РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет/институт: Факультет Искусственного интеллекта, РУДН

Кафедра: Искусственного интеллекта

ОТЧЁТ О ЛАБОРАТОРНОЙ РАБОТЕ

по дисциплине «Прикладная статистика и анализ данных»

Лабораторная работа № 2

Тема: Регрессия, мультиколлинеарность, множественные сравнения, байесовские выводы и ресемплинг

Студент:	Тараканов Борис Александрович, ЗФИмд-01-25, Управление данными и Искусственный интеллект
Преподаватель:	Курашкин Сергей Олегович, доцент, кандидат технических наук
Дата выполнения:	«26» октября 2025 г.
Оценка/подпись:	_____

Москва — 2025

СОДЕРЖАНИЕ

Введение

Цель работы — построить и сравнить несколько спецификаций множественной линейной регрессии, диагностировать мультиколлинеарность и применить регуляризацию; корректно проверять семейства гипотез (контрасты/коэффициенты) с контролем FDR; провести байесовскую регрессию с априорами и постериорной проверкой предсказаний; оценить доверие к выводам с помощью бутстрепа и перестановочных тестов.

Объект исследования — датасет MovieLens-Rating 100k. Он представляет собой выборку исторических оценок фильмов пользователями сервиса MovieLens.

Ожидаемые результаты уметь (1) задавать и интерпретировать линейные модели с взаимодействиями; (2) диагностировать и снижать мультиколлинеарность (VIF, Ridge/Lasso с подбором λ по CV); (3) применять FDR-контроль к семействам проверок; (4) формулировать априоры и читать апостериор (PyMC), выполнять PPC; (5) строить бутстреп-ДИ и перестановочные p-значения; (6) оформлять воспроизводимый отчёт.

Теоретические основы

Множественная линейная регрессия описывает зависимость количественной целевой переменной от нескольких факторов. Модель оценивает вклад каждого признака при фиксированных значениях остальных, что позволяет интерпретировать «чистый» эффект предикторов. При наличии категориальных данных они кодируются в набор бинарных индикаторов. Взаимодействия признаков позволяют учитывать изменение эффекта одного

фактора в зависимости от другого (например, изменение популярности жанра с годами).

ANOVA и сравнение моделей применяются, когда нужно проверить, улучшает ли добавление группы признаков качество модели. Сравниваются вложенные спецификации: сначала строится базовая модель, затем более сложная. Анализ показывает, снижает ли новая группа факторов остаточную изменчивость. Если улучшение статистически значимо, расширенная модель считается более подходящей.

Контрасты и множественные проверки используются, когда оценивается сразу много коэффициентов или сравниваются группы факторных эффектов, например, жанры. При множественных проверках увеличивается риск ложных находок. Для контроля ошибки применяется FDR-коррекция (например, метод Бенджамини–Хохберга), которая ограничивает долю ложноположительных выводов среди всех признанных значимыми обнаружений.

Мультиколлинеарность возникает, когда признаки сильно связаны между собой. Это не мешает прогнозированию напрямую, но делает оценки параметров нестабильными, увеличивает вариацию коэффициентов и снижает доверие к их интерпретации. Диагностика выполняется через показатели инфляции дисперсии (VIF), отражающие, насколько у конкретного признака ухудшается точность оценки из-за связи с другими предикторами.

Регуляризация решает проблему нестабильных коэффициентов. В моделях Ridge и Lasso выполняется штрафование слишком больших по абсолютному значению коэффициентов. Ridge стабилизирует оценки и полезен при высокой корреляции признаков. Lasso может обнулять неинформативные

коэффициенты, выполняя отбор признаков. Значение штрафа подбирается по кросс-валидации.

Байесовская регрессия рассматривает параметры модели как случайные величины с априорными распределениями. Наблюдения обновляют априорные предположения, формируя апостериорное распределение коэффициентов. Вместо одной оценки параметров получается распределение возможных значений, что позволяет анализировать неопределённость. Предсказания также становятся вероятностными. Постериорно-предсказательная проверка оценивает, насколько модель способна воспроизводить наблюдаемые данные.

Ресемплинг используется для оценки стабильности выводов и качества прогноза без строгих теоретических предположений. В бутстрепе множество подвыборок создаются случайным выбором с возвращением, что позволяет строить доверительные интервалы для метрик и коэффициентов. Перестановочные тесты проверяют гипотезы, пересмешивая значения так, чтобы имитировать мир без искомого эффекта. Это даёт честные p-значения даже при нарушении классических допущений.

Описание данных и инструментов

В исследовании используется таблица оценок u.data, объединённая с метаданными фильмов u.item. Таблица u.data содержит строки вида «пользователь–фильм–оценка–временная метка», где user_id и movie_id выступают в роли идентификаторов, rating (целевой признак) принимает значения от 1 до 5, а timestamp позволяет восстановить год выставления оценки. Таблица u.item обеспечивает информацию о фильмах: название, дата релиза и набор бинарных жанровых колонок (по типу one-hot), каждая из которых принимает значение 0/1, показывая принадлежность фильма к конкретному жанру.

Из даты выпуска фильмов выделяется год релиза, а из временной метки рейтинга — год самой оценки. Оба признака центрируются (year_release_c, year_rate_c), чтобы уменьшить корреляцию с константой и избежать избыточной линейной зависимости при добавлении взаимодействий. Таким образом, в объединённом датафрейме каждая строка — это действие пользователя, снабжённое признаками фильма и моментом, когда оценка была выставлена.

Основные группы признаков включают бинарные индикаторы жанров, два временных континуума (год релиза и год оценки), а также позднее введённые out-of-fold leave-one-out признаки: средний рейтинг фильма (movie_mean_loo_oof) и мера активности пользователя (user_count_loo_oof). Эти признаки вычисляются строго без «подсматривания» в будущие строки: для каждого фолда по пользователям средние значения берутся только из тренировочной части. Такой дизайн необходим для предотвращения утечки информации, особенно критичной в задачах, где поведение пользователя само по себе является сигналом.

Для оценки моделей и построения признаков используются групповые разбиения (GroupKFold) по идентификатору пользователя, так как предсказание заранее неизвестных пользователей значительно усложняет задачу и уничтожает оптимистическое смещение. Итоговая матрица признаков включает: жанровые индикаторы, LOO-среднее фильма, LOO-активность пользователя и центрированный год оценки. Целевая переменная — численная оценка рейтинга. Таким образом, объект исследования представляет собой высокоразмерную смешанную панель транзакций «пользователь–фильм–время» с категориальной, числовой и временной структурой, что создаёт естественные условия для линейного моделирования, тестирования гипотез, регуляризации и байесовского анализа.

Программное обеспечение и библиотеки:

- **Python 3.12**

- **NumPy и Pandas** применяются для работы с данными: чтения файлов, объединения таблиц, обработки признаков, группировок и вычислений. NumPy обеспечивает эффективные операции над массивами, а Pandas — удобный табличный формат, необходимый для подготовки матрицы признаков.
- **Pathlib и datetime** используются для корректной работы с файловыми путями и преобразованием временных меток в календарные даты, что важно при построении временных признаков и трендов.
- **Scikit-learn** обеспечивает инструменты машинного обучения: кросс-валидацию (GroupKFold), стандартизацию признаков, построение регрессионных моделей (OLS, Ridge, Lasso), подбор гиперпараметров и вычисление метрик качества (MAE, RMSE, R^2). Через Pipeline объединяется предобработка и модель, что делает эксперимент воспроизводимым.
- **Statsmodels** используется для статистической регрессии, анализа коэффициентов и проверки гипотез. Позволяет работать с линейной моделью в классической постановке, получать стандартные ошибки, p-значения, таблицы ANOVA, а также выполнять статистические тесты на контрасты.
- **SciPy** дополняет статистический функционал, предоставляя распределения и функции для расчёта p-значений и критических уровней. Это нужно для корректного тестирования гипотез, включая F- и t-критерии.
- **Инструменты контроля множественных сравнений** (multipletests из statsmodels) применяются для FDR-коррекции, что необходимо при одновременной проверке большого числа коэффициентов, например, влияния жанров.
- **Variance Inflation Factor** из statsmodels используется для диагностики мультиколлинеарности, так как позволяет количественно оценить связь признаков между собой.
- **Matplotlib** отвечает за визуализацию: графики распределений, остатки регрессии, сравнения моделей и доверительные интервалы. Он используется для иллюстрации выводов и качества модели.

