**Национальный исследовательский университет «Высшая школа экономики» Московский институт электроники и математики им. А.Н. Тихонова, Москва**

Направление 01.03.04. Прикладная математика Бакалаврская программа "Прикладная математика"

Отчет о самостоятельной работе по дисциплине «Методы анализа стохастических взаимосвязей»

Бригада №24

Новиков Глеб Сергеевич, 3 курс

Москва 2024

Оглавление

[1 Общая постановка задачи 3](#_Toc169387406)

[1.1 Основные гипотезы, которые планируется проверить в рамках исследования 4](#_Toc169387407)

[2 Предварительный анализ собранных данных 4](#_Toc169387408)

[2.1 Анализ количественных переменных 4](#_Toc169387409)

[2.2 Анализ качественных переменных 8](#_Toc169387410)

[2.3 Анализ статистической связи 11](#_Toc169387411)

[2.3.1 Графический анализ пары «числовая зависимая переменная – качественная 11](#_Toc169387412)

[независимая переменная» 11](#_Toc169387413)

[2.3.2 Анализ наличия корреляции между независимыми переменными 12](#_Toc169387414)

[2.3.3 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная» 12](#_Toc169387415)

[2.3.4 Предварительная проверка гипотез 15](#_Toc169387416)

[3 Проверка гипотез с помощью моделирования 15](#_Toc169387417)

[3.1 Построение базовой модели 15](#_Toc169387418)

[3.2. Проверка гипотез с помощью моделирования 16](#_Toc169387419)

[3.3 Оптимизация итоговой модели, сравнение качества моделей 17](#_Toc169387420)

[3.4. Проверка прогностических способностей модели 17](#_Toc169387421)

[4 Заключение 18](#_Toc169387422)

# 1 Общая постановка задачи

В таблице 1 представлена информация о приложениях Google Play

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Характеристика объекта/явления | Название переменной | Шкала измерения | Роль: целевая/объясняющая |
| 1 | Название приложение | App | - | - |
| 2 | Категория | Category | Номинальная | Объясняющая |
| 3 | Общий рейтинг приложения пользователями | Rating | Относительная | Объясняющая |
| 4 | Количество отзывов пользователей о приложении | Reviews | Относительная | Объясняющая |
| 5 | Размер приложения | Size |  | Объясняющая |
| 6 | Количество пользовательских установок приложения | Installs | Относительная | Объясняющая |
| 7 | Платное или бесплатное | Type | Номинальная | Объясняющая |
| 8 | Цена приложения | Price | Относительная | Целевая |
| 9 | Возрастная группа | Content Rating | Номинальная | Объясняющая |
| 10 | Дата последнего обновления приложения в Play Store | Last Updated | Номинальная | Объясняющая |
| 11 | Текущая версия приложения доступная в Play Store | Current Ver | Номинальная | Объясняющая |
| 12 | Минимальная требуемая версия Android | Android Ver | Номинальная | Объясняющая |

Таблица 1. Описание факторов, учтенных в анализе.

Исследуемые данные были получены из набора данных, взятых с сайта: <https://www.kaggle.com/datasets/lava18/google-play-store-apps>

## 

## 1.1 Основные гипотезы, которые планируется проверить в рамках исследования

Простая гипотеза:

1. С увеличением оценки на приложение увеличивается цена на приложение.

С увеличением рейтинга приложения будет увеличиваться его качественность и полезность, вследствие чего повысится цена.

Сложные гипотезы:

1. Сила влияния оценки приложения на цену зависит от количества оценок. Чем больше оценок, тем влияние больше.

Чем больше количество оценок, тем их ценность и справедливость больше, поэтому при увеличении их количества влияние самой средней оценки на цену повысится.

1. Сила влияния размера приложения на его цену зависит от типа приложения. Если тип приложения функциональный (TOOLS), то влияние больше.

Так как, чаще всего, чем больше размер приложения, тем он качественнее и совершеннее, а эти критерия наиболее ценны для приложений функционального типа. (Например, для приложений-игр, это будет не так значимо, так как далеко не всегда большой размер игры сказывается на ее качестве).

# 2 Предварительный анализ собранных данных

## 2.1 Анализ количественных переменных

1. **Rating (рейтинг)**

Переменная «Rating» представляет собой рейтинг/оценку, оставленные пользователями приложениям из Google Play.

Количество столбцов в гистограмме здесь, и далее, в других переменных, рассчитывалось по формуле Стерджесса и Фридману-Диаконису.



