СТРЮКОВ П. В., ГЕРБЕРТ Д. В. ЭКОНОМЕТРИЧЕСКИЙ АНАЛИЗ ФАКТОРОВ ФОРМИРОВАНИЯ ПОТРЕБИТЕЛЬСКИХ ОЦЕНОК НА РЫНКЕ КИНОПРОДУКЦИИ: ИНТЕГРАЦИЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ТЕКСТОВОГО АНАЛИЗА

УДК 330.43:004.8, ГРНТИ 28.17.19
Статья поступила в редакцию 06.03.2026

Эконометрический анализ факторов формирования потребительских оценок на рынке кинопродукции: интеграция методов машинного обучения и текстового анализа

П. В. Стрюков, Д. В. Герберт

Ухтинский государственный технический университет, г. Ухта

В статье исследуются детерминанты пользовательских оценок в сфере стриминговых платформ и рекомендательных систем на основе массивов данных MovieLens 32M и Tag Genome. Научная новизна работы заключается в применении гибридной методологии: классический эконометрический инструментарий (МНК) дополнен методами обработки естественного языка (NLP) для оценки тональности текстовых рецензий и методом главных компонент (PCA) для извлечения латентных стилевых характеристик фильмов из пользовательских тегов. На выборке из 35 927 кинокартин показано статистически значимое влияние «эффекта престижа» и эмоционального фона отзывов на количественный рейтинг. Выявлен «парадокс блокбастера», при котором высокая технологичность картины (спецэффекты) оказывает

Econometric analysis of the factors of formation of consumer valuations in the film production market: integration of machine learning and text analysis methods

P. V. Stryukov, D. V. Gerbert

Ukhta State Technical University, Ukhta

The article examines the determinants of user ratings in the field of streaming platforms and recommendation systems based on MovieLens 32M and Tag Genome datasets. The scientific novelty of the work lies in the application of a hybrid methodology: the classical econometric toolkit (OLS) has been supplemented by natural language processing (NLP) methods for evaluating the tonality of text reviews and the principal component method (PCA) for extracting latent stylistic characteristics of films from user tags. A sample of 35,927 films proved the statistically significant influence of the "prestige effect" and the emotional background of reviews on the quantitative rating. The "blockbuster paradox" has been revealed, in which the high technology of the picture (special effects) has a negative impact on the average audience rating.

отрицательное влияние на среднюю оценку зрителей.

Ключевые слова: *эконометрика, машинное обучение, анализ тональности, метод главных компонент, рекомендательные системы, MovieLens, потребительское поведение*

***Keywords*:** *econometrics, machine learning, tonality analysis, principal component method, recommendation systems, MovieLens, consumer behavior*

Введение

В условиях цифровизации индустрии развлечений и перехода к экономике внимания, понимание механизмов формирования пользовательских рейтингов является критически важной задачей для настройки алгоритмов рекомендательных систем. Классические эконометрические подходы к анализу рынка киноиндустрии традиционно опирались на ограниченный набор метаданных: бюджет, жанр, кассовые сборы и наличие актеров-звезд.

Однако развитие методов глубокого обучения и анализа больших данных (Big Data) позволяет перейти от макропоказателей к микроуровню – анализу семантики пользовательских рецензий и краудсорсинговых тегов. Целью данного исследования является построение комплексной эконометрической модели, объясняющей вариацию рейтингов фильмов с учетом их жанровой специфики, скрытых стилевых факторов (генома) и тональности текстового отклика аудитории.

Данные и методология

Эмпирической базой исследования послужили два набора данных исследовательской группы GroupLens: *MovieLens 32M*, содержащий 32 млн пользовательских оценок [1], и *MovieLens Tag Genome 2021*, содержащий текстовые рецензии IMDB и статистику применения 1094 уникальных тегов [2]. Итоговая выборка после очистки данных и объединения массивов составила **35 927 фильмов**.

Концепция «Генома тегов» (Tag Genome)

Одной из главных проблем классического эконометрического анализа киноиндустрии является ограниченность метаданных. Стандартное деление на макро-жанры (драма, комедия, боевик) не способно уловить тонкие стилистические, атмосферные и нарративные различия картин. Для решения этой проблемы в исследовании применяется концепция «Генома тегов» (Vig et al., 2012) [3].

