Тема: Прогнозирование финансового успеха фильма: влияния жанров, студий и сезона на кассовые сборы на примере 5000 фильмов.

Группа: Mi-213  
Выполнил: Гарнов Александр

Аннотация

В данном исследовании изучаются факторы, влияющие на финансовый успех фильмов. Основное внимание уделяется таким аспектам, как сроки выпуска, репутация студии-производителя и жанровые предпочтения. Методология включает создание регрессионной модели для анализа взаимосвязи между различными переменными и кассовыми сборами. Результаты исследования подчеркивают сложность прогнозирования кассового успеха фильмов и указывают на необходимость учета множества факторов для принятия обоснованных стратегических решений в киноиндустрии.

# Введение

Прогнозирование финансового успеха фильма имеет решающее значение для киноиндустрии и влечет за собой значительные последствия для кинематографистов, продюсеров, дистрибьюторов и инвесторов. Сложность и затраты, связанные с производством фильмов, включая предпродакшен, постпродакшен и маркетинг, подчеркивают важность понимания и предвидения потенциального успеха фильма. Учитывая значительные финансовые вложения, студии и производственные компании должны оценить потенциальную отдачу от инвестиций, чтобы принимать обоснованные решения о поддержке проекта, распределении бюджета и стратегии распространения.

## Цель

Данное исследование направлено на изучение влияния переменных сопутствующие любому фильму - сезона выхода, производственной студии, жанра фильма и другие - на кассовые сборы. Изучая взаимодействие между этими факторами, исследование стремится дать представление о принятии стратегических решений в киноиндустрии.

## Тема

Финансовый успех фильма тесно связан с временем его выхода, репутацией и стратегиями студии-производителя, а также предпочтениями аудитории в отношении жанра. Изучая эту тематику, данное исследование факторы влияющие на успех кино. Благодаря анализу и моделированию, исследование намерено внести практический вклад для заинтересованных сторон, ориентирующихся в проблемах и возможностях киноиндустрии.

# Методология

## Набор данных

### Источник данных

Набор данных для исследования был скачан с сайта kaggle. Он был создан методом веб-скрапинга. Данные были извлечены из официальных сайтов дистрибьюторов и кинокомпаний таких как: Disney, Warner Bros., Paramount, 20 Century Fox, Lionsgate и другие.

### Описание набора данных

Набор данных состоит из 5000 записей о фильмах, в котором существует 8 переменных:

* budget: В этом столбце указывается бюджет, выделенный на производство фильма. Он отражает финансовые ресурсы, вложенные в создание фильма.
* genres: в этом столбце указывается жанр, к которым относится фильм. Жанры классифицируют фильмы по их тематическому содержанию, например, боевики, приключения и т. д.
* production\_companies: В этом столбце указаны производственная компания, ответственная за создание фильма. Это дает представление об организациях, которые снимают фильмы.
* production\_countries: В этом столбце указаны страны, в которых были сняты фильмы. Это дает информацию о географическом происхождении фильмов.
* revenue: В этом столбце представлены кассовые сборы каждого фильма. Он служит целевой переменной для анализа, отражающей финансовый успех фильмов.
* runtime: В этом столбце указываются продолжительность фильма в минутах.
* spoken\_languages: В этом столбце указываются языки, на которых говорят в фильмах, что позволяет проанализировать лингвистические аспекты фильмов.
* release\_month: В этом столбце указывается месяц, в котором был выпущен фильм. Он обеспечивает временной аспект анализа, позволяя изучить влияние сезонности на кассовые сборы.

## Предварительная обработка данных

Главные задачи, которые необходимо было реализовать в предварительной обработке данных это: очистка и извлечения данных из JSON формата, обработка отсутствующих значений, приведение данных к общему формату, идентификация и удаление выбросов. Предварительная обработка данных необходима для того, чтобы обеспечить точность, надежность и согласованность данных, что, в свою очередь, способствует корректным и надежным результатам в дальнейшем анализе.

Все эти и последующие изменения с набором данных были сделаны на языке R и Python. Для очистки и извлечения данных из JSON формата я использовал 2 библиотеки jsonlite и stringr. jsonlite - предлагает простые и гибкие инструменты для работы с JSON в R[[1]](#footnote-0), а stringr позволяет манипулировать отдельными символами и словами в строках[[2]](#footnote-1). Используя эти библиотека я прошёлся по всем переменным в наборе данных и очистил их от всех ненужных тэгов, символов и слов.

В данном наборе данных пропущенных значений оказалось мизерное количество в сравнении с общим количеством записей. Почти во всех случаях запись отсутствовала в колонках genres и production countries. Было решено удалить записи, где значение в какой-либо колонке было пустым. В результате было удаленно 74 записи.

### Категориальные переменные.

Важнейшим этапом работы стало приведение всех переменных к единому формату. Главная сложность заключалась в том, что большинство переменных (5 из 8) в наборе данных являются категориальными. А количество разных значений в одной колонке доходит до 50. Однако после подсчета уникальных значений в каждой категориальной переменной оказалось, что большее количество значений встречаются крайне редко. В результате, во всех категориальных переменных, всем редко встречающимся переменным было присвоено значение “other”. Данная процедура была проведена в genres, production\_companies, production\_countries и spoken\_languages.

После объединения всех незначительных категорий в одну в каждой переменной необходимо все категориальные переменные превратить в индикаторные переменные (indicator variables). Для это была использована библиотека pandas[[3]](#footnote-2). Pandas - это библиотека программирования на языке Python, предназначенная для обработки и анализа данных. Из данной библиотеки была взята функция pandas.get\_dummies()[[4]](#footnote-3). Используя эту функцию каждая переменная преобразуется в столько переменных 0/1, сколько существует различных значений. В результате 5 категориальных переменных были поделены на 45 индикаторных переменных.

### Числовые переменные.

Главная задача с числовыми переменными это проверить их распределение и в случае необходимости привести распределение в норму. Во всех моих числовых переменных наблюдается длинный хвост слева. Чтобы выйти из этой ситуации я провел логарифмирование всех числовых переменных:“budget”, “runtime”, “revenue”. А также удалил хвосты у этих переменных. Теперь распределение числовых переменных в норме [Рис. 1]

## EDA

EDA - это подход к анализу и обобщению наборов данных для понимания их основных характеристик, часто с помощью статистических графиков и других методов визуализации данных.[[5]](#footnote-4) С помощью этого подхода мы сможем изучит что из себя представляют переменные из набора данных и как они влияют на друг друга.

В данной работе EDA было выполнено с помощью библиотеки ggplot[[6]](#footnote-5). Ggplot - это библиотека в языке R для декларативного создания графиков на основе The Grammar of Graphics. Для проведения EDA из этой библиотеки были использованы гистограммы, линейные графики и скаттерплоты (scatterplot’s).

## Выбор переменных

Отбор признаков - важный шаг в машинном обучении, который заключается в выборе подмножества релевантных признаков (переменных или атрибутов) из исходного набора признаков для повышения эффективности модели и снижения риска перебора.[[7]](#footnote-6) Правильный выбор признаков может привести к созданию более эффективных моделей, сокращению времени обучения и улучшению обобщения.

