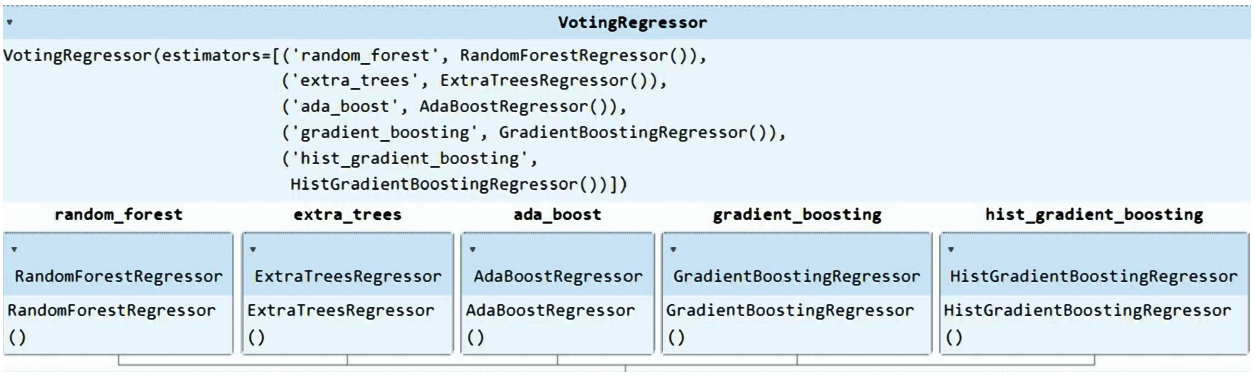

Рис. 15. Метамодель на базе VotihgRrgressor
Fig. 15. Metamodel based on VotihgRgressor

Гиперпараметры моделей, на которых достигнуты коэффициенты R2, MAE, MSE для прогнозирования

Гиперпараметры для модели random_forest:
{‘bootstrap’: True, ‘ccp_alpha’: 0.0, ‘criterion’: ‘squared_error’, ‘max_depth’: None, ‘max_features’: 1.0, ‘max_leaf_nodes’: None, ‘max_samples’: None, ‘min_impurity_decrease’: 0.0, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘min_weight_fraction_leaf’: 0.0, ‘n_estimators’: 100, ‘n_jobs’: None, ‘oob_score’: False, ‘random_state’: None, ‘verbose’: 0, ‘warm_start’: False}

Гиперпараметры для модели extra_trees:
{‘bootstrap’: False, ‘ccp_alpha’: 0.0, ‘criterion’: ‘squared_error’, ‘max_depth’: None, ‘max_features’: 1.0, ‘max_leaf_nodes’: None, ‘max_samples’: None, ‘min_impurity_decrease’: 0.0, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘min_weight_fraction_leaf’: 0.0, ‘n_estimators’: 100, ‘n_jobs’: None, ‘oob_score’: False, ‘random_state’: None, ‘verbose’: 0, ‘warm_start’: False}

Гиперпараметры для модели ada_boost:
{‘base_estimator’: None, ‘learning_rate’: 1.0, ‘loss’: ‘linear’, ‘n_estimators’: 50, ‘random_state’: None}

Гиперпараметры для модели gradient_boosting:
{‘alpha’: 0.9, ‘ccp_alpha’: 0.0, ‘criterion’: ‘friedman_mse’, ‘init’: None, ‘learning_rate’: 0.1, ‘loss’: ‘squared_error’, ‘max_depth’: 3, ‘max_features’: None, ‘max_leaf_nodes’: None, ‘min_impurity_decrease’: 0.0, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘min_weight_fraction_leaf’: 0.0, ‘n_estimators’: 100, ‘n_iter_no_change’: None, ‘random_state’: None, ‘subsample’: 1.0, ‘tol’: 0.0001, ‘validation_fraction’: 0.1, ‘verbose’: 0, ‘warm_start’: False}