Геном тегов представляет собой плотное векторное представление (embedding) семантики фильма. В отличие от бинарного подхода (где тег либо присутствует, либо нет), геном определяет *степень релевантности* каждого

из 1094 уникальных тегов конкретному фильму. Теги охватывают широкий спектр характеристик: от визуального стиля («atmospheric», «neo-noir») и тропов («twist ending», «mindfuck») до субъективных оценок аудитории («overrated», «masterpiece»). Таким образом, каждый фильм i описывается вектором в 1094-мерном пространстве, что позволяет математически измерять смысловую близость картин и выделять микро-жанры.

Исследование проводилось в три этапа:

1. **Анализ тональности (Sentiment Analysis):** С помощью алгоритма VADER (Valence Aware Dictionary and sEntiment Reasoner) из библиотеки NLTK был обработан массив текстовых рецензий на платформе IMDB. Методология анализа тональности базируется на фундаментальных подходах к классификации эмоциональной окраски естественного языка [4].Для каждого фильма был вычислен индекс imdb_sentiment в диапазоне от -1 (максимально негативный) до +1 (максимально позитивный).

2. **Снижение размерности и извлечение латентных признаков (PCA):** Исходная матрица частот применения тегов является сильно разреженной и подвержена смещению из-за неравномерной активности пользователей. Для стабилизации дисперсии к исходным частотам $c_{i,j}$ (где c – количество применений j-го тега к i-му фильму) было применено логарифмическое сглаживание:

$$x_{i,j} = \ln(1 + c_{i,j}) \qquad (1)$$

Далее, для преодоления проблемы мультиколлинеарности и «проклятия размерности» в регрессионной модели, матрица сглаженных частот X была стандартизирована, после чего к ней применен метод главных компонент (PCA) [5]. Задача PCA заключалась в поиске ортогональных линейных комбинаций исходных тегов, максимизирующих объясненную дисперсию:

$$F_k = XW_k \qquad (2)$$

где F_k – вектор значений k-й главной компоненты (латентного стилевого фактора), X – стандартизированная матрица тегов, W_k – собственный вектор ковариационной матрицы X^TX, соответствующий k-му по величине собственному значению.

В результате пространство из 1094 тегов было редуцировано до 20 главных компонент (k=1...20). Анализ факторных нагрузок (W_k) позволил интерпретировать первые три латентные переменные:

Фактор 1 (Престиж/Атмосфера): максимальные веса у тегов «imdb top 250», «atmospheric», «stylized».

Фактор 2 (Технологичный блокбастер): «sci-fi», «special effects», «action».

Фактор 3 (Интеллектуальный триллер): «mindfuck», «twist ending», «complicated».

3. Эконометрическое моделирование: Оценка параметров уравнения множественной линейной регрессии производилась методом наименьших квадратов.

Результаты разведочного анализа данных (EDA)

Анализ распределения эмоционального фона (Рисунок 1) выявил выраженную J-образную кривую с доминирующим пиком в зоне экстремально положительных оценок (+1.0). Данный феномен объясняется эффектом «смещения самоотбора»: зрители склонны инвестировать время в написание текстовой рецензии преимущественно в случае сильного эмоционального отклика (восторга), в то время как «нейтральное» кино чаще игнорируется комментаторами.