Ключевым значением в данном исследовании являются кассовые сборы (“revenue”). “revenue” - зависимая переменная (Y), значение который будет предсказываться. Независимыми переменными (X) будут все остальные значения. Однако, если числовые переменные (“budget”, “runtime”) остались без изменений, то категориальные переменные (“genres”, “production\_companies”, “production\_countries”, “spoken\_languages”, “release\_month”) были разделены на 45 индикаторных переменных. В сумме 47 независимых переменных были подвержены 3 способами отбора:

### Ridge

Отбор переменных в контексте Ridge регрессии связан с использованием самой регрессии для автоматического отбора переменных. Ridge регрессия вводит регуляризацию, которая помогает предотвратить переобучение модели и улучшить ее обобщающую способность. Регуляризация Ridge добавляет штраф к сумме квадратов коэффициентов (кроме свободного члена) к целевой функции, что способствует уменьшению влияния избыточных переменных. [[8]](#footnote-7)

### Lasso

Отбор переменных с использованием Lasso (Least Absolute Shrinkage and Selection Operator) регрессии является эффективным методом для создания более простых и интерпретируемых моделей, обнуляя (или близко к нулю) веса некоторых признаков. В отличие от Ridge регрессии, Lasso использует L1-регуляризацию, что способствует созданию разреженных моделей и, таким образом, выполняет более строгий отбор переменных.[[9]](#footnote-8)

### RandomForest

Random Forest предоставляет встроенные механизмы для оценки важности переменных, что делает его эффективным инструментом отбора признаков. В Random Forest, важность переменной измеряется на основе ее вклада в уменьшение неопределенности (например, прироста Gini в случае классификации или уменьшения среднеквадратичной ошибки в случае регрессии).[[10]](#footnote-9)

### Отобранные Переменные

В результате проведения отбора каждый метод выбрал свою переменную. Однако собрав все результаты и создав график [Рис. 13] можно увидеть какие переменные были выбраны чаще всего. “budget”, “runtime”, “genres\_Adventure”, “genres\_Animation”, “genres\_Horror”, “production\_countries\_UK”, “production\_countries\_USA”, “release\_month\_12”, “release\_month\_5”, “release\_month\_6”, “release\_month\_7” - были отобраны 3 способами отбора.

## Регрессионная модель

Регрессионная модель - это статистический метод, используемый в машинном обучении и статистическом моделировании для изучения взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными (предикторами или признаками).[[11]](#footnote-10) Основная цель регрессионной модели - понять и количественно оценить связь между переменными и сделать прогнозы или выводы.

В исследовании применяется метод линейной регрессии так как зависимая переменная является числовой (“revenue”). В данном виде регрессии взаимосвязь моделируется в виде линейного уравнения, и цель состоит в том, чтобы найти наиболее подходящий результат, который минимизирует разницу между прогнозируемыми и фактическими значениями зависимой переменной.

Для реализации линейной регрессии существует несколько способов. В исследовании используется RandomForest из библиотеки языка Python sklearn.[[12]](#footnote-11) RandomForest - это алгоритм машинного обучения, относящийся к категории ансамблевого обучения. Он используется для решения задач классификации и регрессии. В процессе обучения алгоритм строит несколько деревьев решений и объединяет их прогнозы для получения более точного и стабильного результата.

После достижения результата будет использоваться значения Шапли. В контексте анализа данных, в частности при интерпретации моделей машинного обучения, значения Шэпли позволяют справедливо распределить вклад каждого признака в предсказание конкретного случая.[[13]](#footnote-12) И после создания графика мы точно сможем определить какая переменная влияет больше всего на предсказываемые значения.

# Результаты

## Анализ графиков по результатам EDA

Все переменные в наборе данных были разделены на числовые и категориальные. С числовыми переменными, которых оказалось всего 3 (“budget”, “runtime”, “revenue”), была создана матрица [Рис. 2]. На графике видно, что кассовые сборы и бюджет имеют сильную зависимость, в то же самое время длительность фильма незначительно влияет на сборы.

Переходя уже к анализу гистограмм и других графиков, можно сделать ряд выводов. В среднем продолжительность фильма составляет примерно два часа [Рис. 3], Соединенные Штаты Америки выступают в роли главного производителя кинематографических шедевров [Рис. 4], а главным языком в кинематографе является английский язык [Рис. 11]. На вершине популярности расположились такие жанры, как Драма, Комедия и Боевик [Рис. 5], при этом стоит отметить, что именно эти категории фильмов чаще всего способны оправдать вложенные в них финансовые ресурсы.

Если мы разделим все фильмы этого набора данных на 3 категории по размеру бюджета, то выясним, что подавляющее количество фильмов имеют маленький бюджет. [Рис. 10] И если сравнить количество фильмов, которые смогли окупить себя или хотя бы вернуть затраты, то увидим, что подавляющее число фильмов проваливаются в прокате. [Рис. 9] Однако, интересным является наблюдение о том, что среди фильмов, которые смогли окупить себя, больше всего фильмов с большим бюджетом [Рис. 6]. А также то, что успешные фильмы, которые окупили свои расходы, чаще всего представлены крупными кинокомпаниями [Рис. 7]. Это свидетельствует о том, что доступ к ресурсам, возможности маркетинга и распространения могут иметь значительное влияние на финансовый успех проекта в кинематографической сфере. А интерес людей к небольшим, авторским фильмам слишком мал для того чтобы окупить даже их небольшой бюджет.

Не менее важным фактором оказывается сезон выхода фильма. Анализ показывает, что именно в июне и декабре фильмы приносят наибольшие кассовые сборы [Рис. 8]. Вероятно, это связано с сезонными трендами, праздничной атмосферой или иными факторами, способствующими привлечению внимания зрителей.

## Результаты создания регрессионной модели.

По описанной выше методологии была создана регрессионная модель. Производительность модели на тестовом наборе умеренно удовлетворительна. Коэффициент R2, равный 0,40095, указывает на то, что модель учитывает 40,1 % изменчивости тестовых данных. Хотя это говорит о приемлемом уровне объясняющей способности, все еще остается значительная часть необъясненной дисперсии.

Остаточная сумма квадратов (RSS) 1797,22577 и средняя квадратичная ошибка (MSE) 2,89875 показывают, что существуют значительные ошибки между предсказанными и фактическими значениями в тестовом наборе. Так же был создан график показывающий распределение ошибок [Рис 14]. Среднеквадратичная ошибка (RMSE), равная 1,70257, представляет собой среднюю оценку этих ошибок и говорит о том, что в среднем прогнозы модели отклоняются от фактических значений примерно на 1,70 единицы.

Если взглянуть на Scatterplot, который был создан для сравнения предсказанных и фактических данных [Рис. 16], где показывается корреляция между кассовыми сборами и бюджетом, то можно увидеть, что прекрадсказаные значения в большинстве своем повторяют силуэт фактических значений. Однако на периферии графика видно много упущений и то как предсказанные значения не могут добраться до фактических.

В дополнение был создан график значений Шапли [Рис. 17]. По его результатам можно сказать, что самым важным параметром, который больше всего влияет на кассовые сборы является бюджет. Далее идет длительность фильма и Соединенные Штаты как страна производства. На графике видно, что некоторые месяца влияют на кассу больше чем жанры фильмов, а анимация занимает последнее место.  
   