Рисунок 1. Гистограмма оценок пользователей

|  |  |
| --- | --- |
|  | Rating |
| Количество | 7534 |
| Среднее | 4,173547 |
| Стандартное отклонение | 0,545417 |
| Минимальное количество | 1 |
| Первый квартиль | 4 |
| Второй квартиль | 4,3 |
| Третий квартиль | 4,5 |
| Максимальное количество | 5 |
| Межквартильный размех | 0,5 |
| Коэффициент ассиметрии | -1,76364 |
| Коэффициент эксцесса | 5,132662 |

Таблица 2. Описательная статистика переменной Rating

По приведенной на рисунке 1 гистограмме можно заметить наличие асимметрии влево, кроме того, об этом свидетельствует таблица основных статистик (т.к. среднее меньше медианы и коэффициент асимметрии равен -1,76364). Это связано с тем, что гораздо чаще приложениям ставят положительные оценки, нежели негативные. Эти результаты соответствуют здравому смыслу, так как разработчикам невыгодно создавать некачественные приложения, а даже если такие и есть, то их малое количество. Согласно гистограмме, выбросы в данных присутствуют. Однако, необходимо проверить это по правилу трех сигм. Имеем: вероятность, что наблюдаемое значение отличается от среднего значения (4,173547) больше, чем на три стандартных отклонения (3\*0,545417), меньше или равна 1/9. То есть все значения меньшие 2.5 можно считать выбросами. В наших данных таких значений 143, что не очень много, так как в общей выборке более 7000 значений.

1. **Reviews (количество оценок)**

Переменная «Reviews» представляет собой количество оценок, оставленных некоторому приложению из GOOGLE PLAY. Так же представляем гистограммы по Стерджессу и Фрид.-Диаконису.



Рисунок 2. Гистограмма количества оценок

|  |  |  |
| --- | --- | --- |
|  | Reviews | |
| Количество | | 7534 |
| Среднее | | 298130,6 |
| Стандартное отклонение | | 1885513 |
| Минимальное количество | | 1 |
| Первый квартиль | | 105 |
| Второй квартиль | | 2253 |
| Третий квартиль | | 38367 |
| Максимальное количество | | 44893888 |
| Межквартильный размех | | 38262 |
| Коэффициент ассиметрии | | 13,49234 |
| Коэффициент эксцесса | | 237,7028 |

Таблица 3. Описательная статистика переменной Reviews

По приведенной на рисунке 2 гистограмме можно заметить наличие асимметрии вправо, кроме того, об этом свидетельствует таблица основных статистик (т.к. среднее много больше медианы и коэффициент асимметрии равен 13,49234). Так же исходя из таблицы и графика видим, что величина первого столбца также объясняется большим значения коэффициента эксцесса (237,7028). В данном случае гистограмма представлена для первых трех квартилей, так как в противном случае график имел бы меньше смысла. Но несмотря на это нетрудно заметить, что количество приложений с минимальным количеством отзывов (хотя бы до 2253) гораздо больше, чем других, что логично, так как число всемирно известных игр и приложений мало. Однако выбросов в данном случае (89) меньше, чем для переменной «Rating». Это можно объяснить тем что, несмотря на то, что известных приложений не так много, количество отзывов на них значительно влияет на среднее значение (1885513), вследствие чего по правилу трех сигм, в выбросы попадают значения большие 6071542.

1. **Size (размер приложения)**

Переменная «Size» представляет собой размер приложения в мегабайтах.



Рисунок 3. Гистограмма размера приложения

|  |  |
| --- | --- |
|  | Size |
| Количество | 7534 |
| Среднее | 23,01407 |
| Стандартное отклонение | 23,49896 |
| Минимальное количество | 0,008 |
| Первый квартиль | 5,2 |
| Второй квартиль | 14 |
| Третий квартиль | 33 |
| Максимальное количество | 100 |
| Межквартильный размех | 27,8 |
| Коэффициент ассиметрии | 1,42968 |
| Коэффициент эксцесса | 1,423395 |

Таблица 4. Описательная статистика переменной Size

По приведенной на рисунке 3 гистограмме можно заметить наличие асимметрии вправо, кроме того, об этом свидетельствует таблица основных статистик (т.к. среднее больше медианы и коэффициент асимметрии равен 1,42968). Согласно гистограмме, выбросы в данных присутствуют. Однако, необходимо проверить это по правилу трех сигм. Имеем: вероятность, что наблюдаемое значение отличается от среднего значения (23,01407) больше, чем на три стандартных отклонения (3\*14), меньше или равна 1/9. То есть все значения большие 94 можно считать выбросами. В наших данных таких значений 147, что не очень много, так как в общей выборке более 7000 значений.

1. **Price (стоимость приложения)**

Переменная «Price» представляет собой цену за приложение из Google Play в долларах. Так как большая часть приложений являются бесплатными гистограмма построена только для платных приложений, а таблица для всех.