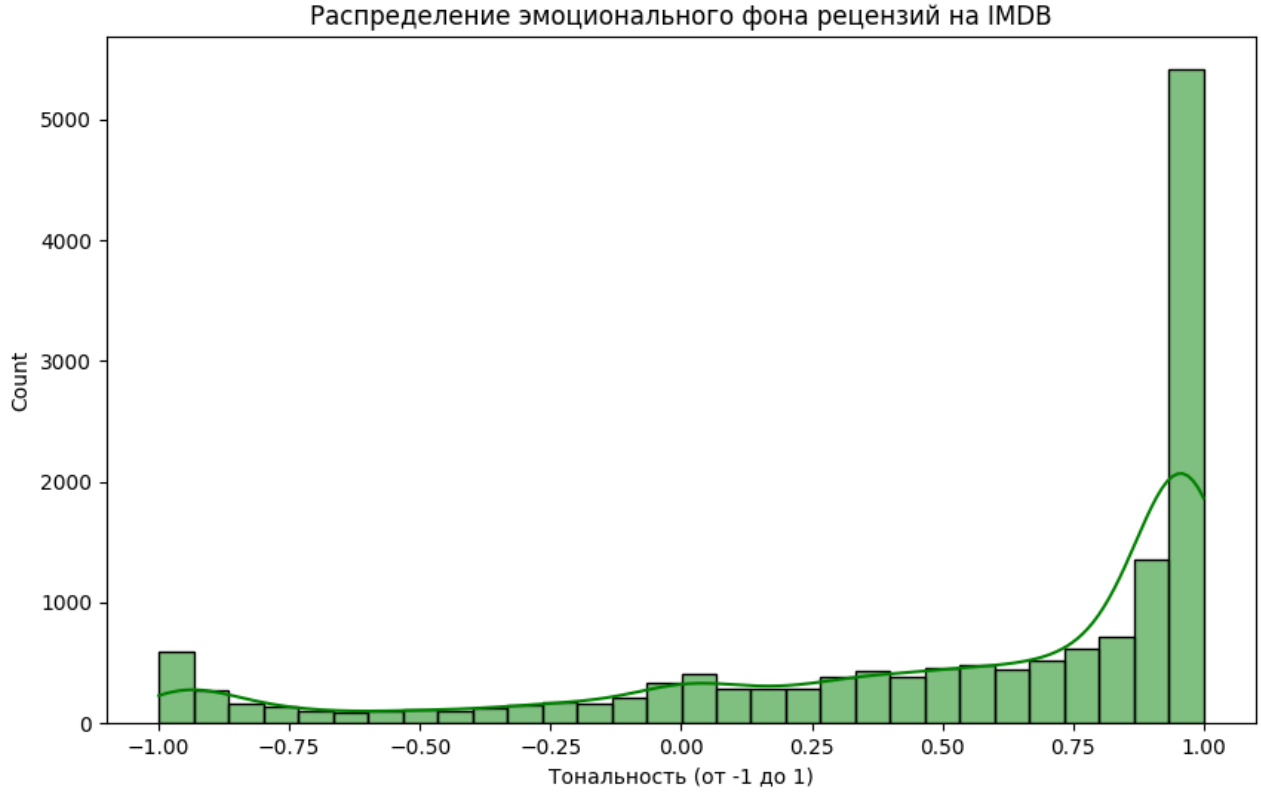

Рисунок 1. Распределение тональности текстовых рецензий на IMDB

Анализ жанровой динамики продемонстрировал устойчивые структурные сдвиги в предпочтениях аудитории. Как видно из тепловой карты (Рисунок 2) и графика рыночных долей (Рисунок 3), жанр «Драма» исторически сохраняет наиболее высокие средние оценки (стабильно выше 3.6 балла) и удерживает лидирующую долю в общем объеме оценок. В то же время жанр «Хоррор» демонстрирует максимальную волатильность и исторически подвергается наибольшему дисконту со стороны пользователей.

