Подготовил
Копнев Максим Михайлович, БПМИ238

Вид проекта - исследовательский

Тип проекта - индивидуальный

Методы машинного обучения в предсказании рейтинга видеоигр

Machine learning methods in predicting video game ratings

Научный руководитель:
Алиев Мишан Хаммад оглы, эксперт, Лаборатория теоретических основ моделей искусственного интеллекта

Описание предметной области

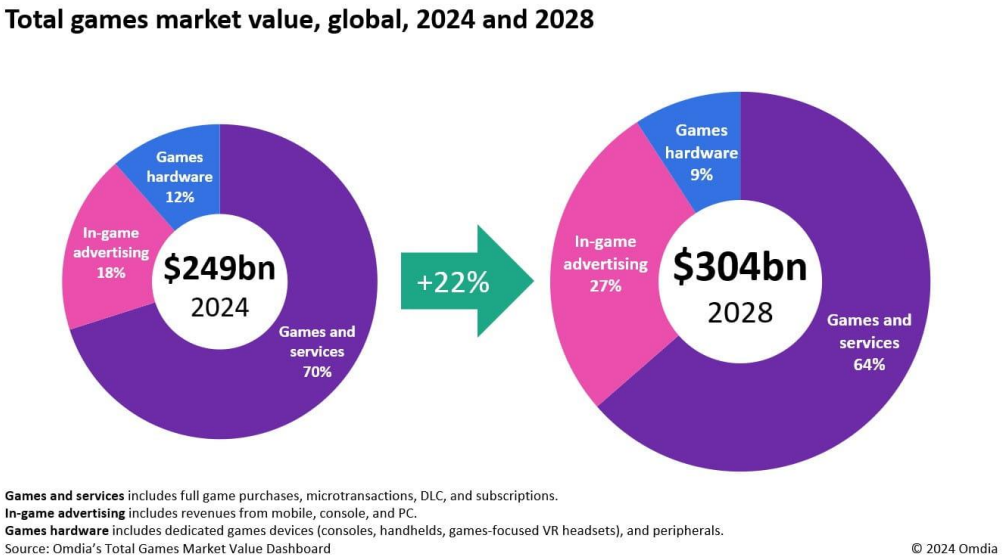
В основе исследования лежат модели машинного обучения: линейная регрессия, решающие деревья и градиентный бустинг, которые применяются для предсказания рейтинга видеоигр по общим данным о ней и отзывам пользователей.

Актуальность работы

Индустрия видеоигр активно развивается, а на рынке ежемесячно появляются десятки новых проектов.

Актуальность работы заключается в следующем:

- Оценка эффективности методов машинного обучения в предсказании рейтинга видеоигр.
- Поиск и обоснование зависимостей в данных.
- Помощь пользователям в выборе лучшего проекта из имеющихся на рынке.



Цель работы

Основной целью исследования является разработка и анализ моделей машинного обучения для прогнозирования рейтингов видеоигр на основе данных в открытом доступе, а также оценка факторов, влияющих на качество предсказания.

Гипотезы

В исследовании выдвинуты и проверены следующие гипотезы:

- Модели градиентного бустинга покажут более высокую точность прогнозирования рейтинга игр по сравнению с линейными моделями.
- Наиболее значимыми факторами в предсказании рейтинга игр являются отзывы пользователей и платформа, на которой выпущена игра.
- Величина студии-разработчика не обязательно находится в прямой зависимости с рейтингом выпускаемых ею игр.

Задачи работы

1. Работа с данными, взятыми с kaggle.com: удаление лишних признаков и пропусков и обработка категориальных признаков с сохранением их интерпретируемости.
2. Определение средней тональности отзывов с использованием NLTK.
3. Обучение моделей на двух датасетах, один из которых содержит данные о тональности отзывов.
4. Оценка эффективности каждой из моделей посредством различных метрик и построения графиков, определение лучшей модели.
5. Построение диаграммы важности признаков и её анализ, подтверждение или опровержение гипотез, выводы о применимости выбранных методов.

Анализ аналогичных работ

Было найдено схожее с нашим исследование ”Prediction and classification of video games” [1], цель которого - создание модели для предсказания глобальных продаж видеоигр на основе общей информации о них и поиск комбинаций характеристик, максимизирующих продажи.

Основное отличие данного исследования от нашего в объекте предсказания: в [1] рейтинг используется для предсказания наряду с остальными данными. Кроме того, в нашем исследовании используется тональность отзывов игроков, что позволяет улучшить качество предсказания.