Рисунок 4. Гистограмма цены приложения

|  |  |
| --- | --- |
|  | Price |
| Количество | 7534 |
| Среднее | 1,149388 |
| Стандартное отклонение | 17,62334 |
| Минимальное количество | 0 |
| Первый квартиль | 0 |
| Второй квартиль | 0 |
| Третий квартиль | 0 |
| Максимальное количество | 400 |
| Межквартильный размех | 0 |
| Коэффициент ассиметрии | 21,89911 |
| Коэффициент эксцесса | 486,1872 |

Таблица 5. Описательная статистика переменной Price

Исходя из таблицы можно заметить, что максимальным значением является 400 долларов, это объясняется тем, что это максимально значение допустимое сервисом. Также видно, что более 75% процентов выборки являются бесплатными, что логично, так как большая часть приложений действительно являются бесплатными. Исходя из этого, можно сделать вывод, что присутствует асимметрия вправо, однако, даже для выборки только из платных приложений, это асимметрия сохраняется (видно на графике). Также в выборке присутствуют выбросы. Все значения большие 80 можно считать выбросами, в наших данных таких значений 17. В основном это либо специальные приложения, используемые в медицине, либо же необоснованно дорогие, приложения-пустышки. Это объясняется тем, что если даже если приложение является платным, его цена невелика, а остальные являются исключением.

## 2.2 Анализ качественных переменных

1. **Category (категория приложения)**

Переменная Category описывает категорию, к которой относится приложение.



Рисунок 5. Гистограмма категорий приложения

Исходя из диаграммы, видно, что самая большая категория — это инструменты, а оставшиеся категории заметно уступают ей. Это логично, так как этими приложениями люди чаще всего пользуются в повседневной жизни.

1. **Installs (количество скачиваний)**

Переменная Installs отвечает за количество скачиваний приложений.



Рисунок 6. Гистограмма количества скачиваний

Исходя из диаграммы, видно, что самая большая категория — это количество скачиваний более 500.000, что можно объяснить тем, что очень плохих и очень хороших приложений мало, и такое значение является средним для всех приложений в целом.

1. **Type (платное или бесплатное приложение)**

Переменная Type показывает, является ли приложение платным или бесплатным.



Рисунок 7. Гистограмма типа приложения

По диаграмме видно, что подавляющее большинство приложений – бесплатные. Очевидно, что бесплатные приложения пользуются большим спросом, так как мало кто готов платить за приложение с немного лучшим функционалом, из-за чего приоритет отдается бесплатному аналогу.

1. **Content Rating (целевая аудитория)**

Переменная Content Rating показывает группы людей, в данном случае возрастные, для которых предназначено данное приложение.



Рисунок 8. Гистограмма целевой аудитории

Исходя из диаграммы, видно, что самая большая группа — это для всех. Очевидно, что приложения для большой группы людей будут пользоваться большим спросом. Кроме того, разработчики часто не задумываются о том, для какого возраста конкретно предназначено данное приложение, кроме случаев, когда необходима метка 18+ в связи со специфическим контентом.

1. **Last Updated (последнее обновление)**

Переменная Last Updated показывает год последнего обновления приложения разработчиками.



Рисунок 9. Гистограмма последних обновлений

Исходя из диаграммы, видно, что подавляющее большинство приложений обновлено в 2018/2017 годах. Количество приложений обновленных в 2010-2013 годах ничтожно мало, а это значит, что различные ошибки в функционировании приложения вряд ли исправлялись, исходя из чего, можно сказать, о том, что данные приложения не будут пользоваться спросом.

1. **Current Ver**

Переменная Current Ver показывает текущую версию приложения, доступную в Play Store.



Рисунок 10. Гистограмма текущих версий

Исходя из данной диаграммы можно заметить, что большее количество приложений, находящихся в Play Store, с первой версией.

1. **Android Ver**

Переменная Android Ver показывает минимально необходимую версию Android версию приложения, доступную в Play Store.



Рисунок 11. Гистограмма минимальных необходимых версий

Исходя из диаграммы, видно, что подавляющее большинство приложений требует версию 4 и выше, так, очевидно, что такие приложения захватывают более широкий круг пользователей в связи с универсальностью требуемой платформы.

## 2.3 Анализ статистической связи

### 2.3.1 Графический анализ пары «числовая зависимая переменная – качественная

### независимая переменная»



Рисунок 12. Диаграмма взаимосвязи категории приложения и рейтинга



Рисунок 13. Диаграмма взаимосвязи категории приложения и рейтинга

Для более удобного восприятия диаграмма была разделена на две части.

Из полученной диаграммы Бокса-Уискера видно наличие зависимости рейтинга приложения от его категории (типа). Несмотря на то, что значения схожи, можно заметить, что у приложений типов “EDUCATION”, “BOOKS\_AND\_REFERENCES”, “HEALTH\_AND\_FITNESS” верхний квартиль находится выше остальных, а это значит, что данным типам приложений более высокие оценки ставят чаще. Также видно, что у приложений “EDUCATION” и “ENTERTAINMENT”, гораздо меньше выбросов, чем у остальных, что означает, что тип таких приложений низкие оценки ставят реже.