 12 Моделей по всем месяцам в году.

По результатам EDA было выявлено, что существенное влияние на кассовые сборы оказывает месяц выхода фильма. Модель, описанная выше, охватывает все месяца, жанры и студии, а также страдает из-за маленького R2 и ошибок. Для увеличения R2 и уменьшения ошибок были созданы ещё 12 новых моделей, разделённые по месяцам выхода фильма.

В результате разделение набора данных и создания новых моделей по месяц в 11 из 12 моделей значение R2 стало меньше и варьируется в районе 0.20 - 0.30 [Рис. 18].

Однако модель для мая отличается от стальных.  
 Если сравнить “майскую” модель с первой моделью, то видно, что “майская” модель лучше по некоторым показателям [Рис. 19]. У нее выше показатель R2 (0.51) , что свидетельствует о лучшем соответствии данным, ниже остаточная сумма квадратов RSS (160.55). Однако, средняя квадратичная ошибка (MSE) равно 3.34 что больше MSE, которые была получена при создании первой модели (2.89).

В целом, несмотря на более высокий MSE, "майская" модель продемонстрировала улучшение в R2 и RSS, что может свидетельствовать о более точном соответствии данным в сравнении с первой моделью.

Изучив структуру и распределение в наборе данных за май не было найдено никаких отличий от остальных месяцев. Скорее всего дело в разделении на обучающий и тестовый набор. Из-за разделения изначального набора данных на 12 частей, на каждый месяц приходилось не более 400 записей. Уже при создании моделей данные неравномерно делились на обучающие и тестовые. И если модель очень хорошо понимала связь между данными на обучающем наборе данных, так как R2 во всех 12 моделях в обучающих наборах колеблется около 0.88 [Рис. 20], то на тестовых данных модель теряется и R2 становится заметно меньше. Видимо в “майской” модели разделение данных произошло удачней, модель определила взаимосвязей и дала более точный результат.

## Выводы

Данное исследование позволяет сделать несколько важных выводов о факторах, влияющих на кассовые сборы фильмов. В частности, выявлена сильная зависимость между бюджетом и кассовыми сборами, подчеркивающая важность финансовых ресурсов для успеха проекта. Жанровые предпочтения зрителей также отражены в данных, где драма, комедия и боевик пользуются наибольшей популярностью. Отмечается, что эти жанры чаще всего оправдывают вложенные в них финансовые ресурсы.

Факторы, такие как размер бюджета, кинокомпания-производитель и сезон выхода, существенно влияют на успешность фильма. Фильмы с крупными бюджетами, выпущенные крупными студиями и выходящие в определенные месяцы, имеют больше шансов на финансовый успех.

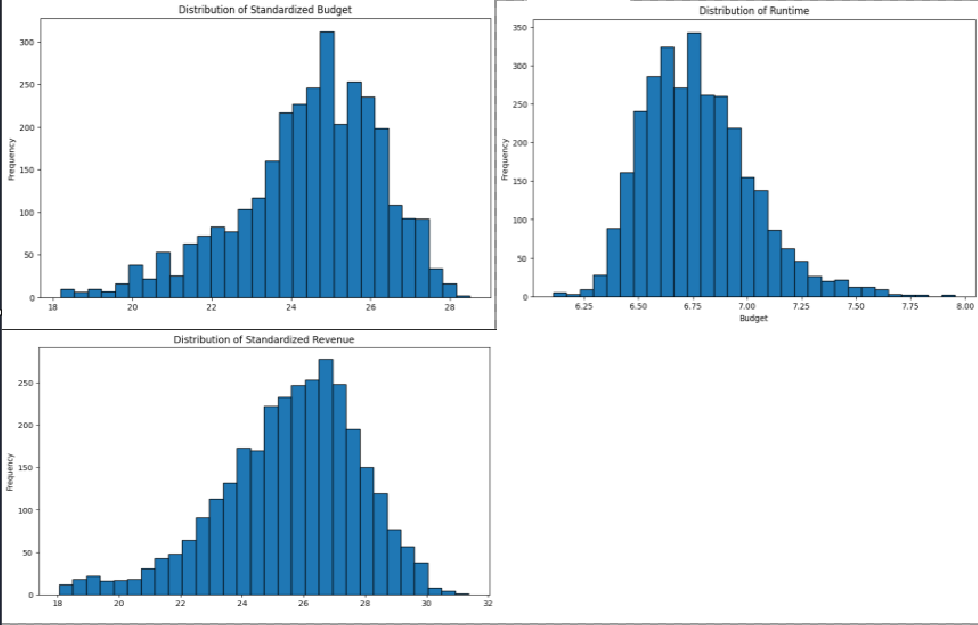
Созданная регрессионная модель, несмотря на умеренно удовлетворительную производительность на тестовом наборе данных, подчеркивает сложность предсказания кассовых сборов, поскольку значительная часть изменчивости остается необъясненной. А важность бюджета, выхода фильма в определенный месяц и жанр фильма как фактора успеха подтверждается графиком значений Шапли.

Дополнительные модели, разделенные по месяцам, не выявили ничего нового. "Майская" модель, несмотря на некоторые увеличения в R2 и улучшения в RSS, демонстрирует ухудшения в виде более высокой MSE. Это может свидетельствовать о том, что даже с улучшением в общем объяснении модели, точность прогнозов все еще ограничена.

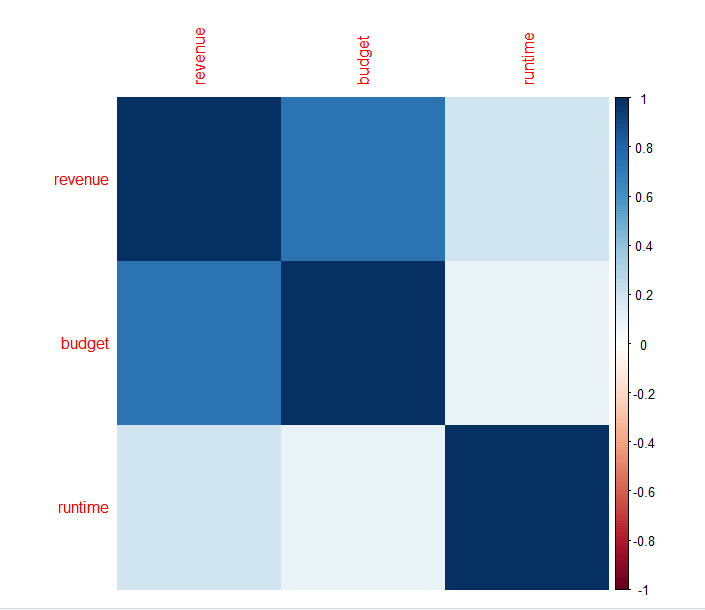
Подобный результат связан с малым количеством записей и ограниченностью изучаемого набора данных. Для улучшения результатов и более точного понимания факторов влияющих на кассовые сборы необходим набор данных в разы крупнее, который будет разделен не только по месяцам выхода фильма, но и по жанру и студиям производителям. Также стоит учесть в разы больше переменных, которые не изучались в этом исследовании.