Гиперпараметры для модели hist_gradient_boosting:
{‘categorical_features’: None, ‘early_stopping’: ‘auto’, ‘l2_regularization’: 0.0, ‘learning_rate’: 0.1, ‘loss’: ‘squared_error’, ‘max_bins’: 255, ‘max_depth’: None, ‘max_iter’: 100, ‘max_leaf_nodes’: 31, ‘min_samples_leaf’: 20, ‘monotonic_cst’: None, ‘n_iter_no_change’: 10, ‘quantile’: None, ‘random_state’: None, ‘scoring’: ‘loss’, ‘tol’: 1e-07, ‘validation_fraction’: 0.1, ‘verbose’: 0, ‘warm_start’: False}

ЛИТЕРАТУРА / REFERENCES

Аракелян А.М. (2016). Формирование условий повышения инвестиционной привлекательности киноотрасли в России [*Arakelyan A. M.* (2016). Formation of conditions for increasing the investment attractiveness of the film industry in Russia] // *Сервис plus*. Т. 10. № 2. С. 74–79. DOI: 10.12737/19459.

Замбалаева Т.Б., Рыжкова М.В., Чиков М.В. (2019). Бизнес-схемы краудфандинговых платформ в обеспечении финансирования инновационного проекта [*Zambalaeva T.B., Ryzhkova M.V., Chikov M.V.* (2019). Business schemes of crowdfunding platforms in providing financing for an innovative project] // *Экономика и управление инновациями*. № 2. С. 70–83. DOI: 10.26730/2587-5574-2019-2-70-83.

Князева И.Г., Иванова Д.М. (2020). Прогнозирование кассовых сборов проката фильма [*Knyazeva I.G., Ivanova D.M.* (2020). Forecasting the box office of the film] // *Вестник Омского университета. Серия: Экономика*. Т. 18. № 2. С. 24–37. DOI: 10.24147/1812-3988.2020.18(2).24-37.

Ноакк Н.В., Неволин И.В., Татарников А.С. (2012). Методика прогнозирования выручки от проката кинофильмов [*Noakk N.V., Nevolin I.V., Tatarnikov A.S.* (2012). Methodology for forecasting revenue from movie rentals] // *Финансовая аналитика: проблемы и решения*. № 48. С. 17–24.

Педяш Е.А. (2013). *Эконометрическое прогнозирование кассового успеха кинофильмов: магистерская диссертация* [*Pedyash E.A.* (2013). Econometric Forecasting of Film Box Office Success: Master's dissertation] М.: НИУ ВШЭ (дата публикации: 06.03.2013). https://www.hse.ru/data/2013/06/03/1285529668/Магистерская_диссертация_Педяш_Евгений.docx (дата обращения: 06.07.2023).

Печегина Т. (2023). Импакт идет в кино. Оценка эффективности инвестиций в российский кинематограф [*Pechegina T.* (2023). Impact goes to the cinema. Evaluation of the effectiveness of investments in Russian] // *Позитивные изменения.* Т. 3. № 2. С. 16–27. DOI: 10.55140/2782-5817-2023-3-2-16-27.

Смекалин И. (2023). Фильмы, меняющие жизнь. Оценка социального импакта кино и практики доказательности в кинопроизводстве[*Smekalin I.* (2023). Movies that change lives. Assessment of the social impact of cinema and the practice of evidence in film production] // *Позитивные изменения.* Т. 3 № 2. С. 28–38. DOI: 10.55140/2782-5817-2023-3-2-28-38.

Сокуренко К.В., Маглинова Т.Г. (2021). Особенности развития рынка киноиндустрии в Китае [*Sokurenko K.V., Maglinova T.G.* Peculiarities of development of the film industry market in China] // *Казанский экономический вестник*. №1 (51). С. 37–41.

Татарников А.С. (2012). Методы прогнозирования кассовых сборов [*Tatarnikov A.S.* (2012). Methods for predicting box office receipts] // *Бюллетень кинопрокатчика*. №10–11 (75–76). С. 50–56.

Татарников А.С., Неволин И.В., Ноакк Н.В. (2012). Выявление спроса на неосязаемые продукты (на примере кинофильмов) [*Tatarnikov A.S., Nevolin I.V., Noakk N.V.* (2012). Identification of demand for intangible products (on the example of films)] / *Труды 55-й научной конференции МФТИ. Инновации и высокие технологии*. — М.: МФТИ. С. 65–67.