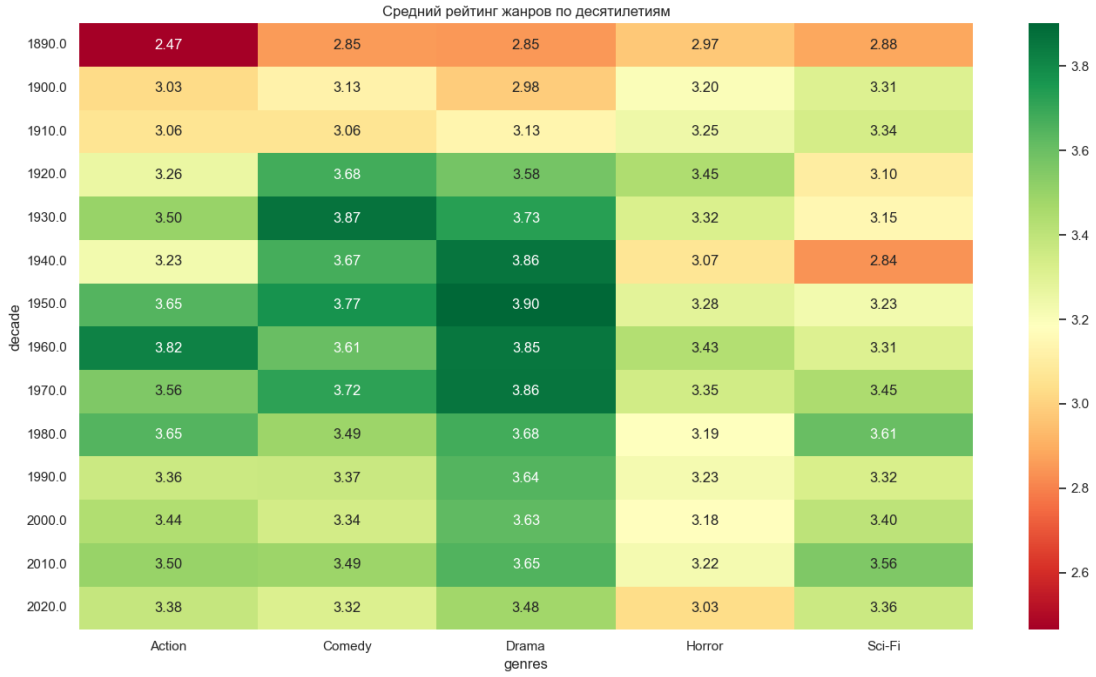

Рисунок 2. Динамика и структура жанровых предпочтений (1950–2023 гг.)