Таким образом, успешность кинопроекта зависит от множества факторов, включая бюджет, кинокомпанию, выбранный месяц релиза и жанр. Однако, предсказание кассовых сборов остается сложной задачей, требующей учета множества переменных и факторов, а также может подразумевать неоднозначные результаты в зависимости от конкретных условий.

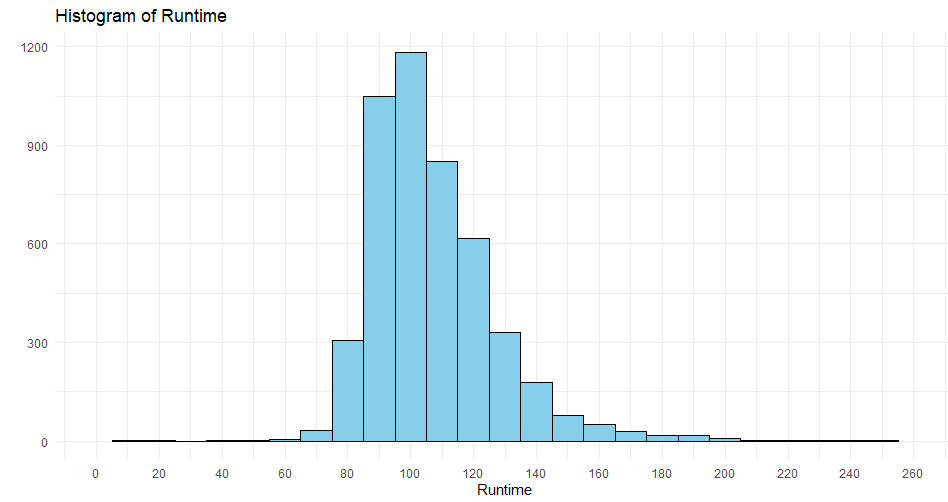
**Приложения**

****

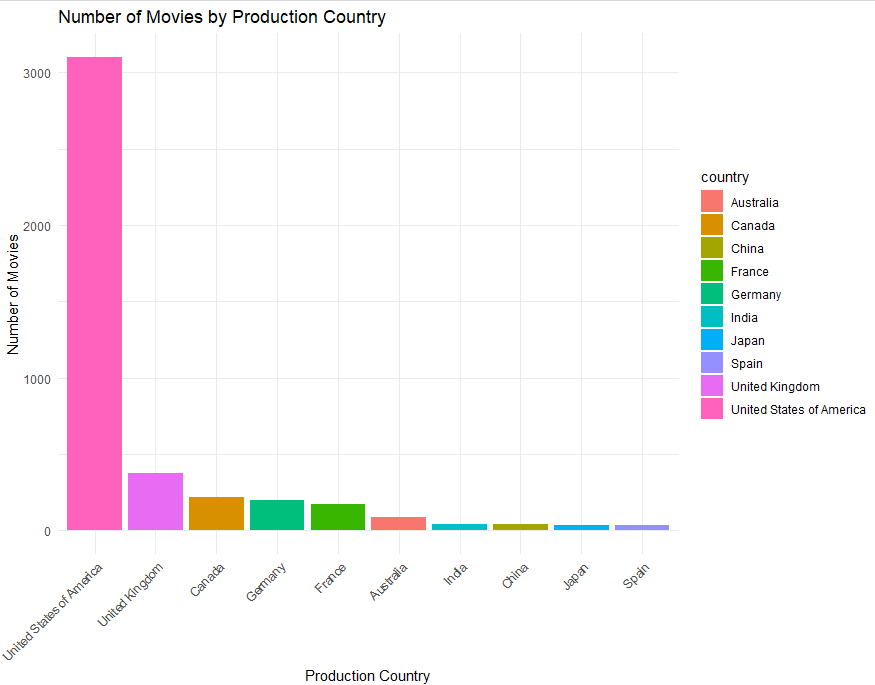
**Рисунок 1: Распределение budget, runtime, revenue после логарифмирования.**