Ясницкий Л.Н., Белобородова Н.О., Медведева Е.Ю. (2017). Методика нейросетевого прогнозирования кассовых сборов кинофильмов [*Yasnitsky L.N., Beloborodova N.O., Medvedeva E.Yu.* (2017). Methodology for neural network forecasting of box office receipts of films] // *Финансовая аналитика: проблемы и решения*. Т. 10. № 4(334). С. 449–463.

Chakraborty P., Zahidur R, Saifur R. (2019). Movie Success Prediction using Historical and Current Data Mining// *International Journal of Computer Applications.* Vol. 178. No. 47. Pp. 1–5. DOI:10.5120/ijca2019919415.

Christoforou C., Papadopoulos T.C., Constantinidou F., Theodorou M. (2017). Your Brain on the Movies: A Computational Approach for Predicting Box-office Performance from Viewer's Brain Responses to Movie Trailers // *Frontiers in neuroinformatics*. December. Vol. 11. Article 72. DOI: 10.3389/fninf.2017.00072.

Gupta C., Chawla G., Rawlley K., Bisht K., Sharma M. (2021). Senti_ALSTM: Sentiment Analysis of Movie Reviews Using Attention-Based-LSTM // *Proceedings of 3rd International Conference on Computing Informatics and Networks. Lecture Notes in Networks and Systems.* A. Abraham, O. Castillo, D. Virmani (eds). Vol. 167. Springer, Singapore. DOI: 10.1007/978-981-15-9712-1_18.

Krauss J., Nann S., Simon D., Fischbach K., Gloor P. (2008). Predicting Movie Success and Academy Awards through Sentiment and Social Network Analysis // *16th European Conference on Information Systems* // AISeL. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1185&context=ecis2008.

Li D., Liu Z.-P. (2022). Predicting Box-Office Markets with Machine Learning Methods // *Entropy.* 24. No. 5: 711. DOI: 10.3390/E24050711.

Meenakshi K., Maragatham G., Agarwal N., Ishitha G. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie // *Journal of Physics:* Conference Series. 1000 012100. DOI: 10.1088/1742-6596/1000/1/012100.

Mestyán M., Yasseri T., Kertész J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data // *PLoS ONE*. No. 8 (8):e71226. DOI: 10.1371/journal.pone.0071226.

Mohan Raj, Prasanna and Aditya, S. (2017). Predictive Model for Movie's Success and Sentiment Analysis// *Research Journal of Management Sciences.* Vol. 6 (6). Available at SSRN: https://ssrn.com/abstract=3194449.

Murschetz P.C., Bruneel C., Guy J.-L, Haughton D., Lemercier N, McLaughlin M.-D., Mentzer K., Vialle Q., Zhang C., Murschetz P.C., & Bakhtawar B. (2020). Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success // *Nordic Journal of Media Management*. No. 1 (3). https://journals.aau.dk/index.php/NJMM/article/view/5871. DOI: 10.5278/njmm.2597-0445.5871.

Olubukola D., Stephen O., Funmilayo A., Ayokunle O., Oyebola A., Oduroye A., Wumi A., Yaw M. (2021). Movie Success Prediction Using Data Mining // *British Journal of Computer, Networking and Information Technology*. No. 4. Pp. 22–30. DOI: 10.52589/BJCNIT-CQOCIREC.

Sivakumar P., Rajeswaren V., Abishankar K., Ekanayake J., Mehendran Y. (2021). Movie Success and Rating Prediction Using Data Mining Algorithms // *Conference: International Research Conference of Uva Wellassa University* (IRCUWU-2020). https://www.researchgate.net/publication/349586116_Movie_Success_and_Rating_Prediction_Using_Data_Mining_Algorithms DOI: 10.13140/RG.2.2.14697.42085.

Souza T., Nishijima M., Pires, R. (2023). Revisiting predictions of movie economic success: random Forest applied to profits // *Multimed Tools Appl*. https://www.researchgate.net/publication/369591522_Revisiting_predictions_of_movie_economic_success_random_Forest_applied_to_profits. DOI: 10.1007/s11042-023-15169-4.