Все библиотеки использовались в последних стабильных версиях, доступных на момент выполнения работы.

Аппаратное обеспечение:

- **Процессор**: Intel Core i5-9600KF
- **GPU**: GeForce RTX 3070
- **Оперативная память**: 32 GB DDR4
- **Среда выполнения**: Jupyter Lab

Данная конфигурация обеспечила достаточную производительность для обработки крупного набора данных и проведения множественных экспериментов с кросс-валидацией.

Методика и план эксперимента

- Загрузка MovieLens-100k, объединение таблиц, первичная чистка и извлечение года релиза/оценки.
- Формирование признаков: жанры (OHE), центрированные временные признаки, OOF-LOO для фильма и пользователя (с GroupKFold).
- Базовая линейная модель OLS; анализ коэффициентов и качества (RMSE, R^2).
- Добавление взаимодействий жанр×год; сравнение вложенных моделей через F-тест, ANOVA и контрасты с FDR-коррекцией.
- Байесовская линейная модель (аналитическое сопряжённое решение, вместо PyMC по причине несовместимости оборудования); оцениваются постериоры и PPC.
- Смешанная модель (случайные эффекты пользователя/фильма), проверка согласованности распределений предсказаний.

- Бутстреп-интервалы для RMSE и коэффициентов (классический и кластерный по user_id); бутстреп-сравнение моделей.
- Перестановочные тесты: значимость признаков, sign-flip сравнение моделей и permutation feature importance.

Результаты и их анализ

Предобработка данных

Для оценки моделей и построения признаков используются групповые разбиения (GroupKFold) по идентификатору пользователя, так как предсказание заранее неизвестных пользователей значительно усложняет задачу и уничтожает оптимистическое смещение. Итоговая матрица признаков включает: жанровые индикаторы, LOO-среднее фильма, LOO-активность пользователя и центрированный год оценки. Целевая переменная — численная оценка рейтинга. Таким образом, объект исследования представляет собой высокоразмерную смешанную панель транзакций «пользователь–фильм–время» с категориальной, числовой и временной структурой, что создаёт естественные условия для линейного моделирования, тестирования гипотез, регуляризации и байесовского анализа.

Линейные модели и ANOVA: спецификация и сравнение

Базовая линейная модель, включающая жанровые индикаторы, усреднённый рейтинг фильма (LOO), активность пользователя и временной тренд, объясняет около 17.5% дисперсии рейтинга. Это ожидаемо для задач рекомендаций: значительная часть вариации зависит от индивидуальных вкусов, которые в модели отсутствуют в явном виде.

Наиболее значимый предиктор — средний рейтинг фильма (movie_mean_loo_oof): его высокий t-статистический результат отражает мощный эффект «коллективных оценок». Активность пользователя

(user_count_loo_oof) также значима, но с отрицательным знаком: больше оценивающие пользователи ставят несколько ниже средний балл. Учитывание временного тренда показывает снижение средних оценок со временем. Некоторые жанровые переменные имеют значимые эффекты (например, Drama, Animation, Children's), но они заметно слабее эффекта LOO-признаков.

Важно отметить предупреждение о мультиколлинеарности: большое значение условного числа предполагает возможную линейную зависимость среди жанров, но на данном этапе это не мешает оценке общей тенденции.

Добавление взаимодействий «жанр × год оценки»: Модель расширяется за счёт взаимодействий, отражающих изменение восприятия жанров во времени. Величина R^2 почти не меняется (0.176 вместо 0.175), прирост минимален, что указывает на небольшой дополнительный вклад взаимодействий. Однако на уровне статистики остатков изменения значимы.

Несколько взаимодействий оказываются статистически значимыми: например, Animation × год оценки и Crime × год оценки. Это можно интерпретировать как изменение популярности отдельных жанров. Тем не менее многие взаимодействия имеют большие стандартные ошибки, что указывает на шум и пересечение жанровых эффектов, а также усиливает мультиколлинеарность (условное число растёт до 10^{15}).

Проверка вложенных спецификаций через частичный F-тест показывает, что добавление взаимодействий статистически улучшает модель ($p \approx 0.0087$). ANOVA-таблица подтверждает, что прирост объяснённой вариации мал, но значим при большом размере выборки.

Таким образом, взаимодействия вносят небольшой, но статистически подтверждённый вклад. Это частично ожидаемо: слабые, но стабильные жанровые тренды могут проявляться только на больших данных.

Контраст Comedy vs Drama показывает, что Drama получает ощутимо более высокую оценку, чем Comedy (разница около −0.047, p ≈ 1e-07). Это важный пример интерпретируемой гипотезы: жанры действительно различаются по восприятию, и этот эффект статистически устойчив.

При контроле доли ложных находок (BH-FDR) значимыми остаются: Drama, Children’s, Animation, War, Western.

После коррекции часть жанров, выглядевших значимыми по обычным p-значениям, уже не проходят порог (например, Romance, Film-Noir). Это подчёркивает важность FDR-коррекции при анализе тематических факторов: простые p-значения переоценивают количество «значимых жанров».

Мультиколлинеарность, VIF и регуляризация

Сначала была проведена диагностика мультиколлинеарности с помощью показателя VIF для расширенной спецификации, включающей жанры, временные признаки и взаимодействия «жанр × год оценки». Таблица VIF показала две аномально проблемные переменные: индикатор жанра unknown и соответствующее взаимодействие unknown_x_year_rate. Для обеих величин VIF формально оказался бесконечным, а в матрице парных корреляций обнаружена почти идеальная по модулю корреляция между ними. Это означает, что один из признаков практически является линейной комбинацией другого (по сути, информация дублируется), что делает дизайн-матрицу близкой к вырожденной и сильно раздувает стандартные ошибки. Остальные признаки имели VIF заметно меньше порогового уровня 10: для временного тренда year_rate_c значение около 9.4, для жанров и взаимодействий — в районе от 1.3 до 3.1. То есть проблема мультиколлинеарности локализована в одной узкой группе признаков.Чтобы проиллюстрировать практическое исправление, был запущен итеративный алгоритм по удалению признаков с максимальным VIF при защите

важных ковариат (год релиза и год оценки). На первой итерации он удаляет именно жанр unknown, после чего число признаков уменьшается, а максимальный VIF значительно падает. В результате дизайн становится условно «здоровым» по критерию VIF, хотя небольшая корреляция между жанрами и их взаимодействиями, естественно, сохраняется. Показательно, что для остальных жанров VIF остаётся в диапазоне 1–2, что говорит об умеренной зависимости и вполне нормальной интерпретируемости коэффициентов.

Далее рассматриваются результаты регуляризации. Для Ridge-регрессии был проведён подбор параметра штрафа по сетке значений на логарифмической шкале с использованием групповой кросс-валидации по пользователям. Кривая CV-ошибки (MSE) практически плоская при малых значениях штрафа и начинает заметно улучшаться при переходе к более крупным альфам, вплоть до максимального значения в сетке (Рис. 1). Это говорит о том, что слабая регуляризация почти не меняет ситуацию по сравнению с OLS, а сильная регуляризация слегка улучшает устойчивость предсказаний. Лучшее значение альфы по кросс-валидации оказалось довольно большим (порядка тысячи), что отражает желание модели агрессивно «подтянуть» коэффициенты к нулю, сглаживая шум в данных и устраняя последствия мультиколлинеарности.