****

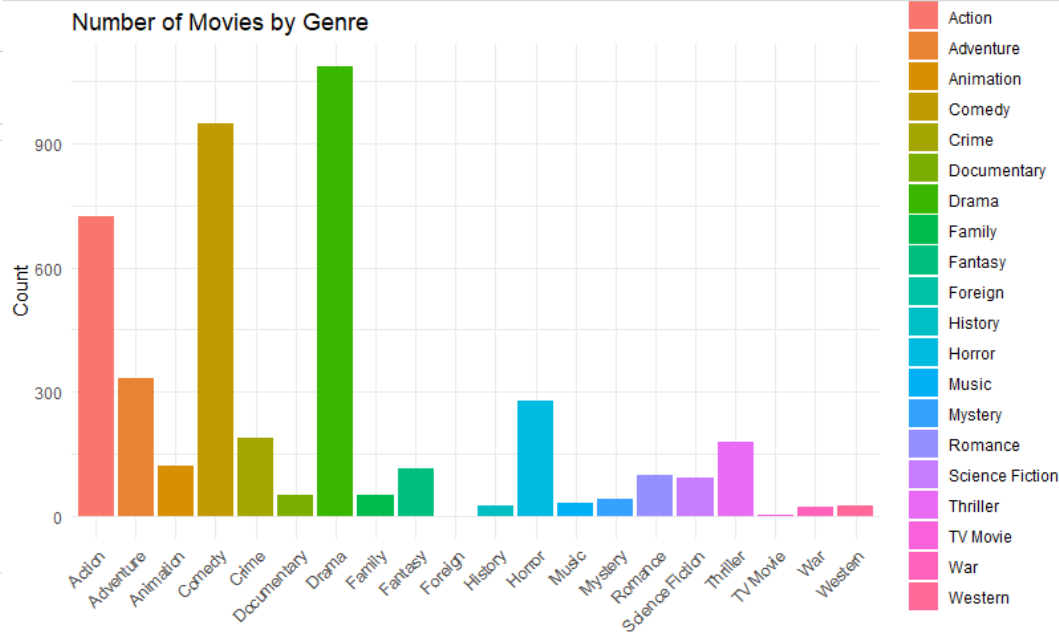
**Рисунок 2: Матрица Корреляции.**

****

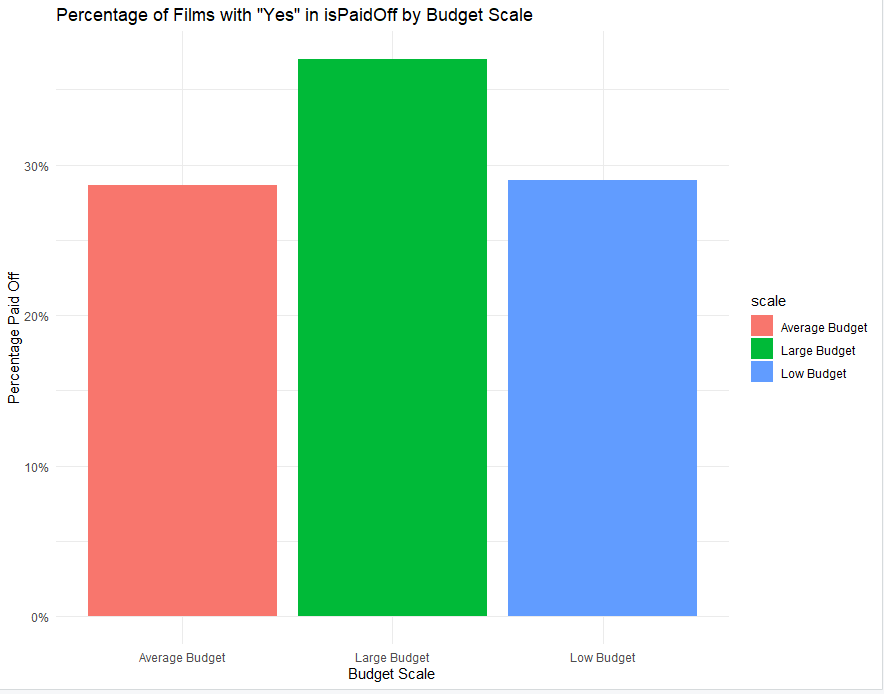
**Рисунок 3: График распределения runtime.**

****

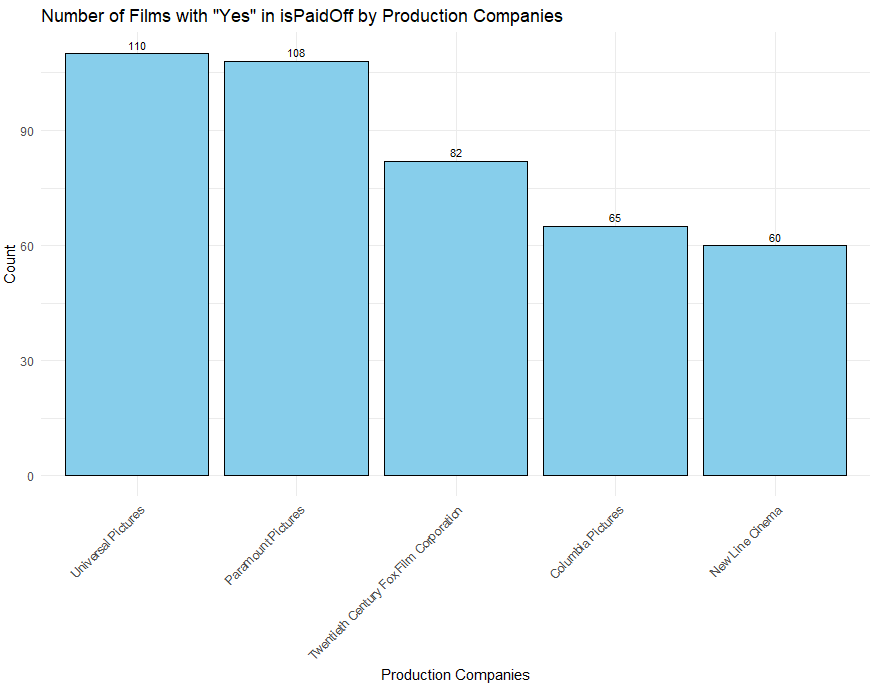
**Рисунок 4: График главных стран-производителей кино.**

****

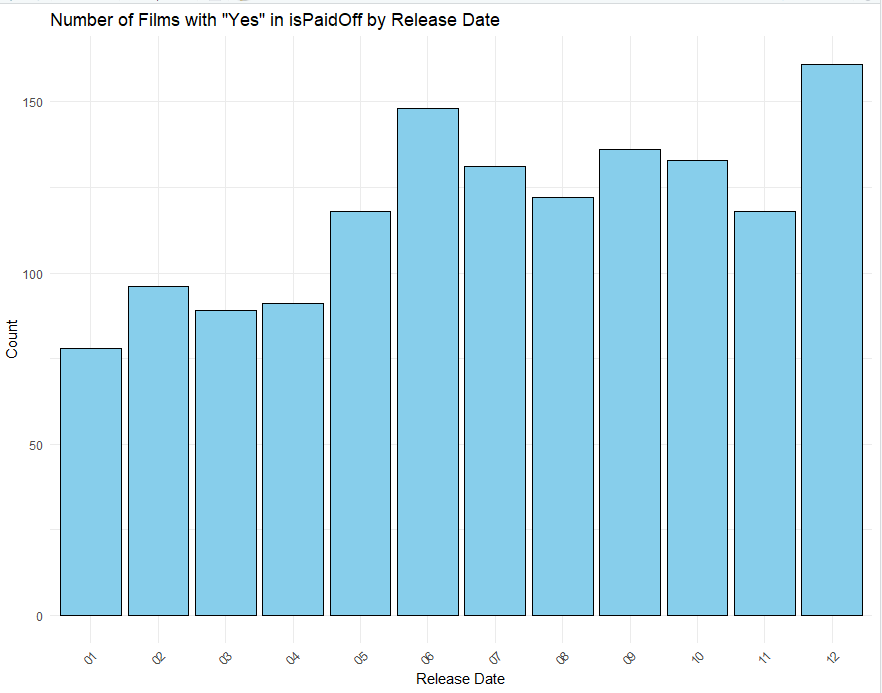
**Рисунок 5: График популярности жанров в кинематографе.**

****

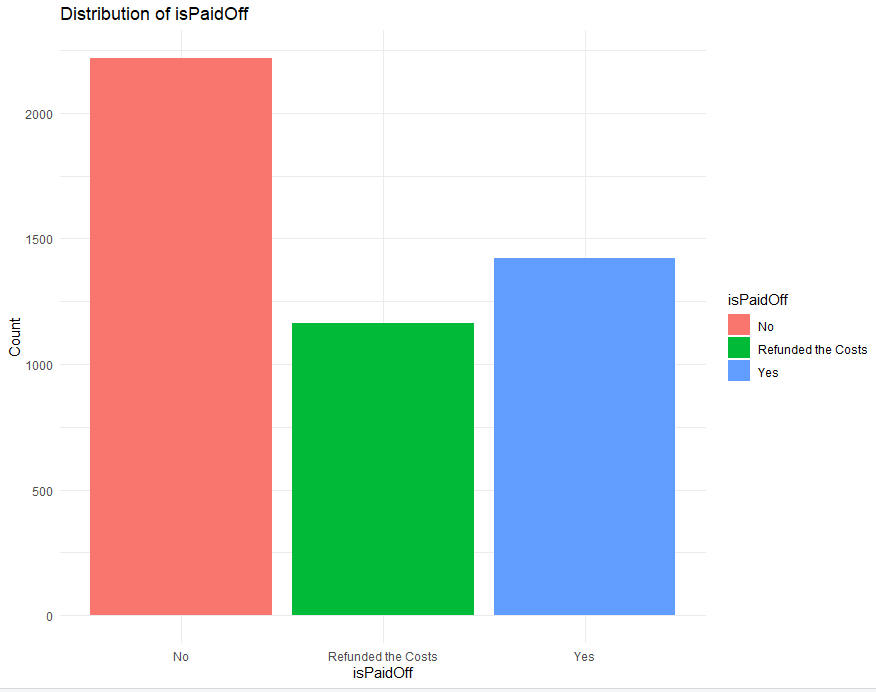
**Рисунок 6: График окупившихся фильмов по размеру бюджета.**