Tripathi J., Tiwari S., Saini A., Kumari S. (2023). Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data//*Indonesian Journal of Electrical Engineering and Computer Science*. March 2023. Vol. 29. No. 3. Pp. 1750–1757. DOI: 10.11591/ijeecs.v29.i3.pp1750-1757.

Wi. J., Jang S., Kim Y. (2020). Poster-Based Multiple Movie Genre Classification Using Inter-Channel Features // *IEEE Access.* Vol. 8. Pp. 66615–66624. DOI: 10.1109/ACCESS.2020.2986055.

Wu S., Zheng Y., Wu F., Zhan C.. (2019). Movie box office prediction based on ensemble learning. *IEEE Symposium on Product Compliance Engineering — Asia* (ISPCE-CN). https://ieeexplore.ieee.org/abstract/document/8958631. DOI: 10.1109/ISPCE-CN48734.2019.8958631.

You J.K., Yun G.Ch., J.ung H.L. (2019). Prediction of a Movie's Success From Plot Summaries Using Deep Learning Models // *Proceedings of the Second Workshop on Storytelling*. Association for Computational Linguistics. Florence, Italy. Pp. 127–135. DOI: 10.18653/v1/W19-3414.

Yu Y., Liu J. (2022). Optimizing Film Companies' Marketing Strategy Using Blockchain and Recurrent Neural Network Model // *Computational Intelligence and Neuroscience*. Pp. 1–13. Available at https://www.hindawi.com/journals/cin/2022/4139074/. DOI.: 10.1155/2022/4139074.

Yun K.O. (2017). The Impact of Initial eWOM Growth on the Sales in Movie Distribution// *Journal of Distribution Science*. No. 15. Pp. 85–93. DOI: 10.15722/jds.15.9.201709.85.

Дождиков Антон Валентинович
antondnn@yandex.ru

Anton Dozhdikov
Ph.D. (politics), independent researcher, Moscow
antondnn@yandex.ru

PREDICTION OF THE RESULTS OF MOVIE RELEASE USING MACHINE LEARNING[12]

Abstract. The subject of the research is the results of distribution of Russian national films. The purpose of the study is to classify projects according to the principle of their success/failure at the box office and predict the characteristics of the box office. The objectives of the study are to create algorithms for selecting (classifying) potentially successful projects into an investment portfolio and predicting (regression) rental characteristics: the number of views, payback, viewer rating. The technique is based on the application of ensemble machine learning models. The empirical base of the study is the entire set of Russian national films in distribution from 2004 to April 2022 (N=1469) and from May 2022 to April 2023 (N=194). Achieved accuracy of 0,95 and 0,89 for two and four class classification and high performance ROC_AUC = 0,97 for two class model and 0,94 — 0,98 for four class model. More complex metamodels (superensembles) can achieve an accuracy of 0,97–0,98 for a two-class classification and 0,96 for a four-class one. Complex regression metamodels predict the absolute values of payback, fees, views with a coefficient of determination (R2) in the range of 0,97–0,98 using synthetic data. As a result, it became possible to form investment portfolios of film projects with an annual historical return of up to 139%. The scope of application is to ensure the selection of films for investment "portfolios of film projects" of state (Ministry of Culture, Cinema Fund) and private investors. Machine learning models can be adapted to the conditions of global and foreign markets by increasing the number of model features, expanding the arsenal of machine learning methods, including the analysis of texts, images, videos, and user data of social networks.

Keywords: *national cinematography, investments, financing, machine learning, prediction, classification, regression, portfolio investment, movie rating, number of movie views, movie payback, movie box office success.*
JEL: G11, G17, Z11, C38, C53, C65, C45.

[12] Acknowledgements: Author expresses his gratitude to the staff of the ITMO Higher School of Digital Culture Mikhailova E.G., Grafeeva E.G., Egorova O.B., Boitsev A.A. Romanov A.A. for the knowledge gained; Organizing Committee of PLUS Media Holding, Business Development Director Grizov K.A. for the opportunity to present the practical results of the study and the industry financial magazine «PLUS» and for speaking at the International Forum «Fintech, Banks and Retail» on June 21–22 in the section «Big Data. Strategic capital of the 21st century»?