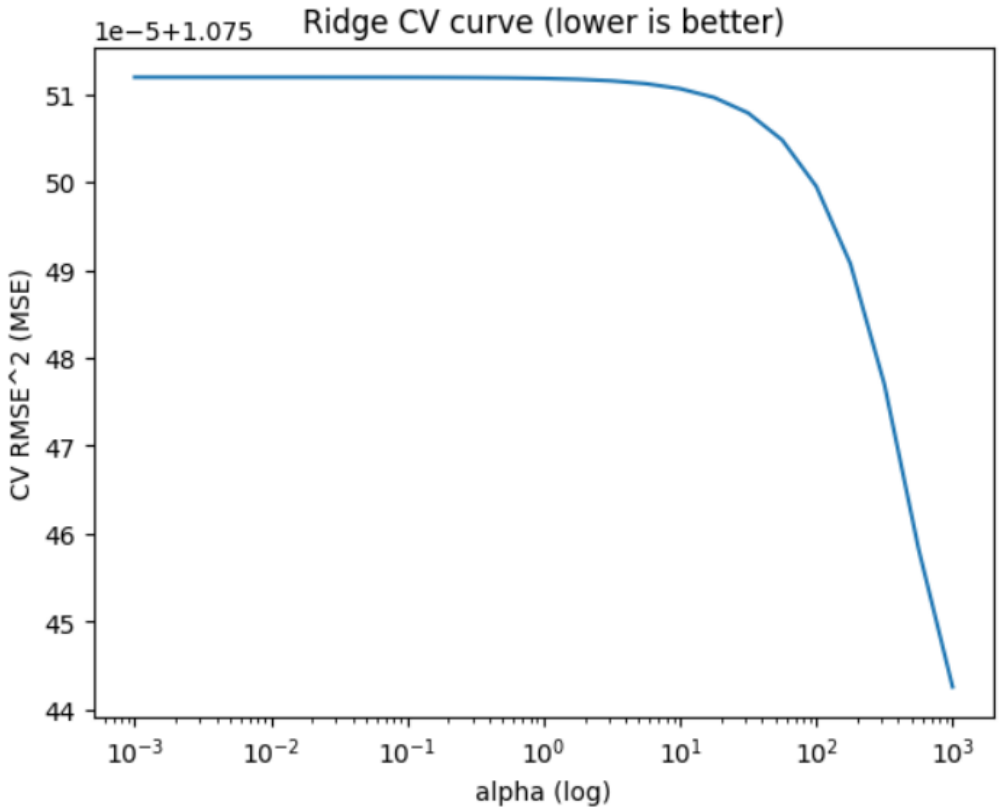

Рисунок 1: Кривая CV-ошибки (MSE) для Ridge

Для Lasso-регрессии картина иная. Кривая кросс-валидации показывает минимум ошибки при очень малых значениях штрафа и резкий рост при увеличении альфы (Рис. 2). Это означает, что жёсткое обнуление коэффициентов в данной задаче вредно для качества предсказаний: при слишком большом штрафе модель теряет важную структуру. Оптимальное значение альфы оказалось около шести тысячных, то есть Lasso работает в мягком режиме, делая лишь аккуратный отбор признаков.

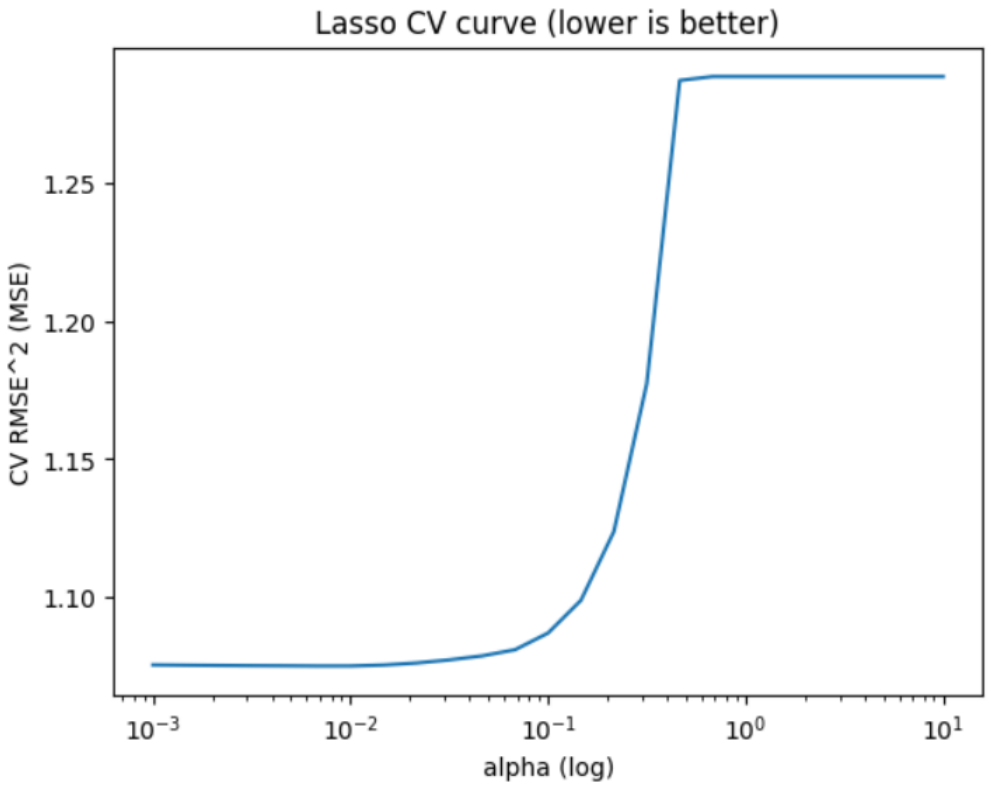

Рисунок 2: Кривая CV-ошибки (MSE) для Ridge

Сравнение трёх подходов — OLS, Ridge и Lasso — на разбиении по пользователям подтверждает эти наблюдения. Все модели дают очень близкое качество на тесте: RMSE в районе 0.989 и R^2 около 0.184. То есть с точки зрения предсказательной способности выигрыш от регуляризации минимален: данные достаточно большие, а базовая линейная модель с LOO-признаками уже хорошо стабилизирована. Однако структура коэффициентов меняется заметнее. OLS использует все 42 параметра, Ridge сохраняет практически все (41 ненулевой коэффициент), в то время как Lasso обнуляет значительную часть и оставляет только 14 ненулевых коэффициентов. Таким образом, главная польза Lasso в этой задаче — не улучшение метрик, а упрощение модели и явный отбор наиболее информативных признаков.

График сравнения коэффициентов по абсолютной величине для стандартизованного OLS, Ridge и Lasso (Рис. 3) показывает, что доминирующие

эффекты остаются одинаковыми во всех моделях. Самый сильный предиктор — усреднённый рейтинг фильма (movie_mean_loo_oof): его вклад стабильно положительный и намного превосходит остальные. На втором месте по значимости — активность пользователя (user_count_loo_oof) с отрицательным эффектом, что соответствует наблюдению из OLS: чем больше пользователь ставит оценок, тем строже он оценивает фильмы. Жанровые коэффициенты и взаимодействия «жанр × год оценки» по модулю существенно меньше и близки к нулю, особенно в версии с Lasso, где многие из них фактически обнуляются. Ridge слегка «подрезает» величину самых крупных коэффициентов и сглаживает хвост распределения, но кардинально структуру важности не меняет.