****

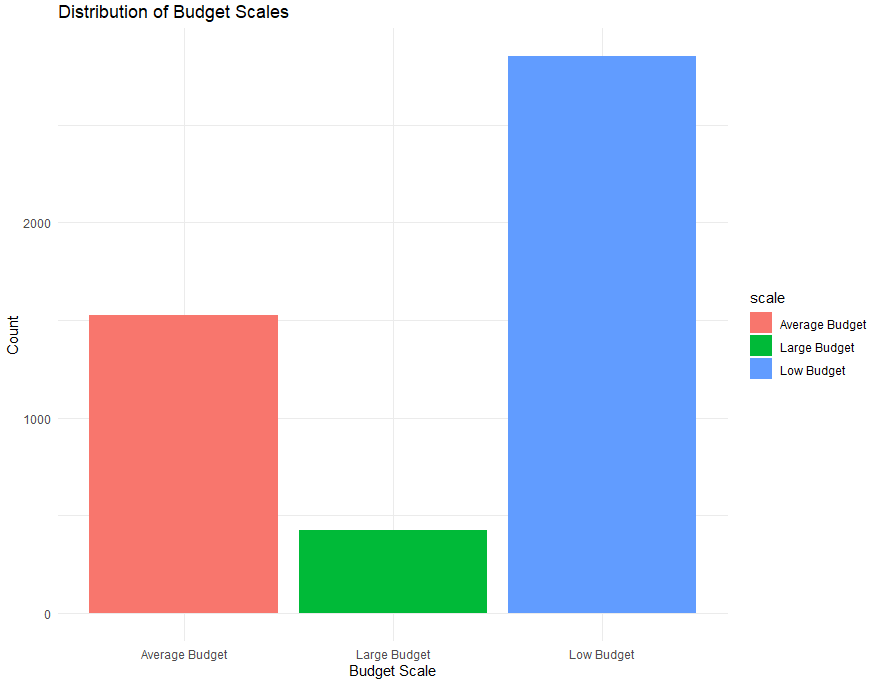
**Рисунок 7: График окупаемости фильмов по 5 самым крупным производителям.**

****

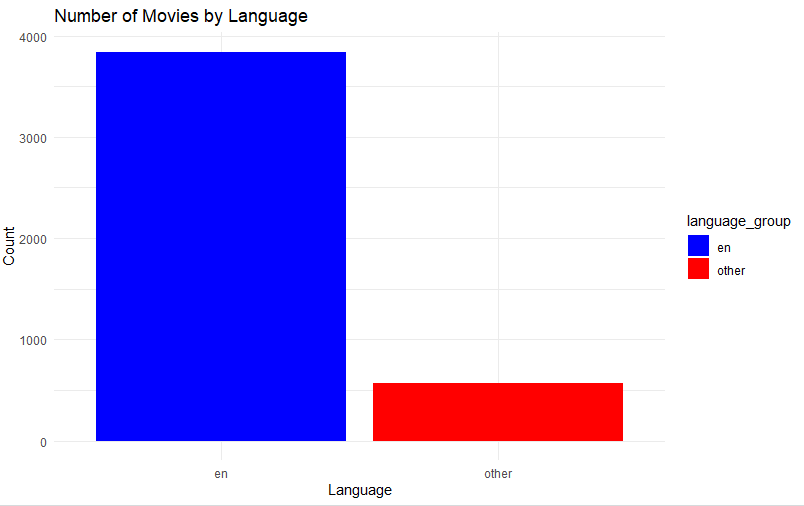
**Рисунок 8. График выхода фильмов в размере по месяцам.**



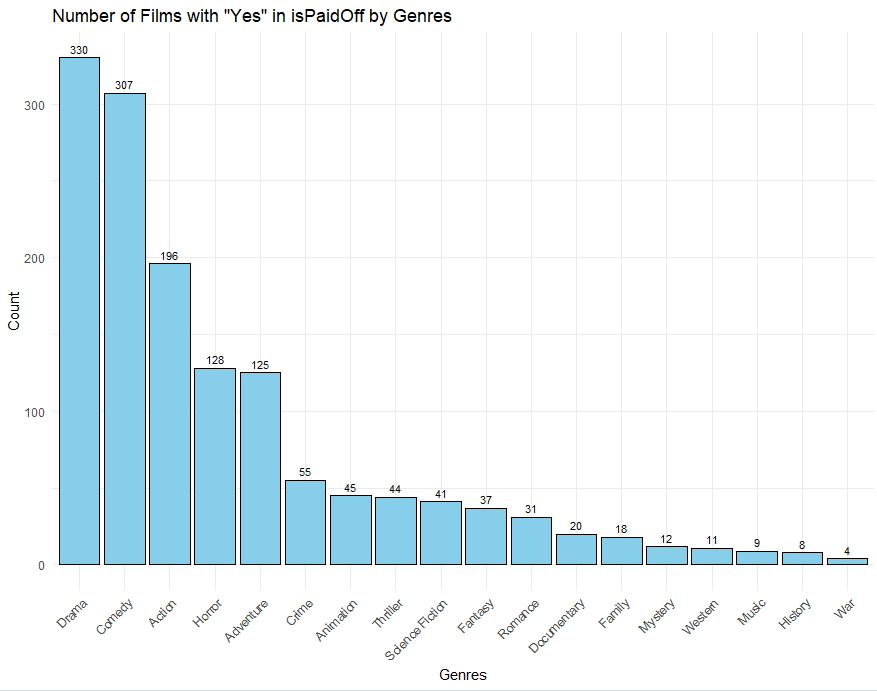
**Рисунок 9: График окупаемости фильмов в разрезе размера бюджета.**

****

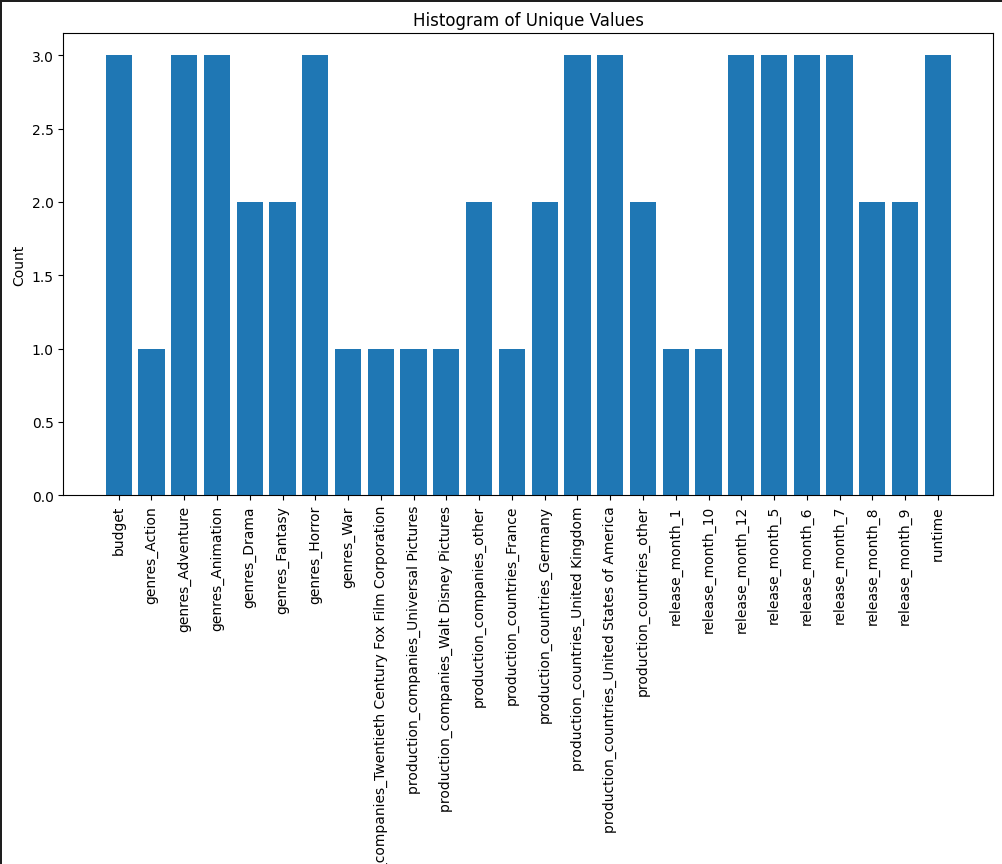
**Рисунок 10: График распределения фильмов по размеру бюджета.**