Проверив наличие статистической связи с помощью непараметрического дисперсионного анализа с использованием критерия Краскела-Уоллиса, получаем значение pvalue=1.1130331508747944e-18. Это говорит о том, что гипотеза об отсутствии связи отвергнута, соответственно категория приложения влияет на его оценку пользователями.

### 2.3.2 Анализ наличия корреляции между независимыми переменными

Категория приложения (Category) и его тип(Type), то есть платное либо бесплатное.

В exel файле приведена таблица сопряжённости для Category и Type. Для

исследования связи между данными переменными воспользуемся статистикой Хи-квадрат: Хи-квадрат = 106.88 (20 ст. свободы, p-значение = 7.227551890309769e-14). Таким образом, очень маленькое p-значение говорит о существовании связи между данными переменными. Исследуем, насколько сильна взаимосвязь между данными признаками с помощью коэффициента V-Крамера. Коэффициент V-Крамера равен 0.08422462946149975, что говорит об умеренной связи исследуемых переменных.

### 2.3.3 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная»

* 1. **Взаимосвязь цены за приложение и размера приложения**

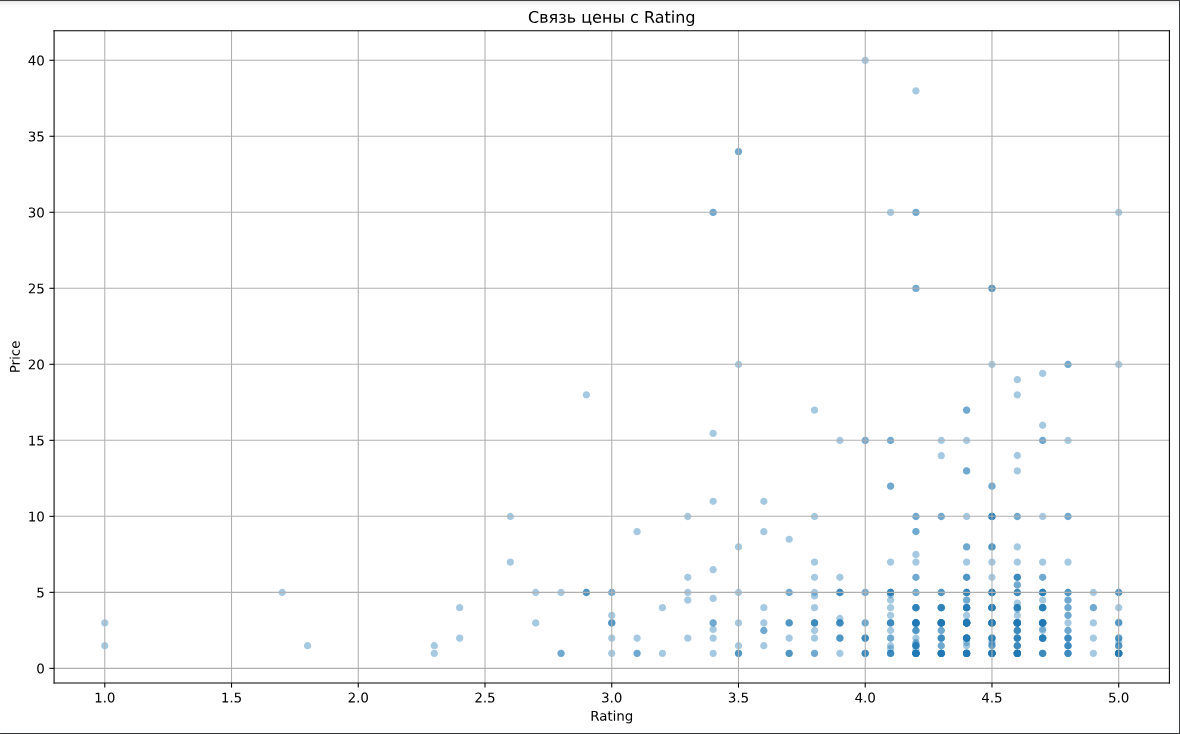


Рисунок 14. Диаграмма взаимосвязи цены приложения и рейтинга

Исходя из данной диаграммы, можно заметить, что зависимость присутствует, хотя и не очень сильная. С увеличением рейтинга увеличивается цена приложения, но это происходит до некоторой оценки (около 4.7), а после этого цена опять начинает падать. Стоит отметить, что приложениям с очень высокой ценой не ставят хорошие оценки, так как цена превышает разумную (данный график представлен, для цен без учета выбросов).