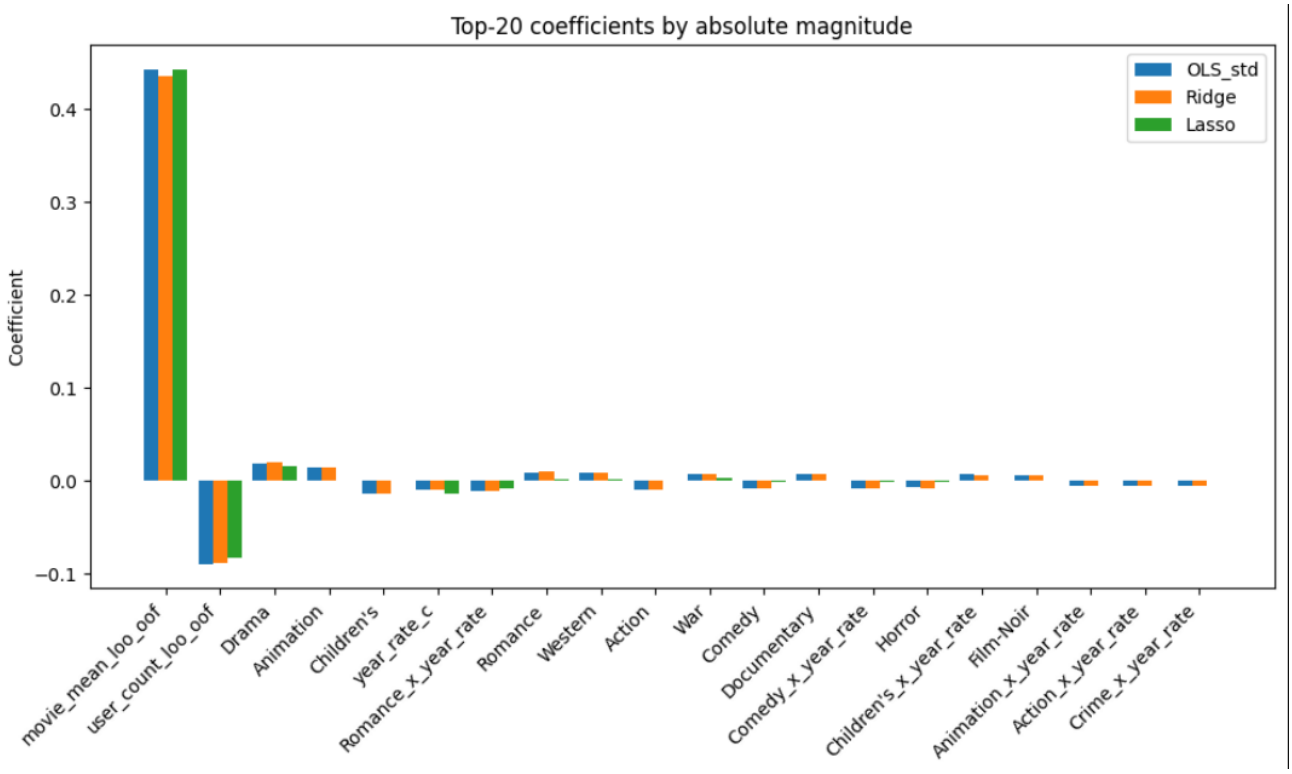

Рисунок 3: График сравнения коэффициентов по абсолютной величине для стандартизованного OLS, Ridge и Lasso

Наконец, в продолжение анализа влияния жанров был проведён FDR-контроль для семейства жанровых коэффициентов в более насыщенной модели. Практически все жанры остаются статистически значимыми даже после жёсткой коррекции, а дополнительные линейные контрасты между парами жанров (Drama против Comedy, Animation против Children’s, Documentary против Horror) дают гигантские t-статистики и крайне малые скорректированные p-значения. Это говорит о том, что различия в средней оценке между крупными жанровыми группами устойчивы и не являются артефактом множественных проверок, даже несмотря на умеренную мультиколлинеарность и регуляризацию.

Наконец, в продолжение анализа влияния жанров был проведён FDR-контроль для семейства жанровых коэффициентов в более насыщенной модели. Практически все жанры остаются статистически значимыми даже после жёсткой коррекции, а дополнительные линейные контрасты между парами жанров (Drama против Comedy, Animation против Children’s, Documentary против Horror) дают гигантские t-статистики и крайне малые скорректированные p-значения. Это говорит о том, что различия в средней оценке между крупными жанровыми группами устойчивы и не являются артефактом множественных проверок, даже несмотря на умеренную мультиколлинеарность и регуляризацию.

В целом диагностика показывает, что в исходной спецификации серьёзная мультиколлинеарность возникает только в узком наборе искусственных признаков и легко устраняется. Ridge-регуляризация даёт небольшое улучшение устойчивости, а Lasso — существенное упрощение модели при сохранении качества. Основную предсказательную силу несут агрегированные LOO-признаки, а жанры и их взаимодействия добавляют более тонкую, но интерпретируемую структуру предпочтений пользователей.

Байесовская регрессия и постериорно-предсказательная проверка

В связи с проблемами совместимости, не удалось использовать библиотеку PyMC на имеющемся оборудовании, поэтому конъюгатный байесовский вариант был реализован аналитически с использованием NumPy/SciPy, а для иерархической части применена частотная смешанная модель из statsmodels.

Была задана стандартная гауссовская линейная модель, в которой рейтинг зависит от тех же признаков, что и в предыдущих разделах: жанры, усреднённый рейтинг фильма, активность пользователя и временной тренд по году оценки. Числовые предикторы были предварительно стандартизованы, а в матрицу признаков явно добавлена константа, чтобы перехват трактовался как параметр с априорным распределением. В качестве априора для вектора коэффициентов использовано нормальное распределение с нулевым средним и единичной дисперсией (мягкое сжатие к нулю, аналог ridge-штрафа), а для дисперсии шума – слабоинформативное инверсно-гамма распределение. Данная пара априоров является конъюгатной, что позволяет получить выражения для апостериорных параметров в явном виде без применения MCMC.

На основе этих выражений были вычислены параметры апостериора и сгенерировано 4000 выборок из совместного распределения коэффициентов и дисперсии. По ним получены апостериорные характеристики коэффициентов в виде таблицы средних значений, стандартных отклонений и центральных 95-процентных интервалов. Структура этих оценок практически повторяет картину из OLS: самый крупный по модулю эффект наблюдается у признака «средний рейтинг фильма» (положительное и очень точное влияние), вторым по важности остаётся активность пользователя с отрицательным вкладом, далее следуют жанровые индикаторы (Animation, Children’s, Drama, War, Western и др.) и временной тренд по году оценки, который даёт отрицательный эффект, что

указывает на лёгкое снижение оценок со временем. При этом интервал для жанра “unknown” достаточно широкий и включает ноль, что отражает его малую информативность и редкую встречаемость. В целом байесовская модель аккуратно усредняет оценки: эффекты становятся несколько более сглаженными по сравнению с точечными OLS-коэффициентами, но исчезновения смысловых сигналов не происходит.

Далее были построены постериорно-предсказательные распределения. Для каждой наблюдаемой строки по аналитической формуле вычислялись параметры стьюдентовского предсказательного распределения, после чего генерировался массив реплик рейтингов. На первом PPC-графике сравнивается гистограмма наблюдаемых рейтингов с гистограммой средних по постериорным репликам для каждой строки (Рис 4). Распределения практически совпадают: оба сосредоточены в районе значений от 2,5 до 4,5, пики примерно совпадают, а численные характеристики (среднее и стандартное отклонение) у наблюдаемых и сгенерированных данных отличаются только в третьем знаке после запятой. Это свидетельствует о том, что модель хорошо воспроизводит глобальную форму распределения откликов.