****

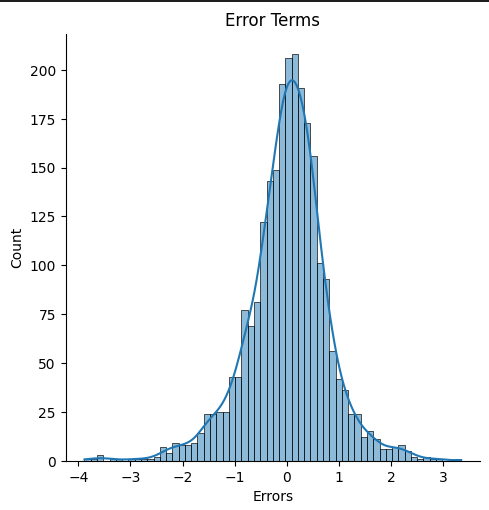
**Рисунок 11: График распределения фильмов по языку.**

****

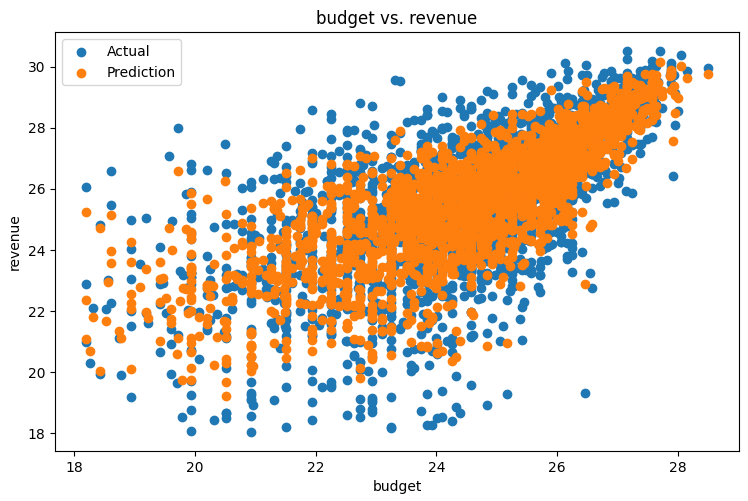
**Рисунок 12: График окупившихся фильмов по жанру.**

****

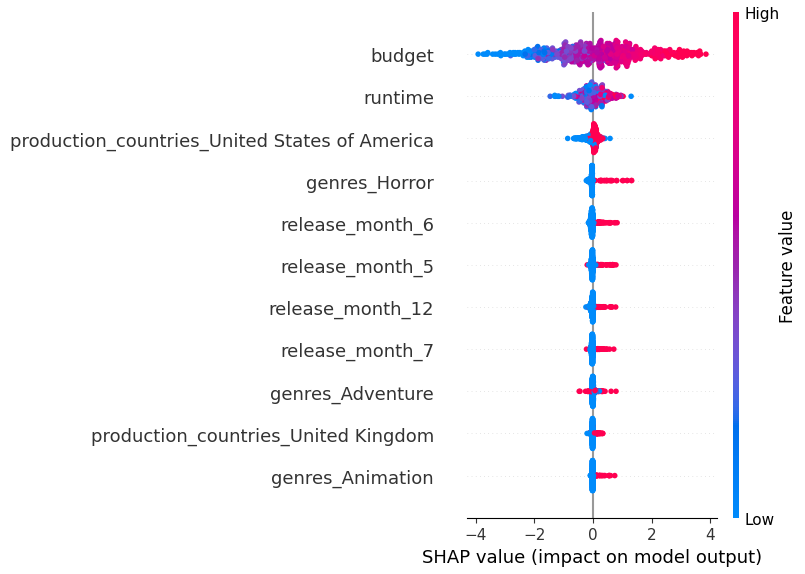
**Рисунок 13. График количества выбора переменных 3 способами отбора.**

****

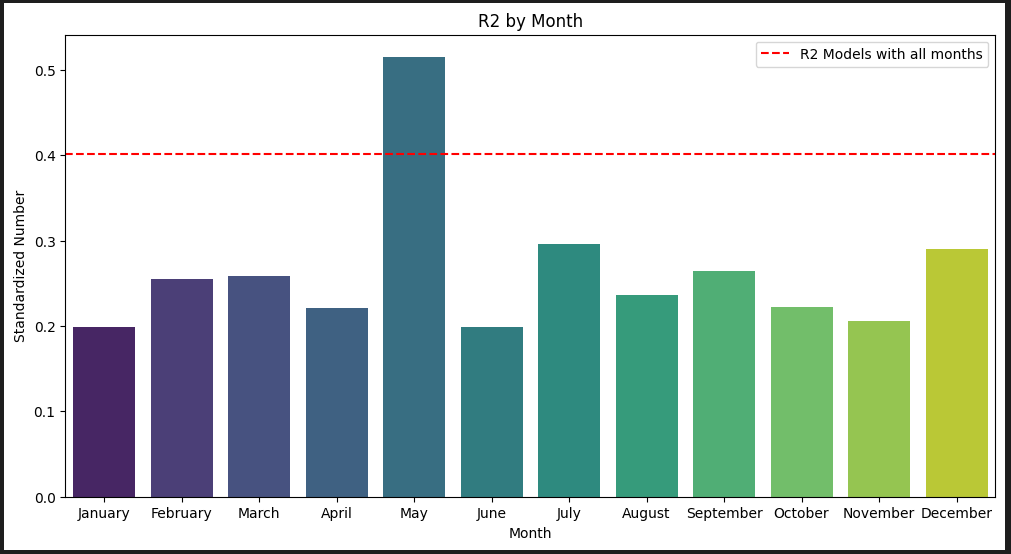
**Рисунок 14. График ошибки модели с отобранными переменными.**