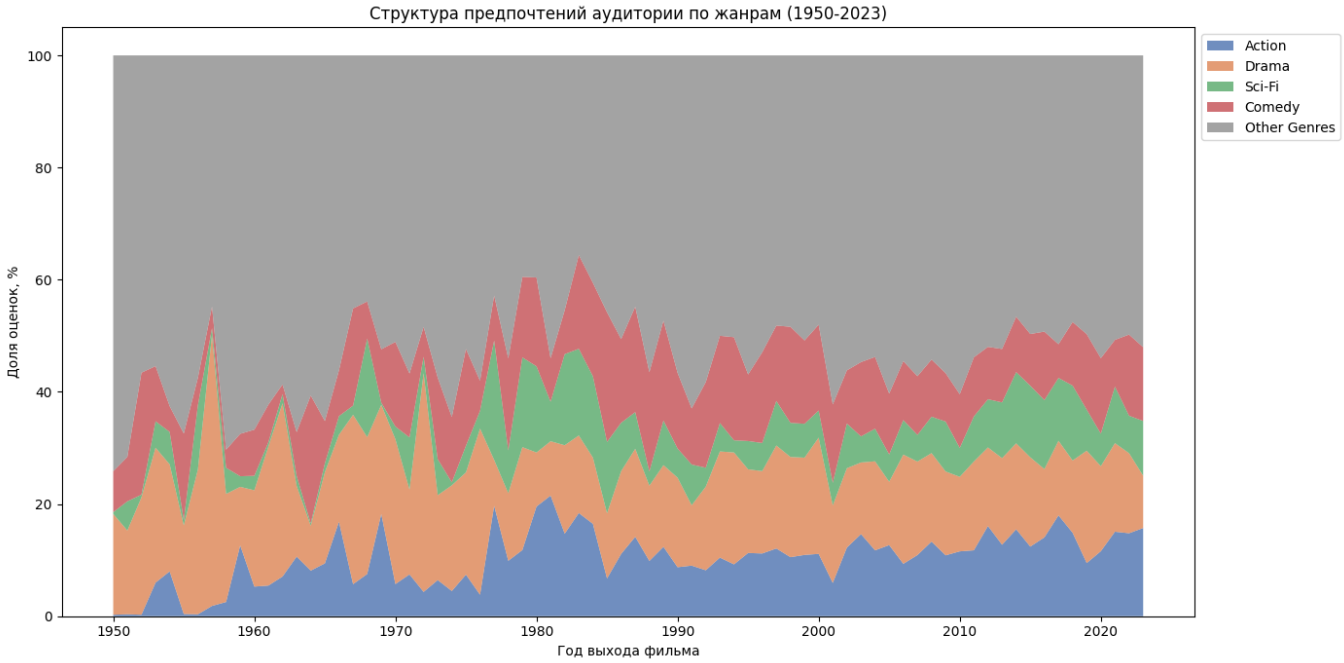

Рисунок 3. Структура предпочтений аудитории по жанрам

При сопоставлении популярности (количества оценок) и среднего рейтинга (Рисунок 4) подтверждается наличие эффекта конвергенции: нишевые фильмы (до 1000 оценок) имеют широчайший разброс качества, однако картины класса «мейнстрим» (от 10 000 оценок) формируют плотное облако в диапазоне от 3.0 до 4.2 баллов, что свидетельствует о сглаживании экстремумов при росте размера выборки зрителей.

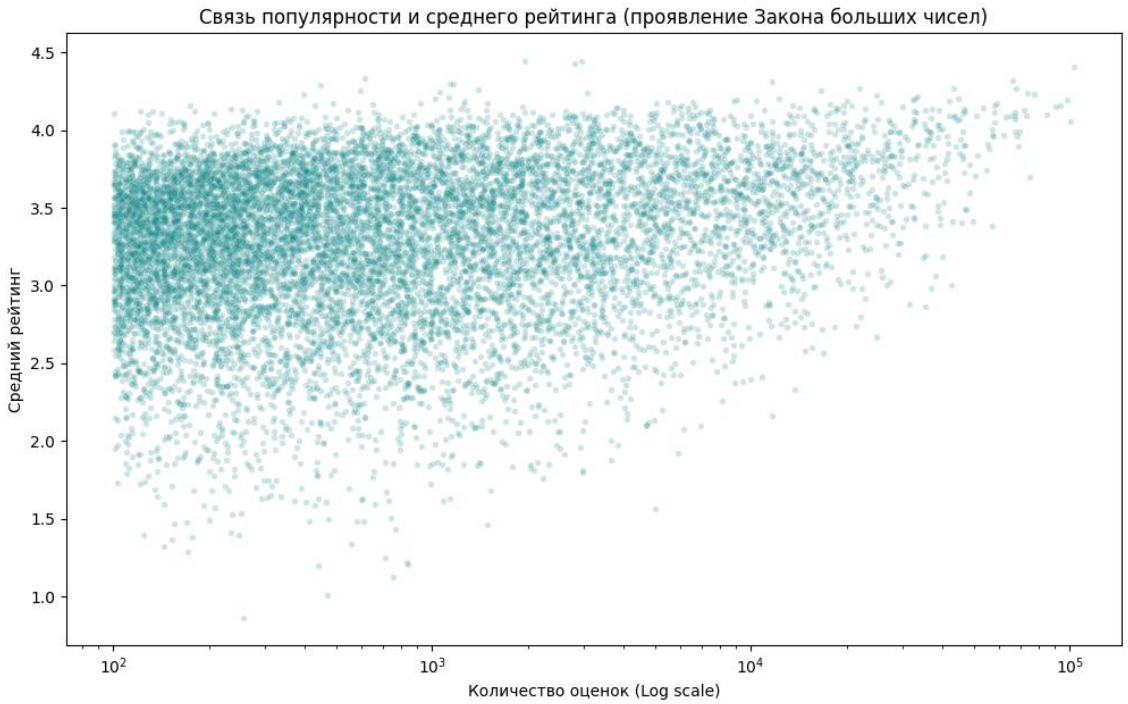

Рисунок 4. Корреляция между популярностью картины и ее средним рейтингом

Результаты регрессионного анализа

Для количественной оценки влияния выявленных латентных характеристик и эмоционального фона на потребительские предпочтения была специфицирована следующая эконометрическая модель множественной линейной регрессии:

$$Rating_i = \beta_0 + \beta_1 Sentiment_i + \sum_{k=1}^{5} \gamma_k F_{k,i} + \delta Year_i + \sum_{m=1}^{M-1} \theta_m Genre_{m,i} + \varepsilon_i \quad (3)$$