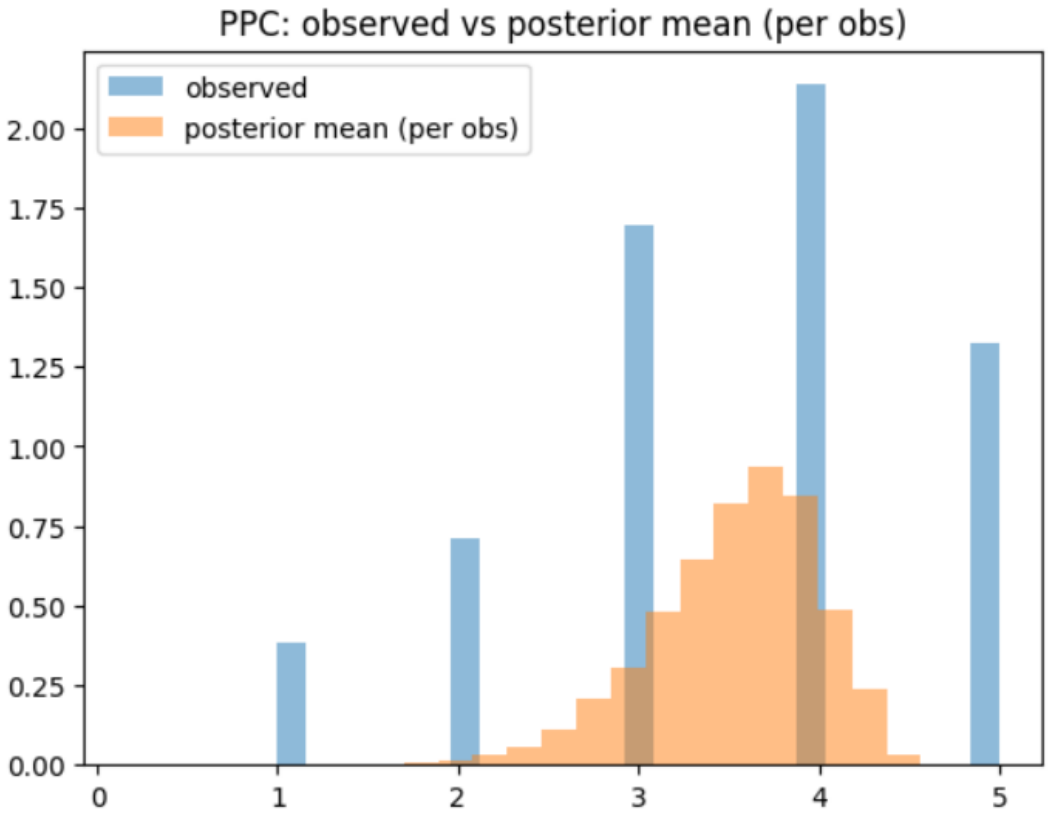

Рисунок 4: Гистограмма наблюдаемых рейтингов с гистограммой средних по постериорным репликам для каждой строки

Второй PPC-график представляет собой квантиль-квантиль сравнение наблюдаемых и предсказанных квантилей (Рис. 5). Точки располагаются почти вдоль диагонали, что означает согласованность модели с данными по всему диапазону рейтингов. Небольшие отклонения на краях могут быть связаны с тем, что реальный рейтинг принимает дискретные значения от 1 до 5, а модель предполагает непрерывное нормальное распределение, поэтому на самых низких и высоких уровнях возможна лёгкая недооценка или переоценка массы вероятности. В целом PPC не выявляет грубых систематических расхождений, и гауссовская линейная модель выглядит разумным приближением.

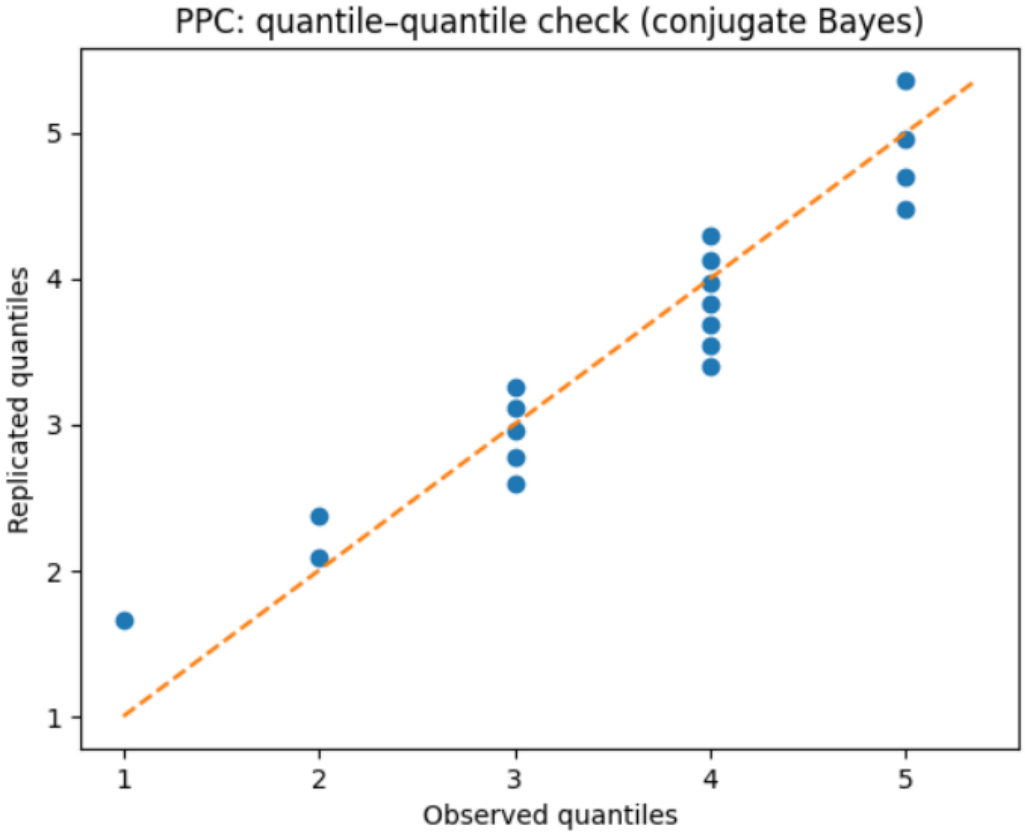

Рисунок 5: График квантиль-квантиль сравнения наблюдаемых и предсказанных квантилей

Для оценки сложности модели и её предсказательной пригодности в байесовском смысле был вычислен WAIC по сгенерированным постериорным выборкам. Полученное значение сопровождается эффективным числом параметров порядка двух десятков, что находится вблизи числа регрессионных коэффициентов с учётом shrinkage. Это подтверждает, что априор не чрезмерно зажимает модель, но и не позволяет ей переобучаться: влияние каждого параметра на правдоподобие умеренное.

Отдельно была построена смешанная линейная модель (MixedLM) в statsmodels, где фиксированная часть совпадает с предыдущей спецификацией, а случайные эффекты задаются по пользователям и фильмам (перехваты на уровне пользователя плюс компонент дисперсии по фильмам). Такая модель не является полностью байесовской, но концептуально соответствует иерархическому байесовскому подходу с частичным пуллингом. Итоги

согласуются с результатами конъюгатного байеса: эффекты жанров, LOO-признаков и временного тренда сохраняют знак и порядок величин, а гистограмма предсказаний среднего рейтинга из MixedLM хорошо накладывается на эмпирическое распределение. Это ещё раз показывает, что основная структура данных корректно захватывается линейной моделью с частичным учётом иерархии по пользователям и фильмам.