Для формальной проверки гипотезы о наличии связи подсчитаем коэффициенты корреляции Пирсона и Спирмена и приведём результаты проверки их значимости.

|  |  |  |
| --- | --- | --- |
|  | Коэффициент  корреляции  Пирсона | Коэффициент  корреляции  Спирмена |
| Полученное значение | -0,02165 | -0,05967 |
| Значимость | 0,060199 | 2,18E-07 |

Таблица 6. Таблица коэффициентов корреляций для Price и Rating

Можно сделать вывод, что гипотеза об отсутствии статистической связи может быть отвергнута. Мы получили относительно значимые коэффициенты. Таким образом, рейтинг приложения влияет на его цену, причем отрицательно, так как коэффициенты отрицательны.

* 1. **Взаимосвязь цены приложения и количества отзывов**

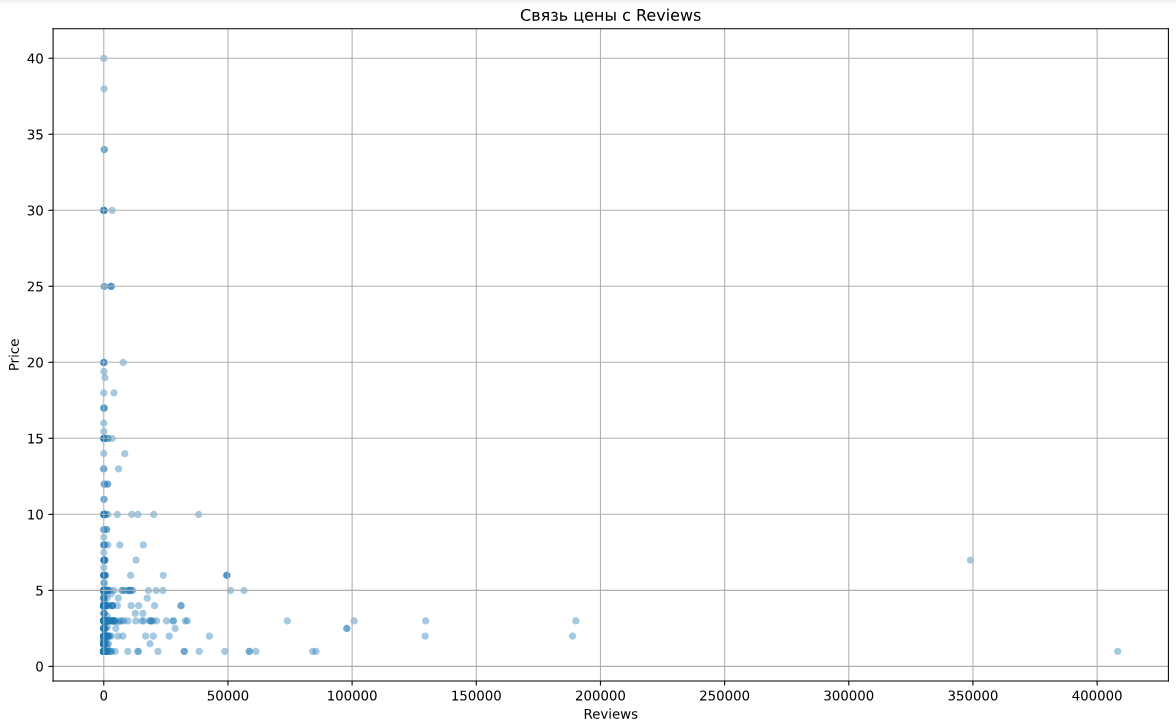


Рисунок 15. Диаграмма взаимосвязи цены приложения и количества отзывов

Исходя из данной диаграммы, также можно заметить, что присутствует очень слабая связь между переменными количество отзывов на приложение и цена за приложение. В отличие от предыдущей диаграммы тут она менее заметна, так как для малоизвестных приложений (количество отзывов на которые стремятся к нулю) соответствую цены всего диапазона. Однако можно выделить то, что среди приложений с большим количеством отзывов, платные почти отсутствуют (данный график представлен без учета выбросов).

Для формальной проверки гипотезы о наличии связи подсчитаем коэффициенты корреляции Пирсона и Спирмена и приведём результаты проверки их значимости.

|  |  |  |
| --- | --- | --- |
|  | Коэффициент  корреляции  Пирсона | Коэффициент  корреляции  Спирмена |
| Полученное значение | -0,01025 | -0,17424 |
| Значимость | 0,37384 | 1,98E-52 |

Таблица 7. Таблица коэффициентов корреляций для Price и Reviews

Можно сделать вывод, что связь присутствует, так как коэффициенты по модулю высокие и положительные, что означает, что с повышением размера приложения увеличивается его оценка, что логично, так как чаще всего, чем больше размер приложения, тем больше он содержит интересного пользователям контента и функционала.