Где:

$Rating_i$ – зависимая переменная, средняя пользовательская оценка i-го фильма на платформе MovieLens;

- $Sentiment_i$ – индекс тональности текстовых рецензий IMDB (от -1 до 1);
- $F_{k,i}$ – значения первых пяти латентных стилевых факторов (главных компонент генома) для i-го фильма;
- $Year_i$ – год выпуска картины (контрольная переменная временного тренда);
- $Genre_{m,i}$ – вектор фиктивных переменных для макро-жанров, где M – общее количество жанров (один жанр выступает базовой категорией во избежание ловушки фиктивных переменных);
- $\beta,\gamma,\delta,\theta$ – подлежащие оценке параметры модели;
- ε_i – случайная ошибка (возмущение), удовлетворяющая классическим предпосылкам Гаусса-Маркова [6].

Оценка параметров модели производилась методом наименьших квадратов (МНК), который обеспечивает наилучшие линейные несмещенные оценки (BLUE) при соблюдении условий гомоскедастичности [7].

Для проверки гипотез была оценена спецификация OLS-модели, где зависимой переменной выступал средний рейтинг фильма. Результаты оценки представлены в Таблице 1. Модель является значимой в целом (F-statistic = 255.0, p-value < 0.001) и объясняет 15.6% вариации ($R2$=0.156), что является высоким показателем для стохастических данных о человеческом поведении.

Таблица 1. Результаты эконометрического моделирования

Объясняющая переменная (Фактор)	**Коэффициент (β)**	**Ст. ошибка (SE)**	**t-статистика**
Константа (Intercept)	4.400***	(0.255)	17.26
Латентные факторы стиля (PCA):			
Фактор 1: Престиж и Атмосфера	0.028***	(0.001)	46.29
Фактор 2: Спецэффекты и Экшн	-0.014***	(0.001)	-14.16
Фактор 3: Сложный сюжет	-0.009***	(0.001)	-8.51
Фактор 4: *Незначимый стиль*	-0.000	(0.001)	-0.02
Текстовая тональность:			
Тональность рецензий (imdb_sentiment)	0.138***	(0.010)	14.45
Временной тренд:			
Год выпуска (year)	-0.001***	(<0.001)	-5.41
Контрольные переменные жанров:			
Документальное кино (Documentary)	0.404***	(0.020)	20.44
Драма (Drama)	0.121***	(0.017)	6.92
Хоррор (Horror)	-0.445***	(0.020)	-21.80
Комедия (Comedy)	-0.020	(0.018)	-1.11
Мелодрама (Romance)	-0.012	(0.042)	-0.27

*Примечание: *** обозначает статистическую значимость на уровне 1% (p < 0.01), ** – на уровне 5% (p < 0.05), * – на уровне 10% (p < 0.1). В скобках указаны робастные стандартные ошибки. Базовой (скрытой) категорией для жанров выступает переменная Action/Без жанра.*

Метрики качества модели:

- Количество наблюдений (N): 35 927
- Коэффициент детерминации (R^2): 0.156
- F-статистика Фишера: 255.0 (p < 0.001)

Эконометрическая интерпретация

1. **Текстовые данные обладают предсказательной силой.** Коэффициент при переменной imdb_sentiment положительный и высокозначимый (β = 0.138). Изменение тональности текстовых рецензий в сторону позитива на 1 единицу коррелирует с увеличением рейтинга в MovieLens на 0.138 балла (при прочих равных).
2. **Парадокс технологичности.** Анализ латентных факторов показал, что «Атмосферность и Престиж» (style_factor_1) дают максимальную положительную премию к рейтингу, выступая индикатором социального доказательства. Однако фактор «Спецэффекты и Экшн» (style_factor_2) имеет значимый отрицательный коэффициент (β = -0.0135). Это означает, что при прочих равных условиях массовые высокобюджетные аттракционы оцениваются пользователями строже, чем кино других стилистик.
3. **Жанровые дисконты.** Документальное кино и Драмы получают наибольшую «надбавку» к оценке, в то время как Хорроры (β = -0.445) и Фантастика (β = -0.373) штрафуются аудиторией сильнее всего.
4. **Временной тренд.** Фактор года (year) демонстрирует отрицательную зависимость, что свидетельствует о наличии «инфляции требований» – со временем средние оценки, выставляемые новинкам, статистически падают.