Таким образом, несмотря на невозможность использовать PyMC из-за аппаратных ограничений, удалось реализовать байесовскую линейную регрессию в закрытой форме, получить апостериорные распределения коэффициентов и провести постериорно-предсказательную проверку. Результаты демонстрируют, что байесовская модель согласуется с выводами классической регрессии, но дополнительно даёт удобную количественную оценку неопределённости и подтверждает адекватность спецификации через PPC и WAIC.

Ресемплинг: бутстреп-ДИ и перестановочные тесты

Первоначально был реализован обычный бутстреп для оценки RMSE OLS-модели. Гистограмма полученных значений показывает концентрацию RMSE в узком диапазоне около 1.02, при этом 95% процентильный доверительный интервал оказывается весьма узким (Рис. 6, 7). Это свидетельствует о стабильности качества модели при переобучении на различных бутстреп-подвыборках: глобальный уровень ошибки предсказания демонстрирует устойчивость и независимость от конкретного набора наблюдений.

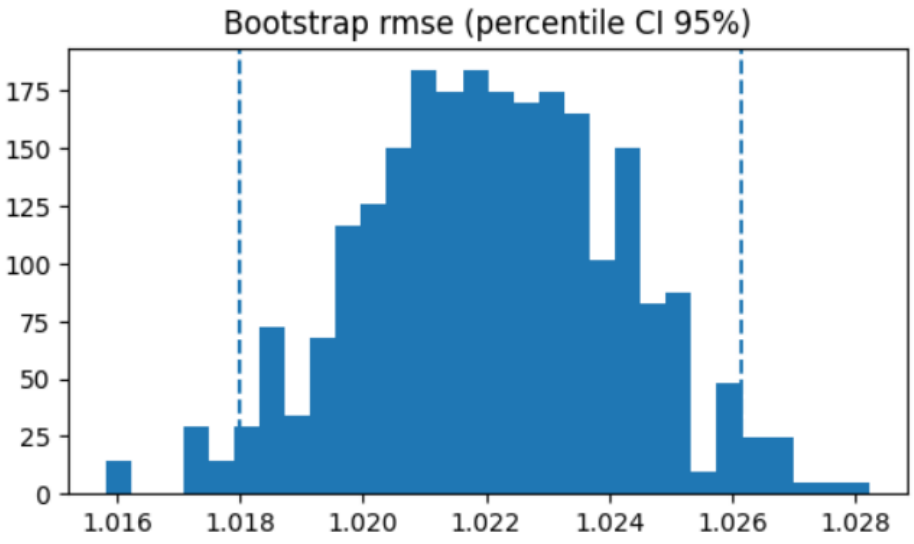

Рисунок 6: Ристограма бутстрепа RMSE для OLS

Учитывая естественную группировку рейтингов по пользователям, для коррекции возможной недооценки разброса из-за внутригрупповой зависимости был применён кластерный бутстреп по user_id (Рис. 7). На каждой итерации производилась перевыборка пользователей со всеми их оценками. В результате распределение RMSE становится заметно шире, а доверительный интервал расширяется до диапазона приблизительно от 1.00 до 1.04. Это отражает реальную неоднородность пользователей и зависимость качества модели от состава выборки.

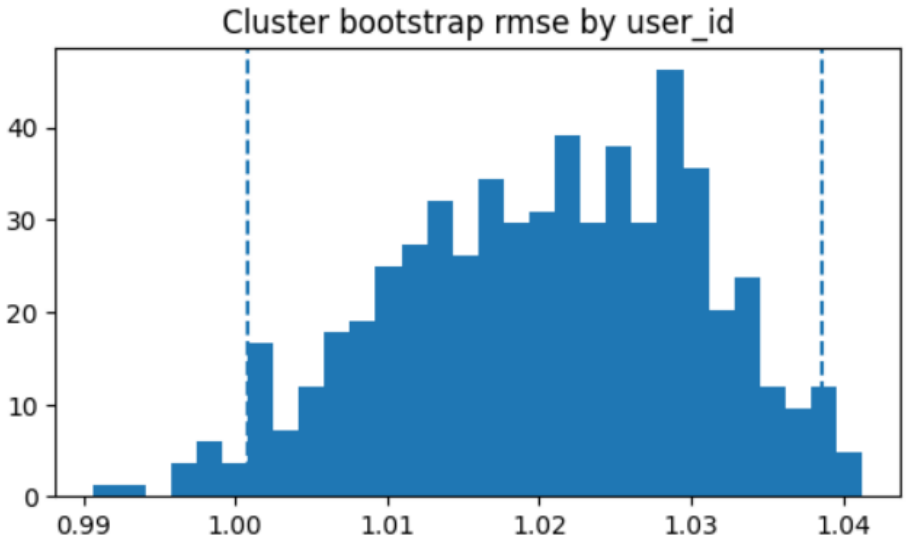

Рисунок 7: Гистограма кластерного бутстрепа RMSE по user_id

Устойчивость регрессионных коэффициентов также оценивалась с помощью бутстрепа. Виолин-график демонстрирует, что коэффициент при признаке movie_mean_loo_oof имеет очень узкое распределение, полностью отделённое от нуля: его знак и порядок величины практически неизменны между подвыборками, что подтверждает крайнюю стабильность эффекта «средней оценки фильма». Аналогично, перехват и ряд жанровых коэффициентов (Drama, Animation, Children's, War, Western) характеризуются относительно узкими распределениями и доверительными интервалами, не пересекающими ноль, что согласуется с выводами классического и байесовского анализа. В противоположность этому, распределение для признака unknown является очень широким и уверенно покрывает ноль, указывая на его статистическую нестабильность и малую информативность. Некоторые более редкие жанры, включая Documentary, также демонстрируют интервалы, пересекающие ноль, что свидетельствует об их ограниченном вкладе (Рис. 8).