Анализ особенностей исследования

Для задачи предсказания рейтинга видеоигра в настоящий момент не найдено наиболее подходящего метода. Это связано в первую очередь с тем, что широкий интерес к видеоиграм возник сравнительно недавно, в том числе из-за резкого увеличения их рынка в 2020-2021 году. Ввиду ограниченного количества данных и обилия характеристик игр, являющихся категориальными признаками, задача предсказания сильно усложняется.

Анализ используемых методов

В исследовании используются наиболее популярные методы машинного обучения, такие как линейная регрессия, решающие деревья и градиентный бустинг, из модуля scikit-learn. Для определения тональности отзывов используется лингвистический метод, реализованный в NLTK. Для обработки данных взяты модули pandas и numpy.

Процесс исследования

Обработка данных

Для обучения моделей были выбраны следующие признаки: год выпуска, жанры, платформа и студия-разработчик. Для их обработки были применены следующие методы:

- Для платформы и года выпуска ввиду низкого количества уникальных значений был применен One-Hot Encoding (OHE).
- Разработчики были поделены на 26 групп по количеству выпущенных игр, чтобы сохранить интерпретируемость признака. Затем был применен OHE.
- Каждый жанр игры был интерпретирован как отдельный бинарный признак.

Каждому отзыву была сопоставлена его тональность, которая затем усреднялась по каждой игре.

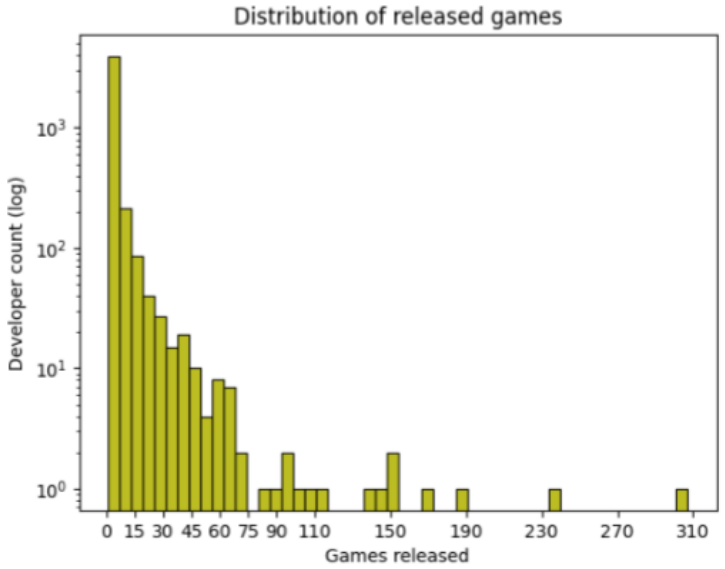


Рисунок 3.1 – Распределение выпущенных игр в зависимости от количества разработчиков

Процесс исследования

Разбиение данных

Обработанные данные были объединены в два датасета, различавшихся лишь наличием признака тональности отзывов.

Для обучения моделей данные были разделены на обучающие и тестовые выборки в пропорции 80/20 с применением параметра stratify для сохранения распределения данных в выборках.

Было обнаружено сильное смещение данных в сторону диапазона [6, 9], объясняемое особенностями выставления рейтинга игр.