* 1. **Взаимосвязь цены за приложение и размера приложения**

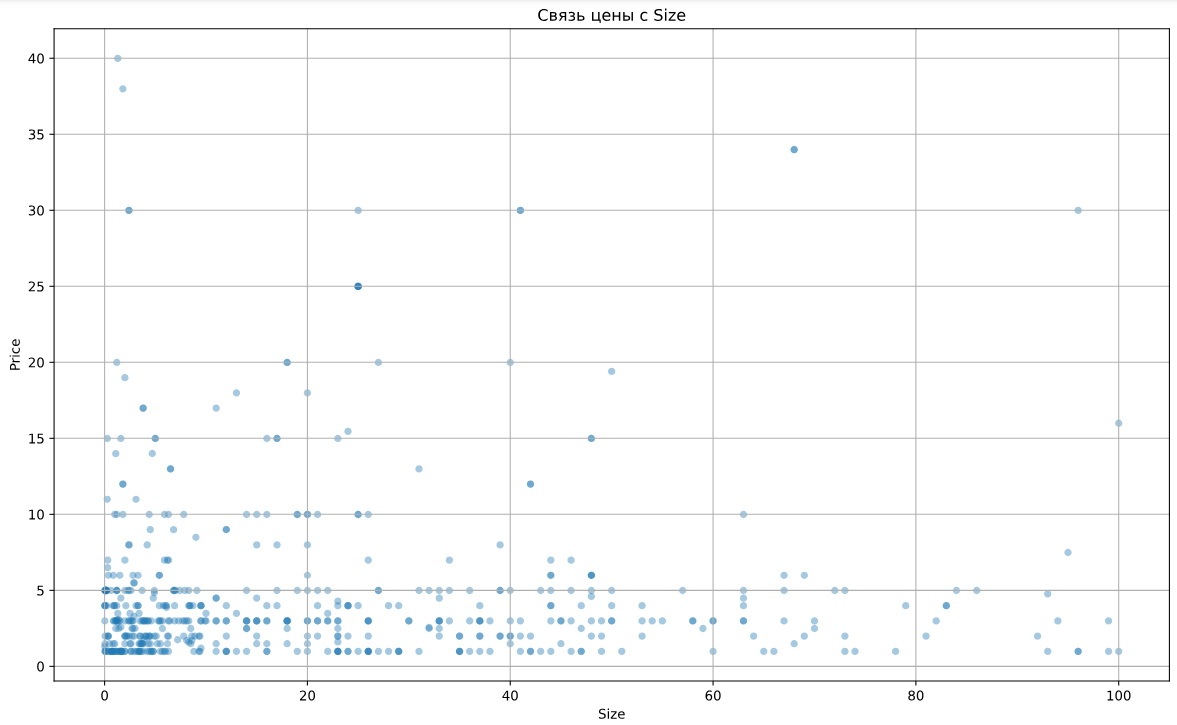


Рисунок 16. Диаграмма взаимосвязи цены приложения и размера

На данной диаграмме можно заметить зависимость цены за приложения и размера. Несмотря на то, что зависимость кажется не самой сильной, она присутствует, что видно по графику. Тем не менее, данные опровергают предположения о том, что при увеличении размера приложения, должна увеличиваться цена. На самом деле наоборот, при увеличении размера цена уменьшается, так же как и количество платных приложений в принципе.

Для формальной проверки гипотезы о наличии связи подсчитаем коэффициенты корреляции Пирсона и Спирмена и приведём результаты проверки их значимости.

|  |  |  |
| --- | --- | --- |
|  | Коэффициент  корреляции  Пирсона | Коэффициент  корреляции  Спирмена |
| Полученное значение | -0,02652 | -0,04796 |
| Значимость | 0,021358 | 3,12E-05 |

Таблица 8. Таблица коэффициентов корреляций для Price и Size

Можно сделать вывод, что гипотеза об отсутствии статистической связи может быть отвергнута. Мы получили относительно значимые коэффициенты. Таким образом размер приложения влияет на его цену.

### 2.3.4 Предварительная проверка гипотез

Таким образом, исходя из предварительного анализа данных, можем объявить об отвержении первой гипотезы, а именно, что с увеличением оценки на приложение, увеличивается цена на приложение. Как бы парадоксально это не казалось, но на самом деле происходит наоборот. Также отметим, что количество оценок на приложение влияет на его цену, а именно с увеличением количества оценок уменьшается его цена. Это замечание пригодится для проверки второй гипотезы. Далее, можем отметить важное замечание для третей гипотезы, что размер приложения влияет на его цену, а именно с увеличением размера уменьшается его цена.