****

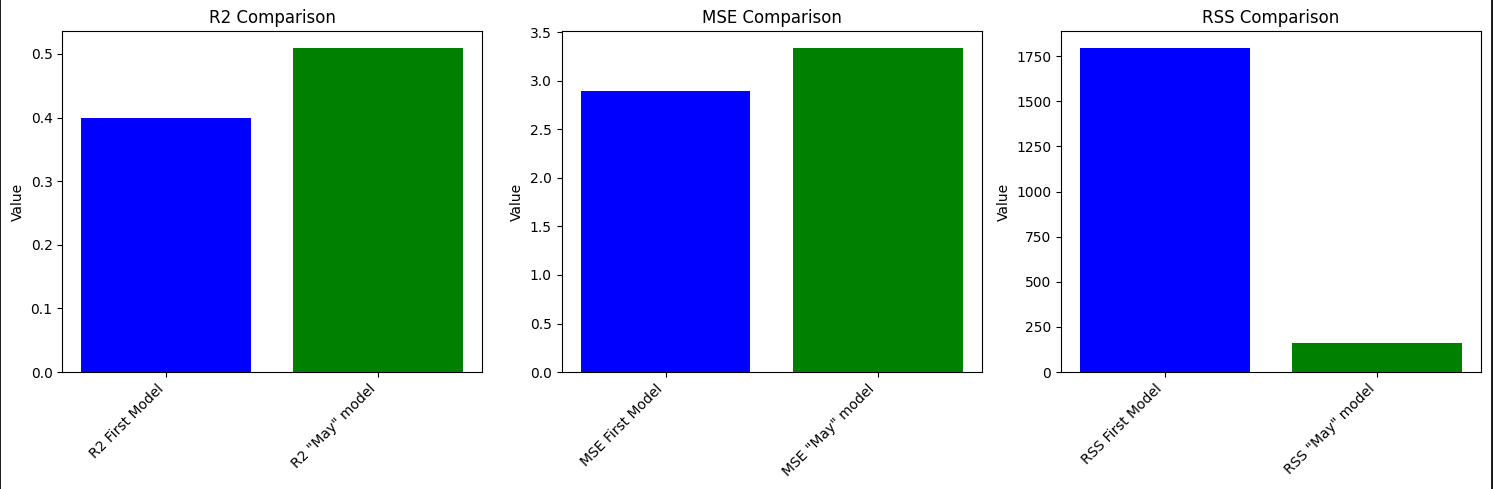
**Рисунок 16. График предсказанных и фактических значений с отобранными переменными.**

****

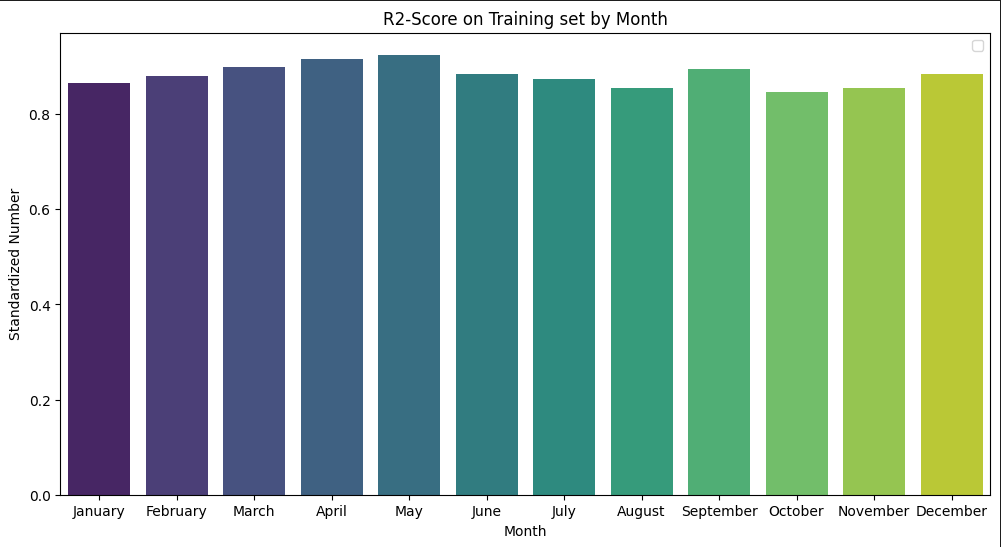
**Рисунок 17. График значений Шапли для модели с отобранными переменными.**

****

**Рисунок 18. R2 моделей по месяцам.**

****

**Рисунок 19. Сравнение показателей 2 моделей.**

****

**Рисунок 20. R2 обучающего набора моделей по месяцам.**

Библиография:

1. Документация к библиотеке jsonlite - <https://cran.r-project.org/web/packages/jsonlite/jsonlite.pdf>
2. Документация к библиотеке stringr - <https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html>
3. Документация к библиотеке Pandas - <https://pandas.pydata.org/docs/>
4. Документация к функции pandas.get\_dummies() <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html>
5. Статья “What is Exploratory Data Analysis ?” - <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>
6. Документация к библиотеке ggplot - <https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf>
7. Статья “Feature Selection Techniques in Machine Learning” - <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/>
8. Статья “Lasso vs Ridge vs Elastic Net” - <https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/>
9. Статья “Lasso vs Ridge vs Elastic Net” - <https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/>
10. Статья “Feature selection with Random Forest” - <https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/>
11. Статья “Linear Regression in Machine learning” - <https://www.geeksforgeeks.org/ml-linear-regression/>
12. Документация к библиотеке scikit-learn - <https://scikit-learn.org/stable/user_guide.html>
13. Статья “An introduction to explainable AI with Shapley values” - https://shap.readthedocs.io/en/latest/example\_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

1. Документация к библиотеке jsonlite - https://cran.r-project.org/web/packages/jsonlite/jsonlite.pdf [↑](#footnote-ref-0)
2. Документация к библиотеке stringr - https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html [↑](#footnote-ref-1)
3. Документация к библиотеке Pandas - https://pandas.pydata.org/docs/ [↑](#footnote-ref-2)
4. Документация к функции pandas.get\_dummies() https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get\_dummies.html [↑](#footnote-ref-3)
5. Статья “What is Exploratory Data Analysis ?” - https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/ [↑](#footnote-ref-4)
6. Документация к библиотеке ggplot - https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf [↑](#footnote-ref-5)
7. Статья “Feature Selection Techniques in Machine Learning” - https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/ [↑](#footnote-ref-6)
8. Статья “Lasso vs Ridge vs Elastic Net” - https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/ [↑](#footnote-ref-7)
9. Статья “Lasso vs Ridge vs Elastic Net” - https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/ [↑](#footnote-ref-8)
10. Статья “Feature selection with Random Forest” - https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/ [↑](#footnote-ref-9)
11. Статья “Linear Regression in Machine learning” - https://www.geeksforgeeks.org/ml-linear-regression/ [↑](#footnote-ref-10)
12. Документация к библиотеке scikit-learn - https://scikit-learn.org/stable/user\_guide.html [↑](#footnote-ref-11)
13. Статья “An introduction to explainable AI with Shapley values” - https://shap.readthedocs.io/en/latest/example\_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html [↑](#footnote-ref-12)