Обсуждение результатов и ограничения исследования

На первый взгляд может показаться, что выявленная положительная связь между тональностью текстовых рецензий и итоговым рейтингом носит тавтологичный характер (проблема эндогенности), так как обе метрики отражают зрительскую симпатию. Однако научная ценность предложенной модели заключается в кроссплатформенной валидации и изоляции эффектов.

Во-первых, модель транслирует качественные оценки с платформы IMDB (где превалируют развернутые тексты критиков и киноманов) в количественные рейтинги MovieLens (где оценки ставят миллионы рядовых зрителей), статистически связывая поведение «пишущего меньшинства» с «голосующим большинством».

Во-вторых, множественная регрессия позволяет оценить влияние стилевых и жанровых факторов при фиксированном эмоциональном фоне. Модель доказывает наличие устойчивых «жанровых пенальти» (например, для фильмов ужасов), которые срабатывают и занижают итоговый рейтинг картины даже в том случае, если она получает высоко-позитивные текстовые рецензии.

Важным методологическим ограничением модели является классическая для рекомендательных систем «проблема холодного старта». Для расчета прогноза алгоритму требуются первичные данные о семантике фильма (матрица пользовательских тегов) и текстовом отклике (рецензии). Следовательно, модель не может быть применена к картинам в день их премьеры. Предложенный эконометрический инструментарий наиболее эффективен для прогнозирования долгосрочного («хвостового») рейтинга фильмов на основе их раннего цифрового следа, когда первоначальный пул тегов и отзывов экстраполируется на массовую аудиторию.

Заключение

Использование методов извлечения признаков из краудсорсинговых тегов и анализа тональности текстов позволило существенно обогатить классическую эконометрическую модель оценки кинопродукции. Результаты исследования доказали, что количественный рейтинг (звезды) неразрывно связан с семантикой вербального отклика аудитории.

Применение полученных результатов целесообразно в алгоритмах коллаборативной фильтрации стриминговых сервисов: учет выявленных стилевых факторов (генома) и "жанровых штрафов" позволит точнее предсказывать зрительский интерес и бороться с проблемой холодного старта для новых фильмов.

Список использованных источников и литературы

1. Harper F. M., Konstan J. A. The MovieLens Datasets: History and Context // ACM Transactions on Interactive Intelligent Systems (TiiS). 2015. Vol. 5, No. 4. P. 19:1–19:19.

2. Kotkov D., Maslov A., Neovius M. Revisiting the tag relevance prediction problem // Proceedings of the 44th International ACM SIGIR conference. 2021. P. 1768-1772.

3. Воронцов К. В. Машинное обучение и анализ данных // М.: МФТИ. – 2020.

4. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2, No. 1-2. P. 1-135.

5. Vig J., Sen S., Riedl J. The tag genome: Encoding community knowledge to support novel interaction // ACM Transactions on Interactive Intelligent Systems. 2012. Vol. 2, No. 3. P. 13:1–13:44.

6. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс. – 8-е изд. – М.: Дело, 2007. – 504 с.

7. Айвазян С. А. Методы эконометрики. – М.: Магистр: ИНФРА-М, 2010. – 512 с.

List of references

1. Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4), 19:1–19:19.

2. Kotkov, D., Maslov, A., & Neovius, M. (2021). Revisiting the tag relevance prediction problem. Proceedings of the 44th International ACM SIGIR conference, 1768-1772.

3. Vorontsov K. V. (2020). Machine learning and data analysis. Moscow: MIPT.

4. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

5. Vig, J., Sen, S., & Riedl, J. (2012). The tag genome: Encoding community knowledge to support novel interaction. ACM Transactions on Interactive Intelligent Systems, 2(3), 13:1–13:44.

6. Magnus Ya. R., Katyshev P. K., Peresetskiy A. A. (2007). Econometrics. Introductory course. Moscow: Delo.

7. Aivazian S. A. (2010). Methods of econometrics. Moscow: Magister: INFRA-M.