# Проверка гипотез с помощью моделирования

## 3.1 Построение базовой модели

В качестве базовой модели рассматривалась линейная регрессия, в качестве регрессоров модели рассматривались те признаки, который были указаны в начальной, также в модель была добавлена константа. Для качественных признаков было произведено кодирование, таким образом каждый из таких признаков разделился на соответствующее число столбцов, значение в которых равно True (1) или False (0). Выборка разделяется случайным образом на обучающую (80% от общего объема) для построения модели и тестовую (20%) для проверки прогностических свойств. Таким образом, уравнение регрессии можно записать в следующем виде:

RSum = a\_0 + a\_1 \* «Category\_BOOKS\_AND REFERENCE» +…+ a\_i \* «Installs 0» +… +…+ a\_j \* «Type\_PAID» + a\_(j+1) \* «Content Rating\_EVERYONE 10» +…+a\_k \* «Last Updated 2011» + a\_m \* «Current Ver\_V.0» +…+ a\_n \* «Android Ver\_V.7 AND UP» + a\_(n+1) \* «Rating» + a\_(n+2) \* «Size» + a\_(n+3) \* «Reviews», где n = 45.

Построив модель, используя все объясняющие переменные, получим следующий результат, получим следующий результат для проверки уровня мультиколлинеарности:

|  |  |  |
| --- | --- | --- |
|  | vars | VIF |
| 0 | Rating | 1,978769 |
| 1 | Reviews | 1,083482 |
| 2 | Size | 2,083401 |

Таблица 9. Таблица уровней мультиколлинеарности

Далее необходимо  было бы оптимизировать структуру модели для повышения ее качества и возможного снижения уровня мультиколлинеарности, если бы значения VIF для некоторых переменных были больше 3, путем удаления некоторых из них и переоценивать модель. Однако, исходя из приведенной таблицы, видим, что все значения VIF меньше 3, что означает, что нам повезло и мультиколлинеарности не наблюдается.

Далее была произведена проверка базовой модели на наличие гетероскедастичности.

|  |  |
| --- | --- |
|  | 0 |
| LM | 1835,978 |
| LM\_P | 3,74E-97 |
| F | 3,193059 |
| F\_P | 1,7E-126 |

Таблица 10. Таблица теста Уайта

Результатом проверки тестов Уайта стало утверждение об отвержении нулевой гипотезы о гомоскедастичности модели (из таблицы видно, значение LM p-value = 3.74e\_97, что очень близко к 0).

Также приведем оценку качества полученной модели (критерий Akaike, R-sq и adjusted R-sq).

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Adjusted R-sq | Akaike | R-sq |
| base\_00 | 0.0573056 | 51553.3 | 0.0651275 |

Таблица 11. Таблица оценки качества базовой модели

Перейдем к проверке гипотез и оптимизации модели и будем следить и сравнивать качества модифицированных моделей.

## 3.2. Проверка гипотез с помощью моделирования

Проверим выдвинутые нами гипотезы

*Гипотеза 1:*

Посмотрим на коэффициент корреляции перед Rating, он отрицательный (-1.111), а значит отвергаем нулевую гипотезу о том, что при увеличении рейтинга увеличивается цена. Получаем, что рост рейтинга отрицательно влияет на цену приложения, что противоречит начальным ожиданиям, однако совпадает с выводом из предварительно анализа.

*Гипотеза 2:*

Для проверки этой гипотезы введем новую переменную “Reviews\_with\_Rating”. Модель для проверки изменится: заместо a\_i \* “Rating”, где a\_i коэффициент для рейтинга приложения заменится на слагаемое (a\_(i0) + a\_(i1) \* “Reviews”) \* “Rating”. Тогда a\_(i0) < 0, a\_(i1) < 0, a\_(i0) < a\_(i1). Здесь a\_(i1) коэффициент у переменной Reviews\_with\_Rating. Посмотрим теперь на результаты моделирования.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | coef | std err | t | P > | t | | 0.025 | 0.095 |
| Rating | -1.1159 | 0.432 | -2.580 | 0.010 | -1.964 | -0.268 |
| Reviews\_with\_Rating | 3.017e-07 | 1.41e-06 | 0.214 | 0.831 | -2.46e-06 | 3.07e-06 |

Таблица 12. Таблица сравнения коэффициентов Rating и Reviews\_with\_Rating

Итак, как видим из таблицы, действительно коэффициент при Rating = -1.1159 < 0, но при Reviews\_with\_Rating = 3.017e-07 > 0. Таким образом, делаем вывод, что гипотеза должна быть отвергнута, более того, заметим из следующей таблицы, что модель стала немного хуже:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Adjusted R-sq | Akaike | R-sq |
| base\_00 | 0.0573 | 51553.3 | 0.0651 |
| hyp\_01 | 0.0571 | 51553.3 | 0.0651 |

Таблица 13. Результаты анализа оптимальности построенной модели для 1 гипотезы

Исходя из этой таблицы, видим, что Adjusted R-sq для base\_00 > hyp\_01, что означает, что модель стала хуже и нужно вернуться к исходной модели.