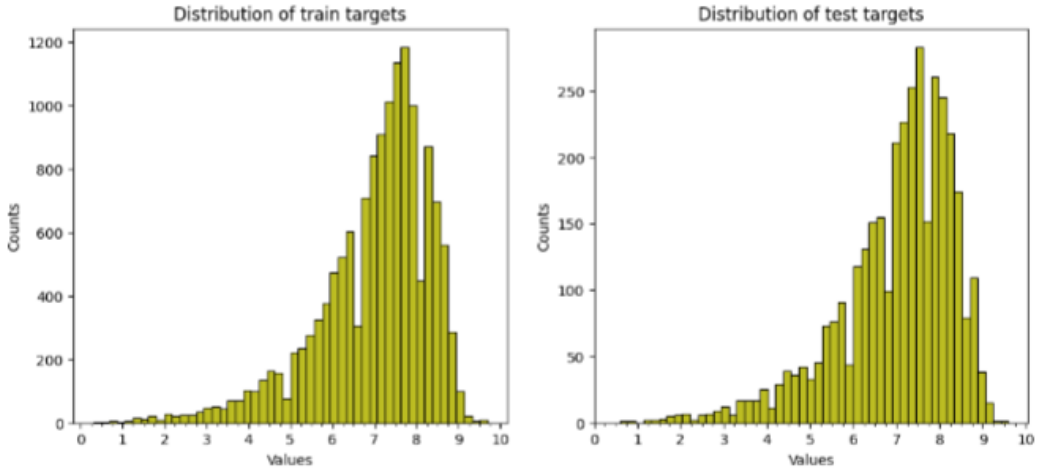


Рисунок 4.2 – Распределение значений целевой переменной на выборках

Процесс исследования

Обучение моделей

Каждая из моделей была обучена на обоих датасетах с использованием модуля sklearn. После для всех моделей были посчитаны средняя абсолютная ошибка в процентах (MAPE), корень из среднеквадратического отклонения (RMSE) и коэффициент детерминации (R^2).

Кроме того, была построена диаграмма сравнения реальных значений рейтинга с предсказанными на тестовой выборке.

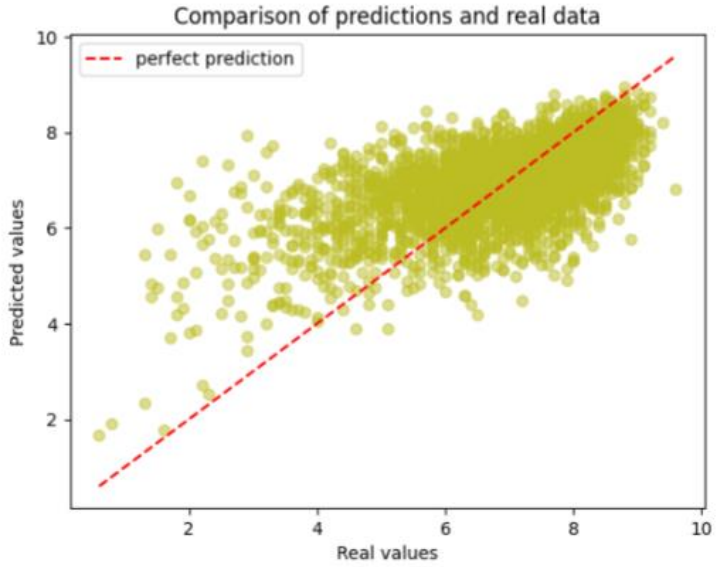


Рисунок 5.3 – Сравнение реальных и предсказанных значений для градиентного бустинга с отзывами. Зеленые точки - положение объекта в координатах (реальный рейтинг, предсказанный рейтинг), красная линия - идеальное предсказание.

Результаты исследования

Качество моделей

При рассмотрении полученных диаграмм для линейной регрессии и градиентного бустинга в динамике было замечено, что добавление данных об отзывах и использование градиентного бустинга дали наилучшие результаты в предсказании. Кроме того, модели с использованием тональности показали в общем лучшие результаты, чем их аналоги без этого признака.

	Предсказания без отзывов			Предсказания с отзывами		
	MAPE	RMSE	R^2-score	MAPE	RMSE	R^2-score
Линейная регрессия	17.7%	1.259	0.164	15.8%	1.144	0.279
Решающее дерево	18.4%	1.305	0.058	15.7%	1.161	0.257
Градиентный бустинг	16.6%	1.209	0.192	**14.2%**	**1.062**	**0.378**

Таблица 5.1 – Сравнение метрик для обученных моделей. Жирным выделены лучшие результаты.

Результаты исследования

Важность признаков

Важность средней тональности отзывов в разы превысила важность остальных признаков. При этом можно заметить:

- Платформы PS2, PS4 и PC часто использовались для предсказания, что соотносится с реальностью, так как PC считается лучшей платформой для гейминга, а данные консоли считаются одними из лучших.
- Группа студий-разработчиков, выпустивших ровно 1 игру, попала в десятку самых важных признаков, что подтверждает одну из гипотез.