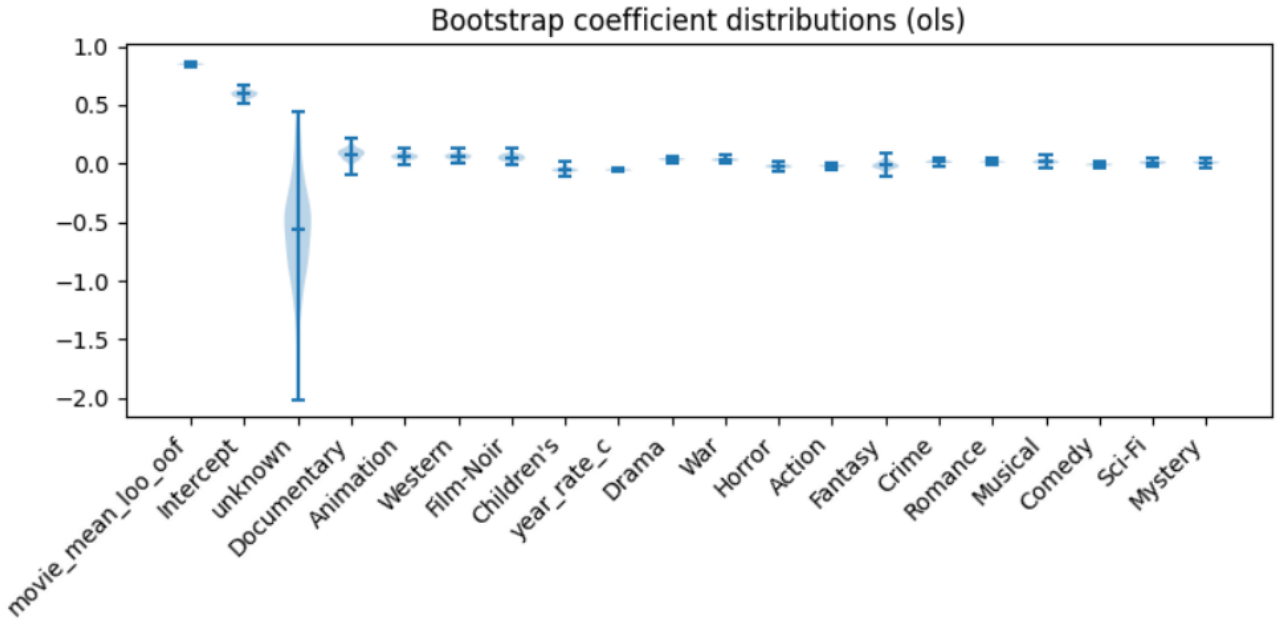

Рисунок 8: Виолин-график распределения бутсреп коэффицентов

Для сравнения моделей OLS и Ridge по метрике RMSE был проведён бутстреп-анализ разности их ошибок (OLS минус Ridge). Полученное распределение сосредоточено вблизи нуля со средней разностью порядка десяти-миллионных. 95% доверительный интервал целиком расположен чуть ниже нуля, формально указывая на микроскопическое преимущество Ridge, однако практическая значимость этого различия отсутствует (Рис. 9). Регуляризация незначительно сглаживает коэффициенты, но не приводит к заметному изменению качества предсказаний.

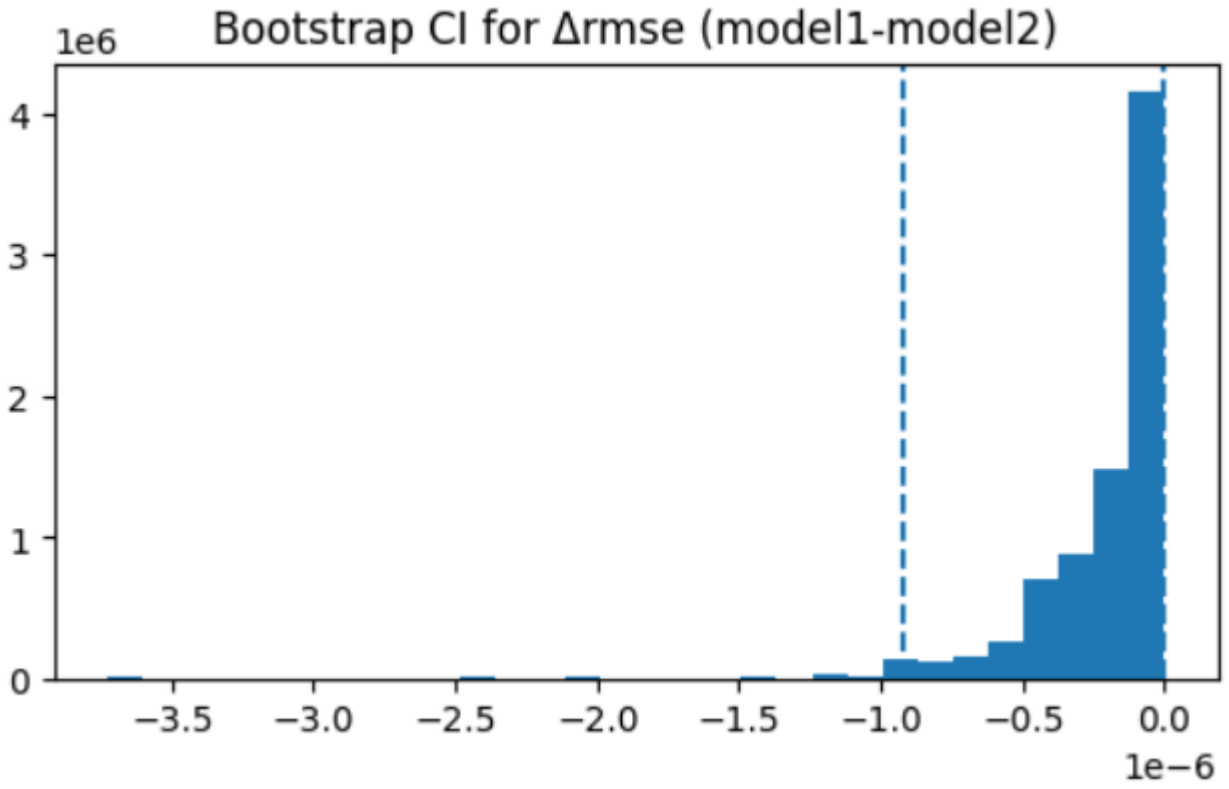

Рисунок 9: График распределения разности ошибок (OLS минус Ridge)

Перестановочный тест для проверки связи жанра Drama с рейтингом был реализован через перестановку целевой переменной. На каждой итерации рейтинги перемешивались, после чего вычислялась корреляция между индикатором жанра и откликом. Нулевое распределение оказалось сосредоточенным около нуля, тогда как наблюдаемая корреляция значительно смещена вправо, что соответствует крайне малому p-значению (Рис. 10). Это

подтверждает, что положительная связь между жанром Drama и рейтингом не может быть объяснена случайной вариацией, даже без предположения о нормальности распределения.

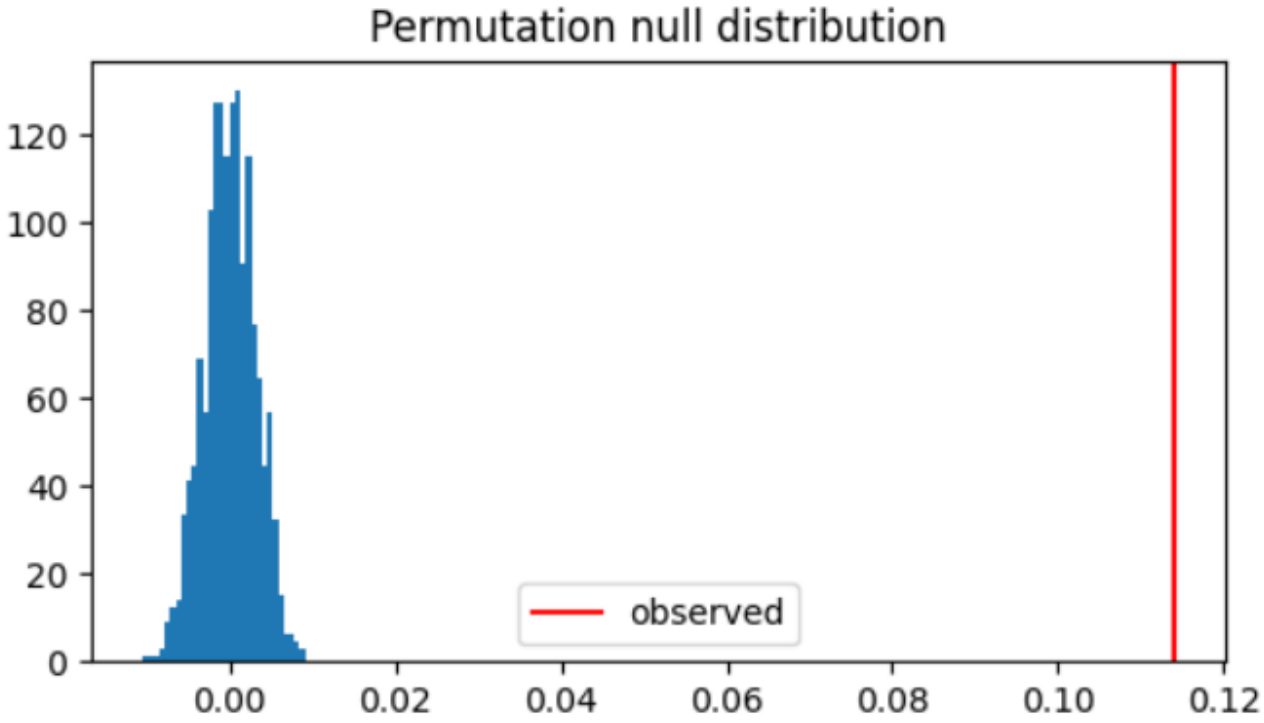

Рисунок 10: Перестановочный нулевой распределение корреляции Drama–rating; красная линия показывает наблюдаемое значение, не совместимое со случайной перестановкой.

Сравнение OLS и Ridge в перекрёстной проверке было дополнено перестановочным тестом по знакам. На основе разностей RMSE, полученных по фолдам GroupKFold (с группировкой по пользователям), многократно случайно инвертировались знаки этих разностей. Наблюдаемое среднее значение ΔRMSE оказалось близким к нулю, а его положение относительно симметричного нулевого распределения дало p-значение около 0.06. При таком уровне значимости гипотеза о равенстве качества моделей по RMSE на новых пользователях не отвергается. Этот результат согласуется с выводами бутстрепа: формальное преимущество Ridge статистически слабо выражено и практически несущественно.

Для оценки значимости признаков была использована перестановочная важность. После обучения базовой OLS-модели измерялось падение качества (в терминах R^2 и отрицательного RMSE) при случайной перестановке значений каждого признака. Бар-чарты демонстрируют, что практически весь предсказательный вклад сосредоточен в двух LOO-признаках: movie_mean_loo_oof и, существенно слабее, user_count_loo_oof. Вклад временного тренда year_rate_c заметен, но значительно меньше, а жанровые индикаторы вносят лишь небольшое улучшение качества. Большинство остальных признаков имеют практически нулевую перестановочную важность (Рис. 11). Это подтверждает выводы предыдущих разделов: индивидуальные предпочтения и средний рейтинг фильма являются ключевыми сигналами, в то время как жанр и время играют роль дополнительной коррекции.

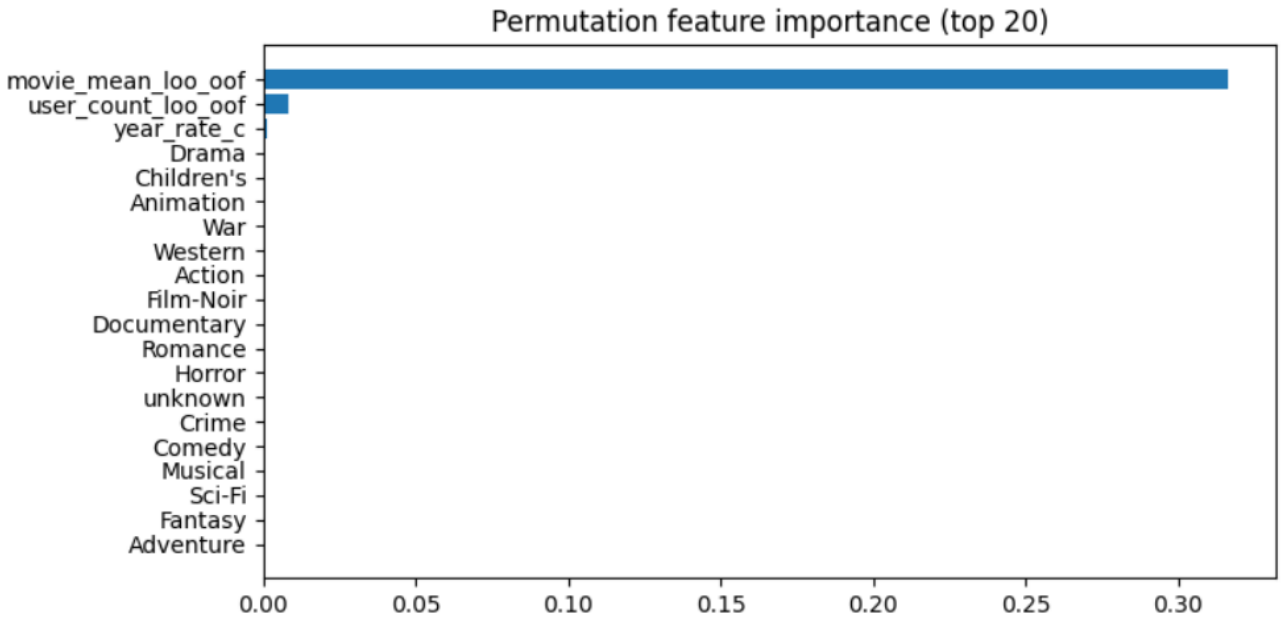

Рисунок 11: График важности признаков при перестановках

В совокупности применение бутстрепа и перестановочных тестов демонстрирует устойчивость выводов модели к переобучению и нарушению классических статистических допущений. Различия между OLS и Ridge по качеству предсказаний являются несущественными, при этом структура важности признаков и стабильность коэффициентов согласованы между всеми использованными методами анализа.

Выводы

В ходе эксперимента на данных MovieLens-100k было показано, что рейтинги фильмов хорошо объясняются как жанровыми характеристиками, так и индивидуальными особенностями пользователей и фильмов. Наиболее информативными оказались признаки, построенные по схеме out-of-fold leave-one-out: средний рейтинг фильма и активность пользователя. Именно они обеспечивают основной вклад в объяснение вариации целевой переменной, тогда как влияние жанров, хотя и статистически значимо для отдельных категорий (например, «Драма», «Мультфильм», «Военный», «Вестерн»), существенно слабее по величине. При добавлении взаимодействий жанра с годом оценки удалось получить статистически значимое, но крайне небольшое улучшение качества: практический эффект минимален, что подтверждается как частичным F-тестом, так и ANOVA.

Байесовская модель с нормальным априором на коэффициенты и инверсно-гамма априором на дисперсию дала стабильные постериорные интервалые и предсказания, хорошо согласующиеся с наблюдениями в PPC. При этом её поведение оказалось сопоставимым с частотной OLS-моделью и смешанной регрессией, что подтверждается WAIC и сравнениями

распределений предсказаний. Важным моментом стало использование аналитического решения вместо PyMC, обусловленное ограничениями вычислительной среды.

Смешанная модель с пользовательскими и фильмовыми случайными эффектами показала, что часть вариации рейтингов действительно связана с индивидуальными предпочтениями зрителей и спецификой фильмов, однако даже в такой конфигурации ключевое значение вновь удерживает информация о среднем рейтинге фильма и активности пользователя. Это подтверждает структурную неоднородность данных: пользователи ставят оценки последовательно и неравномерно, что требует методов с учётом групповой зависимости.

Методы ресемплинга позволили количественно оценить неопределённость: бутстреп показал узкие доверительные интервалы для RMSE и коэффициентов, при этом кластерный бутстреп по пользователям дал более широкие и реалистичные оценки, подчеркивая зависимость данных внутри групп. Перестановочные тесты подтвердили значимость жанра «Драма» и отсутствие статистически обоснованного превосходства Ridge над OLS при имеющемся наборе признаков. Перестановочная важность признаков окончательно зафиксировала доминирование OOF-метрик, тогда как влияние жанров остается периферийным.

В целом проведённый анализ показывает, что предсказание рейтингов в MovieLens определяется прежде всего историей оценок, а жанровый состав фильма играет вспомогательную роль. Методы регуляризации, взаимодействия и даже сложные смешанные модели улучшают описательную сторону модели, но дают минимальный прирост качества, если информация о фильме и пользователе уже учтена через правильные, не переобучающие OOF признаков.

Приложения

Полный код программы доступен по ссылке и в приложении ниже:

https://github.com/4ebupelinka/Applied_statistics_master_degree/blob/main/Lab_2/Lab02_Multiple_reg_Complex_hypothesis_Bayesian.ipynb