*Гипотеза 3:*

Для проверки этой гипотезы введем новую переменную “ Size\_with\_TOOLS'”. Модель для проверки изменится: заместо a\_j \* “Size”, где a\_j коэффициент для размера приложения заменится на слагаемое (a\_(i0) + a\_(i1) \* “Category\_TOOLS”) \* “size”. a\_(i0) < a\_(i1). Здесь a\_(i1) коэффициент у переменной Size\_with\_TOOLS. Посмотрим теперь на результаты моделирования.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | coef | std err | t | P > | t | | 0.025 | 0.095 |
| Size | -0.0281 | 0.012 | -2.247 | 0.025 | -0.053 | -0.004 |
| Size\_with\_TOOLS | 0.0371 | 0.030 | 1.234 | 0.217 | -0.022 | 0.096 |

Таблица 14. Таблица сравнения коэффициентов при Size и Size\_with\_TOOLS

Итак, как видим из таблицы, действительно коэффициент при Size = -0.028, а при Size\_with\_TOOLS = 0.0371> 0. Таким образом, так как -0.028 < 0.0371, делаем вывод, что гипотеза справедлива, более того, заметим из следующей таблицы, что модель стала лучше:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Adjusted R-sq | Akaike | R-sq |
| base\_00 | 0.0573 | 51553.3 | 0.0651 |
| hyp\_01 | 0.0571 | 51553.3 | 0.0651 |
| hyp\_02 | 0.0573 | 51553.8 | 0.0653 |

Таблица 15. Результаты анализа оптимальности построенной модели для 2 гипотезы

Исходя из этой таблицы, видим, что R-sq для base\_00 < hyp\_02, также как и критерий Akaike, что означает, что модель стала лучше.

## 3.3 Оптимизация итоговой модели, сравнение качества моделей

Подвергнем итоговаую модель оптимизации за счет пошагового удаления

незначимых переменных. На каждом шаге будем переоценивать модель. Для финального варианта оценим качество модели с использованием критерия Akaike и adjusted R-sq.

Итак, для наших данных получили следующий вывод:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Adjusted R-sq | Akaike | R-sq |
| base\_00 | 0.0573 | 51553.3 | 0.0651 |
| hyp\_01 | 0.0571 | 51553.3 | 0.0651 |
| hyp\_02 | 0.0573 | 51553.8 | 0.0653 |
| final | 0.0675 | 51553.5 | 0.0683 |

Таблица 16. Результаты анализа оптимальности построенных моделей

Для объяснения этой таблицы посмотрим на критерий Adjusted R-sq. В оптимизированной модели он самый большой, что говорит о том, что вероятность верного прогноза выше (так же и с критерием R-sq), поэтому можем сделать вывод, что итоговая оптимизированная модель является лучшей. Это объясняется тем, что мы учли как результаты проверки гипотез, так и результаты предварительного анализа данных. Данное качество модели было достигнуто путем удаления незначимых фиктивных переменных.

## 3.4. Проверка прогностических способностей модели

Проверка прогностических способностей осуществляется для всех включенных в отчет моделей. Подсчитаем значения прогнозов для элементов тестовой

выборки и построить для них центральные доверительные интервалы на основе

нормального распределения для доверительной вероятности 95%. Для результатов рассчитаем среднеквадратическую погрешность прогнозирования и максимальную абсолютную погрешность прогнозирования. Итого имеем:

|  |  |  |
| --- | --- | --- |
|  | Среднеквадратичная погрешность | Абсолютная погрешность |
| Базовая модель | 16.0173 | 2.83868 |
| Итоговая модель | 16.0048 | 2.56111 |

Таблица 17. Сравнение прогностических способностей моделей

Как видим, проведенные изменения в модели улучшили ее прогностические способности, так как видно, что ошибка уменьшалась при вносимых изменениях.

# 4 Заключение

С использованием построенной модели потенциальный разработчик сможет оценить стоимость приложений на рынке на основе их различных характеристик. Наиболее точной в предсказании оказалась итоговая модель (с очень низкими погрешностями). Во время исследования были проверены гипотезы, касающиеся влияния тех или иных характеристик на цену приложения. В частности, подтвердилась гипотеза о влиянии размера приложения размере на его стоимость с учетом конкретного типа (в нашем случае функциональное приложение). Несмотря на то, что остальные гипотезы не подтвердились, их отрицание несет важный смысл. Например, что как бы это парадоксально не казалось, но цена приложения может отрицательно зависеть от его рейтинга.