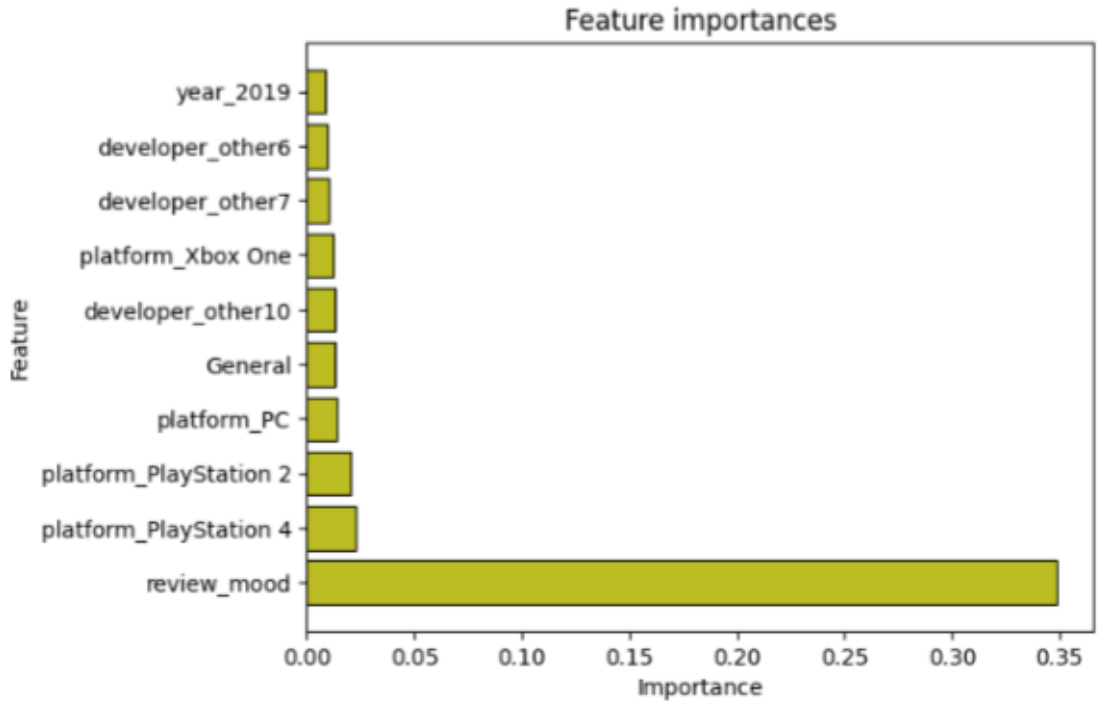


Рисунок 5.4 – Диаграмма важности признаков

Выводы

Из трёх рассмотренных методов лучше других справился с задачей метод градиентного бустинга на решающих деревьях. Однако качество предсказания лишь приблизилось к приемлемому. Это обуславливается наличием следующих проблем:

- Неравномерное распределение данных.
- Отсутствие числовых признаков кроме тональности.
- Сильная корреляция между тональностью отзывов и рейтингом, уменьшающая ценность модели как инструмента предсказания.

Были подтверждены все выдвинутые гипотезы.

На основе работы сделать вывод о сложности видеоигр как объекта для исследования: они, как объект искусства, плохо поддаются строгому математическому описанию.

Направления дальнейшей работы

Данное исследование предполагает продолжение работы в следующих направлениях:

- Самостоятельный сбор данных и конструирование признаков (feature engineering).
- Использование продвинутого NLP для оценки тональности не специализированных текстов (комментариев и постов в социальных сетях), а также обработка кратких описаний видеоигр.
- Использование иных методов предсказания, в том числе глубинного обучения в случае наличия достаточного количества данных.

Список использованных источников

[1] Hannah Igboke. Prediction and classification of video games., 2024. (дата обр. 1.04.2025).

[2] Kaggle: Machine learning and data science community. URL: https://kaggle.com. (дата обр. 1.04.2025).

[3] Steam store. https://store.steampowered.com/, 2003. (дата обр. 1.04.2025).

[4] Учебник по машинному обучению. Яндекс Образование. https://education.yandex.ru/handbook/ml, 2021. (дата обр. 1.04.2025).

[5] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning - From Theory to Algorithms. Cambridge University Press, 2014.

[6] Кострикин А.И. Введение в алгебру в 3 томах. МЦНМО, 2012.

[7] Он Ч. Дайзенрот М., Фейзал А. Математика в машинном обучении. Питер, 2024.

[8] Фихтенгольц Г. М. Основы математического анализа в 2 томах. Наука, 1968.

[9] Самигулин Тимур Русланович и Джурабаев Анвар Эркин Угли. Анализ тональности текста методами машинного обучения. Научный результат. Информационные технологии,2021.


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •