МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский технологический университет «МИСиС»»

ИКН кафедра АСУ

КУРСОВАЯ РАБОТА

по дисциплине

«Прикладной статистический анализ»

на тему

«Разработка модели прогнозирования количества покупателей в торговом центре»

Выполнил:

студент 3-го курса, гр. БИВТ-21-4

Савенко Е.И.

Научный руководитель:

К.т.н., доцент, ученый

секретарь кафедры ИКТ

Маркарян А О.

Москва 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4

1. АНАЛИЗ ХАРАКТЕРИСТИК ОБЪЕКТА ИССЛЕДОВАНИЯ 5

1.1 Описание объекта исследования 5

1.2 Анализ объекта исследования с помощью статистических показателей 5

1.3 Выявление причинно-следственных связей 6

1.4 Постановка задачи моделирования 7

2. МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ 9

2.1 Формализация и классификация переменных 9

2.2 Проверка гипотезы о нормальном распределении выходной величины 9

2.3 Корреляционный анализ 10

2.4 Построение регрессионной модели 11

2.4.1 Структурная идентификация модели 11

2.4.2 Параметрическая идентификация модели 11

В соответствии с методом наименьших квадратов, задача заключается в аппроксимации кривой известной функцией. Вычисление параметров уравнения множественной линейной регрессии будет произведено с помощью алгоритма МНК 11

3. ИССЛЕДОВАНИЕ МОДЕЛИ 12

3.1 Анализ статистической значимости уравнения регрессии 12

3.3 Исследование мультиколлинеарности факторов 13

3.4 Применение шагового регрессионного анализа для улучшения модели 14

4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ 16

4.1 Обоснование выбора и описание программного обеспечения 16

4.2 Описание основных модулей программы 16

4.3 Численное исследование результатов моделирования 21

ВЫВОДЫ 23

СПИСОК ЛИТЕРАТУРЫ 24

ПРОГРАММНЫЕ ПРИЛОЖЕНИЯ 25
ПРИЛОЖЕНИЕ А 26
ПРИЛОЖЕНИЕ Б 27
ПРИЛОЖЕНИЕ В 29

ВВЕДЕНИЕ

Современная розничная торговля подвергается значительным изменениям под воздействием развития технологий, изменениями потребительского поведения и динамикой рыночных трендов. В таком динамичном окружении для эффективного управления торговым предприятием становится необходимым разработка надежных инструментов прогнозирования. Одним из важных аспектов в этом контексте является прогнозирование количества покупателей в торговых центрах. В рамках данной курсовой работы будет разработана модель прогнозирования количества покупателей в торговом центре.

Цель данной курсовой работы заключается в разработке модели прогнозирования количества покупателей в торговом центре с использованием статистических методов.

Актуальность данной задачи обусловлена стремительными изменениями в розничной сфере, где понимание и предвидение потребительского спроса играют ключевую роль в стратегическом управлении.

Задачами работы являются:

- анализ характеристик объекта исследования;
- моделирование статистических зависимостей;
- исследование модели;
- программная реализация и численное исследование результатов моделирования.

Предметом исследования является модель прогнозирования количества покупателей в торговых центрах на основе статистических методов.

Объектом исследования торговые центры в современной розничной торговле.

1. АНАЛИЗ ХАРАКТЕРИСТИК ОБЪЕКТА ИССЛЕДОВАНИЯ

1.1 Описание объекта исследования

Объектом исследования являются торговые центры в современной розничной торговле. На примере одного из крупнейших торговых центров Москвы ТРЦ «Европейского», была найдена первична статистическая информация о количестве посещений. Данные собраны за каждый месяц с января 2018 года до декабря 2022 включительно [3]. В графическом виде можно отследить зависимость посещений от времени (рисунок 1). Данные находятся в Приложении А.

Рис. 1 – Зависимость количества посещений ТРЦ «Европейский» от времени.

1.2 Анализ объекта исследования с помощью статистических показателей

Вычислив абсолютный прирост, равный -4624400 человек, можем выявить убывающую тенденцию.

По среднему темпу прироста, равному -7.5% можно увидеть во сколько раз в среднем уменьшились посещения ТРЦ за 4 года.

Предсказание по среднему абсолютному приросту на 2022 год составило 47830943 человек, а по среднему темпу роста: 48508684.

Из уравнение прямой: $y = -6040054.5 \cdot t + 71338859$ можно предсказать методом аналитического выравнивания количество посещений на 2022 год: 41138587 человек.

На графике также можно заметить убывающую тенденцию и отличие линейной регрессии от реальных значений (рисунок 2).

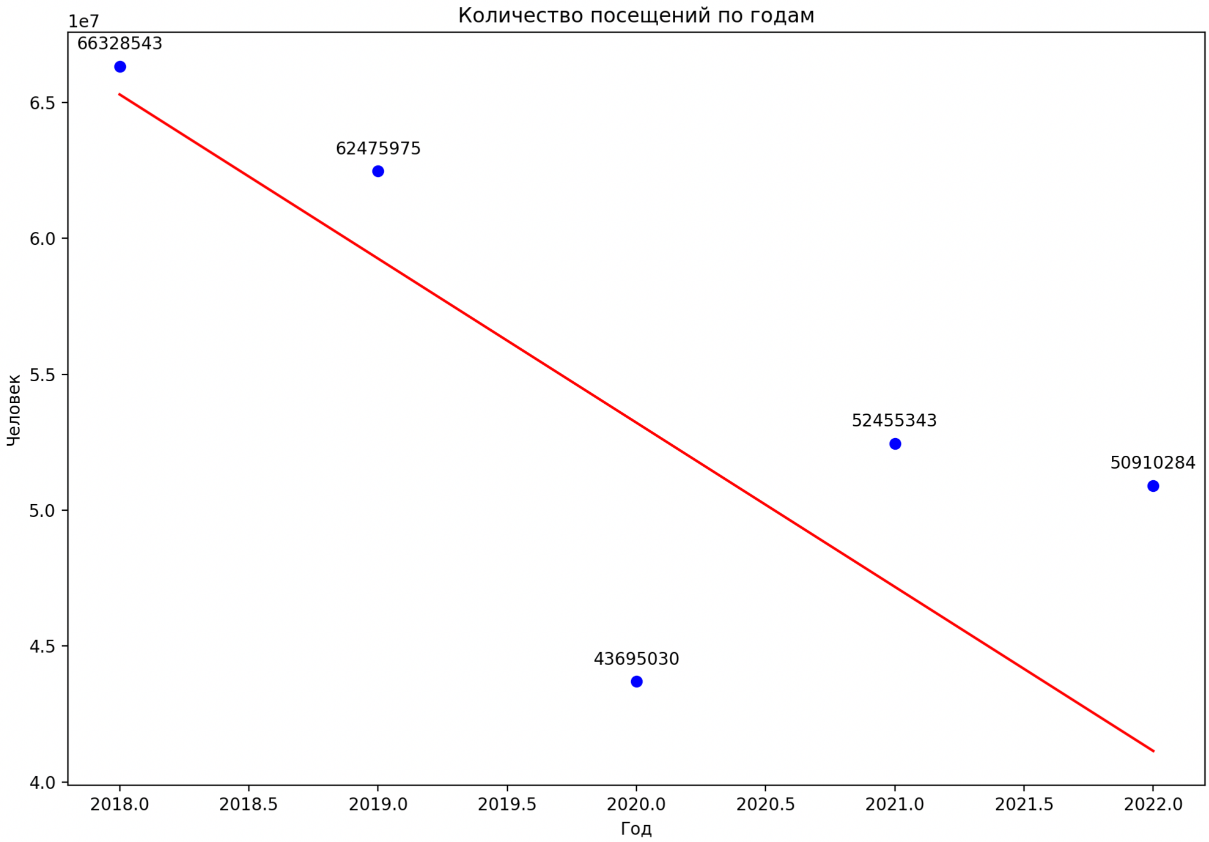

Рис. 2 – Количество посещений ТРЦ по годам

Подробные статистические показатели представлены в Приложении Б.

1.3 Выявление причинно-следственных связей

Изучение причинно-следственных связей для прогнозирования количества покупателей в торговом центре включает в себя анализ нескольких ключевых факторов, таких как:

- Средняя зарплата горожан, которая представляет собой значимый аспект, поскольку её увеличение может сопровождаться повышением покупательской активности.
- Индексы потребительских цен, которые отражают влияние инфляции на покупательскую способность.

- Уровень безработицы, который может служить индикатором экономической нестабильности и оказывать влияние на посещаемость торгового центра.
- Рекламные затраты, которые, в свою очередь, могут существенно влиять на привлекательность торгового центра. Большие рекламные кампании способны привлечь больше посетителей и, следовательно, повысить его общую посещаемость.
- Прирост населения, который может оказать влияние на посещение ТРЦ. При увеличении населения города возрастает потенциальная клиентская база, что может способствовать увеличению посещаемости торгового центра. Новые жители могут приносить с собой новый спрос на товары и услуги, что в свою очередь может сказаться на общем объеме продаж.
- Анализ данных о продажах в предыдущие периоды, который является важным компонентом модели, поскольку успешные периоды продаж могут привлечь больше посетителей в будущем.
- Временной аспект, который остается наиболее влиятельным фактором. Сезонные колебания, праздничные периоды или даже дни недели могут существенно варьировать количество посетителей.

Комплексный анализ факторов способствует лучшему пониманию причинно-следственных связей в прогнозировании количества покупателей в торговом центре, что позволяет выделить эффективнее управлять ресурсами и предлагать специальные предложения в периоды повышенного спроса. Для получения точных результатов, требуется проведение статистического анализа и моделирования, учитывая взаимосвязи между этими различными факторами.

1.4 Постановка задачи моделирования

Постановка задачи моделирования направлена на создание и обучение модели, способной прогнозировать количество покупателей в торговом центре. Для достижения этой цели предполагается использование специального набора данных, содержащего информацию о различных параметрах торгового процесса, таких как средняя зарплата горожан, праздничные дни и другие ключевые переменные.

Первоочередной задачей является подготовка и очистка данных, а также определение признаков, имеющих наибольшее влияние на количество покупателей. Для эффективного моделирования необходимо также провести анализ структуры данных, выявить возможные пропуски или выбросы, которые могут повлиять на качество модели.

Следующим этапом является выбор подходящего алгоритма, способного учесть особенности предсказания количества покупателей в зависимости от различных параметров.

В качестве математической модели лучше всего подойдет уравнение множественной регрессии, которое способно учесть факторы, влияющих на точность прогнозирования. Обучение модели будет проводиться на обучающем наборе данных, а затем ее эффективность будет проверена на тестовой выборке.

Оценка качества модели включает в себя анализ ее точности, чувствительности и специфичности, а также других метрик, адаптированных к задаче предсказания количества покупателей в торговом центре, с помощью построения графиков зависимостей и расчета коэффициентов корреляции и корреляционных отношений, а также выбора вида парной зависимости. Для улучшения точности модели необходимо провести отсев незначимых переменных с использованием шагового регрессионного анализа.

После анализа и интерпретации результатов будет выявлена основная цель данного моделирования, а именно предоставление торговым компаниям инструмента, способного на раннем этапе прогнозировать потенциальное количество посетителей в торговом центре.

2. МОДЕЛИРОВАНИЕ СТАТИСТИЧЕСКИХ ЗАВИСИМОСТЕЙ

2.1 Формализация и классификация переменных

Даны статистические данные о зависимости количества посещений ТЦ от потенциально полезных факторов $x_1, \ldots\ x_5$, на основании наблюдений за 5 лет.

x_1 – количественная дискретная переменная, показывающая зависимость от числа праздничных и выходных дней в каждом месяце, дни.[2]

x_2 – количественная дискретная переменная, показывающая зависимость от средней зарплаты по Москве, руб. [6]

x_3 – количественная дискретная переменная, показывающая зависимость от оборота розничной торговли непродовольственных товаров по месяцам по Москве, руб. [5]

x_4 – количественная дискретная переменная, показывающая зависимость от численности занятых в возрасте 15-72 лет по Москве, чел. [7]

x_5 – количественная дискретная переменная, показывающая зависимость от времени, месяц.

y – выходная количественная дискретная переменная (отклик), отражающая количество посещений ТЦ, чел.

Первичные статистические данные по каждому фактору представлены в Приложении В.

2.2 Проверка гипотезы о нормальном распределении выходной величины

Проверка гипотезы о нормальном распределении была осуществлена с помощью «Правила трёх сигм» и критерия Пирсона.

Для простоты вычислений перейдем от количества человек к количеству тысяч человек и округлим до целого числа.

Правило трёх сигм гласит, что с высокой вероятность случайная величина не отклонится от своего среднего значения более, чем на 3σ, то есть на 3 среднеквадратических отклонения. Более точно – случайная величина подчинена распределению $N(a, \sigma)$, тогда около 68% ее реализации лежат в интервале $(a - \sigma,\ a + \sigma)$, около 95% ее реализаций лежат в интервале $(a{-}2\sigma{,}a{+}2\sigma)$, а 99.7% ее реализаций лежат в интервале $(a{-}3\sigma{,}a{+}3\sigma)$. Применяя данное правило, были вычислены значения: 85.92, 95.78, 97.19 соответственно, что не соответствует правилу трёх сигм и говорит о том, что выходная величина не подчинена закону нормального распределения.

Проверим гипотезу о нормальности случайной величины с помощью критерия Пирсона. Критерий Хи-квадрат Пирсона используется для проверки гипотезы о соответствии эмпирического распределения предполагаемому. Пусть гипотеза H_0 – величина распределена нормально. Если вычисленный наблюдаемый Хи-квадрат не превышает критическое значение Хи-квадрат, то можно будет принять гипотезу H_0. В данном случае, $\chi^2_{\text{набл}}$ в 54 раза больше $\chi^2_{\text{крит}}$, следовательно, мы отвергаем гипотезу H_0 о нормальном распределении выходной величины.

В итоге, выходная величина – количество посещений ТЦ, распределена не нормально, это может ухудшить качество модели, поэтому стоит провести выравнивание для лучших результатов предсказания.

2.3 Корреляционный анализ

Корреляционный анализ используется для определения того, насколько изменения в одной переменной коррелируют с изменениями в другой. Основной инструмент в корреляционном анализе - коэффициент корреляции, чаще всего коэффициент Пирсона.

Коэффициент корреляции принимает значения от -1 до 1 и позволяет оценить характер взаимосвязи между переменными. Значение близкое к 1 указывает на положительную линейную корреляцию, тогда как значение близкое к -1 указывает на отрицательную линейную корреляцию. Коэффициент, близкий к 0, свидетельствует о слабой или отсутствующей линейной связи. Матрица корреляций показывает коэффициенты корреляции между несколькими переменными (рисунок 3).

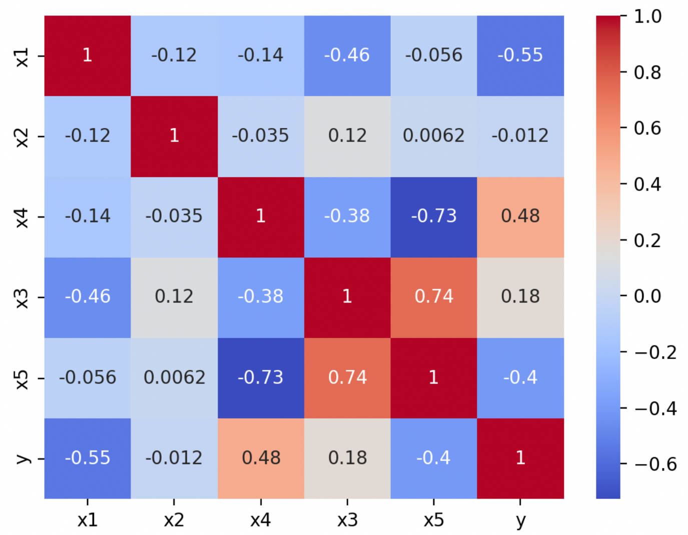

Рис. 3 – Корреляционная матрица

Можно заметить, что переменная x_2 и x_3 имеют слабую корреляцию с выходной, а остальные переменные неплохо коррелируют с выходом. Корреляция между независимыми переменными называется мультиколлинеарностью, такая связь введет к неопределенности и плохим результатам предсказания.

2.4 Построение регрессионной модели

2.4.1 Структурная идентификация модели

Зависимой переменной является количество смертей. Независимыми переменными являются 5 признаков: модель самолёта, стадия полёта, тип полёта, место падения, год производства самолёта, страна, регион мира, кол-во пассажиров на борту, кол-во персонала на борту, причина происшествия.

Рассмотрим уравнение множественной линейной регрессии $Y = \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_n \cdot X_n + \varepsilon$, где Y – зависимая переменная, $X_1 \dots X_n$ – независимые переменные, $\beta_0, \beta_1 \dots \beta_n$ - коэффициенты регрессии, ε – случайная ошибка. Данная функциональная форма отлично подойдет для рассматриваемой задачи.

2.4.2 Параметрическая идентификация модели

В соответствии с методом наименьших квадратов, задача заключается в аппроксимации кривой известной функцией. Вычисление параметров уравнения множественной линейной регрессии будет произведено с помощью алгоритма МНК.

```
omega_0: -435446.403
omega_1: -67873.406
omega_2: -1.641
omega_3: 3.988
omega_4: 0.244
omega_5: -67320.491
```

Рис. 4 – Результаты МНК

Сразу заметно низкое влияние некоторых коэффициентов. Однако на данной стадии нас интересует лишь полученное уравнение множественной линейной регрессии, которое имеет вид:

=-435446.403-67873.406X_1-1.641X_2+3.988X_3+0.244X_4-67320.491X_5

$$Y = -435446.403 - 67873.406X_1 - 1.641X_2 + 3.988X_3 + 0.244X_4 - 67320.491X_5$$

3. ИССЛЕДОВАНИЕ МОДЕЛИ

3.1 Анализ статистической значимости уравнения регрессии

Общая сумма квадратов отклонений переменной y от среднего значения $\overline{y}$ может быть разложена на две составляющие:

$$S_y = S_{\text{факт}} + S_e,$$

где $S_y = \sum_{i=1}^{n}(y_i - \overline{y})^2$ - общая сумма квадратов отклонений; $S_{\text{факт}} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$- сумма квадратов отклонений, объясненная регрессией; $S_e = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$ – остаточная сумма квадратов отклонений (необъясненная).

Выдвинем гипотезу о равенстве нулю коэффициентов регрессии. В том случае выходная переменная y не зависит от факторов, и вариация y обусловлена только воздействием ошибок: $S_y = S_e$. Противоположным является случай, при котором выходная переменная y функционально зависит от факторов: $S_y = S_{\text{факт}}$.

Для сравнения $S_{\text{факт}}$ и S_e их необходимо разделить на соответствующее число степеней свободы, получив таким образом средний квадрат отклонений на одну степень свободы – дисперсию: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$, $s_{\text{факт}}^2 = \frac{1}{m}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$, $s_e^2 = \frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$

Статистическая значимость уравнения регрессии определяется условием $s_{\text{факт}}^2 > s_e^2$. Задача сводится к проверке нулевой гипотезы $H_0: D_{\text{факт}} = D_e$ при конкурирующей гипотезе $H_1: D_{\text{факт}} > D_e$. Оценка статистической значимости уравнения регрессии выполняется с помощью F-критерия Фишера: $F = \frac{s_{\text{факт}}^2}{s_e^2}$.

Уравнение регрессии является статистически значимым, если:

1. F попадает в критическую область при заданном уровне значимости α, то есть $F > F_{\text{кр}}$
2. Уровень значимости α_F, для которого F является критической точкой (вероятность нулевой гипотезы, P-значение) меньше заданного уровня значимости α, то есть $\alpha_F < \alpha$.

Для данной модели $F = 26.966,\ \alpha_F = 1.4566126083082054e - 13$ при $F_{\text{кр}} = 2.386, \alpha = 0.05$, что удовлетворяет заданным условиям. Следовательно, текущее уравнение регрессии можно назвать статистически значимым.

3.2 Анализ статистической значимости коэффициентов уравнения регрессии

Для проверки значимости коэффициентов формулируются гипотезы: $H_0: \beta j = 0$ (коэффициент незначим), $H_1: \beta_j \neq 0$ (коэффициент значим). В качестве критерия выбирается случайная величина T_j, распределенная по закону Стьюдента с $n - m - 1$ степенями свободы: $T_j = \frac{\beta_j}{s_j}$, где β_j – коэффициент уравнения регрессии при факторе x_j, s_j – стандартная ошибка коэффициента β_j. $s_j = s\sqrt{[(C^T C)^{-1}]_{jj}}$, где $[(C^T C)^{-1}]_{jj}$ – j-й диагональный элемент матрицы $(C^T C)^{-1}$, $s = \sqrt{s_e^2}$.

Коэффициент β_j статистически значим, то есть значимо отличается от нуля (принимается гипотеза H_1 на уровне значимости α), если:

1. T_j попадает в критическую область при заданном уровне значимости α, то есть $|T\_j| > T_{\text{кр}}$;
2. Уровень значимости α_{T_j}, для которого T_j является критической точкой (P-значение) меньше заданного уровня значимости α: $\alpha_{T_j} < \alpha$.

Интервальная оценка для коэффициентов β_j определяется с помощью доверительного интервала $\left(\beta_j - t_\gamma s_j;\ \beta_j + t_\gamma s_j\right)$, где $t_\gamma = t(a, n - m - 1)$.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.354e+05	2.35e+07	-0.019	0.985	-4.75e+07	4.66e+07
x1	-6.787e+04	3.01e+04	-2.256	0.028	-1.28e+05	-7552.430
x2	-1.6406	0.890	-1.844	0.071	-3.424	0.143
x3	3.9879	0.644	6.194	0.000	2.697	5.279
x4	0.2443	3.284	0.074	0.941	-6.340	6.828
x5	-6.732e+04	1.04e+04	-6.481	0.000	-8.81e+04	-4.65e+04

Рис. 5 – Статистическая значимость коэффициентов регрессии

Работая с уровнем значимости α=0.05, заметны коэффициенты, которые являются статистически незначимыми, то есть они оказывают незначительно влияние на нашу модель. Избавление от таких коэффициентов может привести к лучшим результатам предсказания.

3.3 Исследование мультиколлинеарности факторов

Мультиколлинеарность модели множественной регрессии – наличие высокой взаимной коррелированности между факторами. Последствия мультиколлинеарности:

- Матрица $(C^T C)$ может являться невырожденной, но величина её определителя мала и, как следствие, элементы обратной матрицы становятся очень большими. В результате получаются большие дисперсии коэффициентов;

- Оценки коэффициентов чувствительны к незначительному изменению результатов наблюдений и объема выборки, что делает модель непригодной для анализа и прогнозирования;
- Уменьшаются t-статистики коэффициентов, и оценка их значимости по t-критерию теряет смысл;

Если в матрице парных коэффициентов корреляции факторов пары переменных имеют высокие коэффициенты корреляции, в модели наблюдает мультиколлинеарность. Если же факторы не коррелированы между собой, матрица парных корреляций является единичной матрицей, и ее определитель равен 1. Но если между факторами существует зависимость, то все коэффициенты корреляции равны единице, а определитель равен нулю. Следовательно, чем ближе определитель матрицы парных корреляций к нулю, тем сильнее мультиколлинеарность факторов и наоборот.

Определитель матрицы равен 0.07583822155474113, что является значением, далёким от нуля. Однако, исходя из построенной матрицы парных корреляций (рисунок 3), сильная корреляция признаков присутствует между x_5 и x_3, x_4 и x_5, а значит можно наблюдать явление мультиколлинеарности в данной модели. Стоит исключить некоторые факторы для улучшения модели.

3.4 Применение шагового регрессионного анализа для улучшения модели

Шаговый регрессионный анализ реализуется двумя способами. С помощью добавления факторов и с помощью их удаления. При добавлении определяется фактор, имеющий наиболее высокий коэффициент корреляции с выходной величиной, а после происходит пошаговое добавление остальных факторов исходя из условия увеличения скорректированного коэффициента детерминации. При удалении факторов берется модель с максимальным числом переменных, на каждом шаге проводится удаление наименее значимого фактора. Изначальный $R^2_{adj} = 0.688$.

Шаги при удалении:

1. Удаление «x_4» со P-значением 0.941 приводит к $R^2_{adj} = 0.693$
2. Удаление «x_2» со P-значением 0.065 приводит к $R^2_{adj} = 0.679$

На втором шаге и далее происходит уменьшение оценки. Следовательно, признаки, которые необходимо использовать: «x_1 », «x_3», «x_5», «x_2».

Шаги при добавлении:

1. Добавление «x_1» с приводит к $R^2_{adj} = 0.288$
2. Добавление «x_5» приводит к $R^2_{adj} = 0.472$

3. Добавление «x_3» приводит к $R^2_{adj} = 0.679$

4. Добавление «x_2» приводит к $R^2_{adj} = 0.693$

Дальнейшее добавление признаков приводит к уменьшению R^2_{adj}. Таким образом, в двух случаях была выявлена необходимость исключения фактора x_4.

4. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ РЕЗУЛЬТАТОВ МОДЕЛИРОВАНИЯ

4.1 Обоснование выбора и описание программного обеспечения

В процессе работы был использован Python 3 в качестве основного языка программирования, который может быть обоснован широкими возможностями, предоставляемыми Python: лаконичным синтаксисом, обширной стандартной библиотекой и активным сообществом разработчиков, что значительно облегчает разработку и поддержку кода.

Для обработки и анализа данных была задействована библиотека Pandas, предоставляющая удобные инструменты для работы с табличными данными. Библиотеки Scikit-learn и Statsmodels использовались для реализации статистических моделей, предоставляя разнообразные алгоритмы для анализа данных и построения моделей.

Matplotlib и Seaborn были использованы для визуализации результатов и изучения структуры данных. Гибкость и функциональность этих библиотек обеспечили создание информативных графиков, способствуя глубокому понимаю данных.

4.2 Описание основных модулей программы

Для начала были импортированы все необходимые библиотеки, которые будут использованы на протяжении всей работы.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy.stats import f
import matplotlib.pyplot as plt
from scipy.stats import chi2
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
```

Листинг 1. Импортирование библиотек.

Первичные статистические данные считываются из файлов и организуются в pandas DataFrame. Добавляется столбец времени и индексируется по времени. Отклик записывается в новую переменную, которая для простоты расчетов переведены в тысячи.

```
df = pd.DataFrame({
    "x1": np.loadtxt(path_x1, converters={0: np.int64}),
    "x2": np.loadtxt(path_x2, converters={0: np.int64}),
```

```
    "x3": np.loadtxt(path_x3, converters={0: np.int64}),
    "x5": np.loadtxt(path_x5, converters={0: np.int64}),
    "y": np.loadtxt(path_y, converters={0: np.int64})
})

# Используйте столбец с метками времени в качестве индекса
df['timestamp'] = pd.period_range(start='2018-01', end='2022-12', freq='M')
df = df.set_index('timestamp')

data = df['y'].values
```

Листинг 2. Обработка первичных данных.

Определение и вывод гистограммы интервального ряда распределения для выходной величины.

```
# Определение дискретного ряда распределения
def discrete_var(data):
    discrete = {}
    for line in data:
        discrete[line] = discrete.get(line, 0) + 1
    discrete = dict(sorted(discrete.items(), key=lambda item: item[0]))
    return discrete

def interval_var(data):
    # Число групп и длина интервалов
    discrete = discrete_var(data)
    m = int(np.ceil(1 + 3.222 * np.log10(max(data) - min(data))))
    h = int(np.ceil((max(data) - min(data)) / m))

    # Создание интервального ряда
    intervals = [(i, i + h) for i in range(min(data), max(data), h)]
    frequencies = [0] * len(intervals)

    # Распределение частот в интервалах
    for i, interval in enumerate(intervals):
        for key, value in discrete.items():
            if interval[0] <= key < interval[1]:
                frequencies[i] += value
    return intervals, frequencies, m, h

# Вывод интервального ряда распределения
print('Интервальное распределение', '\n')
intervals, frequencies, m, h = interval_var(data)
print("Интервал\tЧастота")
for i, interval in enumerate(intervals):
    print(f"[{interval[0]}, {interval[1]})\t{frequencies[i]}")

# Построение гистограммы

def hist_xi(data, m):
    plt.hist(data, bins=m, range=(min(data), max(data)), edgecolor='black')
    plt.xlabel('Тыс. человек')
    plt.ylabel('Частота')
    plt.title('Гистограмма количества посещений')
    plt.show()

hist_xi(data, m)
```

Листинг 3. Гистограммы интервального ряда распределения отклика

Были реализованы функции для проверки нормальности выходной переменной. Функция проверки правила трёх сигм выводит проценты вхождений в интервалы: одной, двух и трёх сигм.

```
# Нормальное распределение по теореме 3-х сигм
def normal_distribution_sigma(data):
    mean_x = (1 / len(data)) * sum(data)
    std_x = np.sqrt((sum([(i - mean_x) ** 2 for i in data]) / len(data)))

    sigma_68 = sum([1 if mean_x - std_x <= i <= mean_x + std_x else 0 for i in
data]) / len(data) * 100
    sigma_95 = sum([1 if mean_x - 2 * std_x <= i <= mean_x + 2 * std_x else 0
for i in data]) / len(data) * 100
    sigma_99 = sum([1 if mean_x - 3 * std_x <= i <= mean_x + 3 * std_x else 0
for i in data]) / len(data) * 100

    return sigma_68, sigma_95, sigma_99

print('Нормальное распределение по теореме 3-х сигм')

# Правило 3-х сигм
sigma_68, sigma_95, sigma_99 = normal_distribution_sigma(data)
print(f"Процент вхождений в интервал 1 сигмы: {sigma_68}")
print(f"Процент вхождений в интервал 2 сигм: {sigma_95}")
print(f"Процент вхождений в интервал 3 сигм: {sigma_99}")

if (sigma_68 > 68) and (sigma_95 > 95) and (float(sigma_99) > 99.7):
    print('Распределение нормальное')
else:
    print('Распределение не нормальное')
```

Листинг 4. Функция правила трёх сигм

Функция проверки нормальности распределения с помощью критерия Пирсона использует вспомогательную функцию, делящую данные на интервалы.

```
# Нормальное распределение по критерию Пирсона
print('Нормальное распределение по критерию Пирсона')

def normal_distribution_pearson(data):
    n = (sum(frequencies))
    xi = [(left + right) / 2 for left, right in intervals]
    def mean_xi(intervals, frequencies):
        return (1 / sum(frequencies)) * sum(
            [((left + right) / 2) * frequencies[i] for i, (left, right) in
enumerate(intervals)])
    def var_xi(intervals, frequencies):
        return (1 / sum(frequencies)) * sum(
            [(((left + right) / 2) - mean_xi(intervals, frequencies)) ** 2 *
frequencies[i] for i, (left, right) in
             enumerate(intervals)])
    def std_xi(intervals, hist_data):
        return np.sqrt(var_xi(intervals, hist_data))
    ui = [(i - mean_xi(intervals, frequencies)) / std_xi(intervals, frequencies)
for i in xi]
    def laplace_xi(x):
```

```
        return np.exp(- (x ** 2 / 2)) / (np.sqrt(2 * np.pi))
    f_ui = [laplace_xi(i) for i in ui]
    ni_teor = [(h * n / std_xi(intervals, frequencies)) * i for i in f_ui]
    chi2_obs = sum([(frequencies[i] - j) ** 2 / j for i, j in
enumerate(ni_teor)])
    r = 2
    chi2_crit = chi2.ppf(1 - 0.05, m - r - 1)
    return chi2_obs, chi2_crit, chi2_obs / chi2_crit

chi2_obs, chi2_crit, chi2_ratio = normal_distribution_pearson(data)
print(f'Хи-квадрат наблюдаемое {chi2_obs}')
print(f'Хи-квадрат критическое {chi2_crit}')
print(f'Рассчитанное значение {chi2_obs / chi2_crit}')
if chi2_obs > chi2_crit:
    print("Отвергаем H0, распределение не является нормальным")
else:
    print("Принимаем H0, распределение является нормальным")
print()
```

Листинг 5. Критерий хи-квадрат Пирсона

Функция, реализующая корреляционный анализ, выводит на экран матрицу парных корреляций и её детерминант.

```
# Составим матрицу парных корреляций
def corr_analysis(df):
    sns.heatmap(df.corr(), annot=True, cmap="coolwarm", annot_kws={"size": 10})
    plt.show()
    print(f'Детерминант матрицы парных корреляций:
{np.linalg.det(df.corr().to_numpy())}')

corr_matrix = df.corr()
```

Листинг 6. Функция корреляционного анализа

Метод наименьших квадратов был реализован с помощью функции OLS из библиотеки Statsmodels. Данная модель была обучена на тренировочных данных, после чего получен массив y_pred – предсказания для тестовых данных.

```
# Составим матрицу признаков и вектор ответов
X = df.drop('y', axis=1)
y = df['y']

# Построение регрессионной модели с помощью statsmodels

# Разделение данных на тренировочные и тестовые
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
random_state=42)

# Добавление константы (интерцепта) к матрице признаков для МНК
X_train_mnk = sm.add_constant(X_train)
X_test_mnk = sm.add_constant(X_test)

# Реализация МНК для тренировочных данных
model = sm.OLS(y_train, X_train_mnk).fit()
```

```
# Получение прогнозов для тестовых данных
y_pred = model.predict(X_test_mnk)
```

Листинг 7. Программная реализация МНК

Проверка значимости уравнения регрессии с помощью критерия Фишера.

```
# Тест Фишера
def fisher_test(y, X, model, n, m):
    # Общая сумма квадратов отклонений
    S2_y = np.sum((y - np.mean(y)) ** 2) / (n - 1)
    # Сумма квадратов отклонений объясненной моделью
    S2_fact = np.sum((model.predict(X) - np.mean(y)) ** 2) / m
    # Сумма квадратов отклонений по остаткам
    S2_e = np.sum((y - model.predict(X)) ** 2) / (n - m - 1)

    # F-статистика
    F_statistic = S2_fact / S2_e

    # Критическое значение для alpha=0.05
    alpha = 0.05
    critical_value = f.ppf(1 - alpha, m, n - m - 1)

    # P-значение
    p_value = 1 - f.cdf(F_statistic, m - 1, n - m)

    return F_statistic, critical_value, p_value

F_statistic, critical_value, p_value = fisher_test(y_test, X_test_mnk, model,
len(y_test), X_test_mnk.shape[1] - 1)

print(f"F-критерий: {F_statistic:.3f}")
print(f"Критическое значение: {critical_value:.3f}")
print(f"P-значение: {p_value}")
```

Листинг 8. Критерий Фишера

Проверка значимости коэффициентов уравнения регрессии.

```
# Вывод результатов регрессии
result_summary = model.summary()
coefficients_table = pd.DataFrame(result_summary.tables[1].data[1:],
columns=result_summary.tables[1].data[0])

# Заменяем имена столбцов
coefficients_table.columns = [' ', 'coef', 'std err', 't', 'P>|t|', '[0.025',
'0.975]']
print('Анализ статистической значимости коэффициентов уравнения регрессии:')
print(coefficients_table.to_string(index=False))
```

Листинг 9. Значимость коэффициентов регрессии

Для оценки качества модели была написана функция, которая выводит на экран абсолютную ошибку среднего, среднеквадратичную ошибку, коэффициент детерминации и его исправленную версию.

```
# Оценка качества модели на тестовых данных
def model_evaluation(y_test, y_pred):
    mse = mean_squared_error(y_test, y_pred)
    print(f"Среднеквадратичная ошибка на тестовых данных: {mse:.4f}")
    mae = mean_absolute_error(y_test, y_pred)
    print(f"Средняя абсолютная ошибка на тестовых данных: {mae:.4f}")
    R_2 = model.rsquared
    print(f"Коэффициент детерминации: {R_2:.3f}")
    R_2_adj = 1 - (n - 1) / (n - m - 1) * (1 - R_2)
    print(f"Адаптивный коэффициент детерминации: {R_2_adj:.3f}")
model_evaluation(y_test, y_pred)
```

Листинг 10. Функция вывода оценок

4.3 Численное исследование результатов моделирования

Оценка модели была произведена по 4 характеристикам, оценивающим качество предсказаний и модели в целом:

- MSE (Mean Squared Error) – средняя квадратичная ошибка. Измерение среднего квадрата разности между предсказанными и фактическими значениями.
- MAE (Mean Absolute Error) – средняя абсолютная ошибка. Измерение среднего значения абсолютных разностей между предсказанными и фактическими значениями.
- R^2 score – коэффициент детерминации. Измеряет долю дисперсии зависимой переменной, которая может быть объяснена моделью. Принимает значения от 0 до 1, где 1 означает идеальное предсказание.
- Adjusted R^2 score – скорректированный коэффициент детерминации, учитывающий количество предикторов в модели и корректирующий $R2$ score в случае наличия избыточных предикторов.

```
Среднеквадратичная ошибка на тестовых данных (MSE): 392545383125.6406
Средняя абсолютная ошибка на тестовых данных (MAE): 530223.0236
Коэффициент детерминации: 0.777
Адаптивный коэффициент детерминации: 0.760
```

Рис. 6 – Характеристики модели

Коэффициент детерминации нормальный, это свидетельствует о не плохом качестве предсказаний. MSE принимает высокое значение. MAE примерно равно 530223, в контексте рассматриваемой задачи это значит, что предсказанной значение потенциального количества посетителей ТРЦ может отличаться от истинного значения на ± 530223.

Пусть коэффициенты детерминации не плохие, значения MSE и MAE достаточно высоки, поэтому следует произвести улучшение модели, чтобы добиться приемлемого качества предсказания.

ВЫВОДЫ

В ходе проведенного исследования были выявлены статистические зависимости между количеством посетителей торгового центра и рядом факторов. Анализ характеристик объекта исследования позволил определить ключевые переменные, оказывающие влияние на исследуемый показатель. В результате формализации и классификации переменных была проверена гипотеза о нормальном распределении выходной величины.

Корреляционный анализ подтвердил наличие статистически значимых связей между переменными, а построение регрессионной модели дало возможность выявить структурные и параметрические характеристики влияющих факторов.

Исследование модели подтвердило статистическую значимость уравнения регрессии, а также позволило провести анализ статистической значимости коэффициентов уравнения. Применение шагового регрессионного анализа способствовало улучшению модели, оптимизации коэффициентов и исключению мультиколлинеарности факторов.

В ходе программной реализации и численного исследования результатов моделирования было обосновано выбор программного обеспечения, представлено описание основных модулей программы и проведено численное исследование результатов.

Разработанная модель прогнозирования посещаемости торгового центра представляет собой достаточно эффективный инструмент для предсказания будущих тенденций. Результаты исследования помогут в принятии управленческих решений в области торговли.

СПИСОК ЛИТЕРАТУРЫ

1. Бослаф С. Статистка для всех. М.: ДМК Пресс, 2015
2. Линник Ю.В. Метод наименьших квадратов и основы теории обработки наблюдений. М.: Государственное издательство физико-математической литературы, 1962
3. Европейский торговый центр. (н.д.). О компании. URL: https://europe-tc.ru/about/ (дата обращения: 18.12.2023)
4. Гарант.Ру. (н.д.). Календарь бухгалтера и юриста. URL: https://www.garant.ru/calendar/buhpravo/ (дата обращения: 20.12.2023)
5. Федеральная служба государственной статистики. (н.д.). Розничная торговля. URL: https://rosstat.gov.ru/statistics/roznichnayatorgovlya (дата обращения: 20.12.2023)
6. Федеральная служба государственной статистики. (н.д.). Рынок труда, занятость и заработная плата. URL: https://rosstat.gov.ru/labor_market_employment_salaries (дата обращения: 20.12.2023)
7. Федеральная служба государственной статистики. (н.д.). Рабочая сила. URL: https://rosstat.gov.ru/labour_force (дата обращения: 20.12.2023)

ПРОГРАММНЫЕ ПРИЛОЖЕНИЯ

Программной реализации алгоритма:

URL: https://github.com/darcysoul/kr_psa_5sem

ПРИЛОЖЕНИЕ А

Статистика посещений ТРЦ «Европейский» с 2017 по 2023 года.

	Посещаемость (чел.)				
месяц\год	2018	2019	2020	2021	2022
январь	5 920 777	5 268 603	5 122 083	4 259 785	4 336 041
февраль	5 107 023	4 630 542	5 025 002	4 272 518	4 360 325
март	5 260 188	5 251 521	3 869 908	4 647 206	4 634 183
апрель	5 241 923	5 394 705	217 114	4 651 527	3 958 694
май	5 513 691	4 714 597	258 315	4 201 100	3 963 504
июнь	5 908 677	5 490 355	2 629 653	4 193 056	4 096 926
июль	5 323 940	4 641 573	3 970 811	4 217 312	4 322 628
август	5 939 614	4 793 950	4 399 658	4 471 314	4 192 339
сентябрь	4 861 658	4 936 602	3 806 806	4 233 799	3 807 424
октябрь	4 632 388	5 558 193	4 723 175	4 269 717	3 831 074
ноябрь	5 334 797	5 718 277	4 688 987	4 036 141	4 454 125
декабрь	7 283 867	6 077 057	4 983 518	5 001 868	4 953 021

ПРИЛОЖЕНИЕ Б

Статистические характеристики данных о посещении ТРЦ.

Код. URL: https://replit.com/@katarix/kursach-punkt-1

Год	Посещения, чел.
2018	66328543
2019	62475975
2020	43695030
2021	52455343
2022	50910284
2023	48972379

Год	Посещения, чел.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2018	66328543	nan	nan	nan
2019	62475975	-3.85257e+06	94.2	-5.8
2020	43695030	-2.26335e+07	65.9	-34.1
2021	52455343	-1.38732e+07	79.1	-20.9
2022	50910284	-1.54183e+07	76.8	-23.2
2023	48972379	-1.73562e+07	73.8	-26.2

Год	Посещения, чел.	Абсолютный прирост	Темп роста, %	Темп прироста, %
2018	66328543	nan	nan	nan
2019	62475975	-3.85257e+06	94.2	-5.8
2020	43695030	-1.87809e+07	69.9	-30.1
2021	52455343	8.76031e+06	120	20
2022	50910284	-1.54506e+06	97.1	-2.9
2023	48972379	-1.9379e+06	96.2	-3.8

Средний уровень ряда: 55173035 (чел.)

Средний абсолютный прирост: -3854564 (чел.)

Средний темп роста: 93.6%

Средний темп прироста: -6.4%

Предсказание по среднему абсолютному приросту: 47055720 (чел.)

Предсказание по среднему темпу роста: 47652072 (чел.)

Уравнение прямой: y = -4085715.0 * t + 67430180.0

Предсказание методом аналитического выравнивания: 42915890 (чел.)

Относительная погрешность по среднему абсолют. приросту: 3.91

Относительная погрешность по среднему темпу роста: 2.7

Относительная погрешность по аналитическому выравниванию МНК: 12.37

ПРИЛОЖЕНИЕ В

Первичные статистические данные о факторах.

	Год	x_1	x_2	x_3	x_4	x_5
2018	январь	14	70251	1202292	7158110	1
2018	февраль	9	80184	1165811	7158110	2
2018	март	11	84082	1278051	7158110	3
2018	апрель	9	89318	1274769	7158110	4
2018	май	11	81064	1309579	7158110	5
2018	июнь	10	90094	1349531	7158110	6
2018	июль	9	80999	1396384	7158110	7
2018	август	8	77618	1469262	7158110	8
2018	сентябрь	10	77274	1452049	7158110	9
2018	октябрь	8	791506	1453254	7158110	10
2018	ноябрь	9	78947	1453568	7158110	11
2018	декабрь	10	113989	1719395	7158110	12
2019	январь	14	79681	1291648	7196190	13
2019	февраль	8	85370	1259407	7196190	14
2019	март	11	95179	1379374	7196190	15
2019	апрель	8	102908	1368527	7196190	16
2019	май	13	89045	1383874	7196190	17
2019	июнь	11	96030	1424197	7196190	18
2019	июль	8	91608	1468437	7196190	19
2019	август	9	86734	1540480	7196190	20
2019	сентябрь	9	86685	1514309	7196190	21
2019	октябрь	8	89129	1530861	7196190	22
2019	ноябрь	8	88657	1542425	7196190	23
2019	декабрь	9	135375	1800000	7196190	24
2020	январь	14	88845	1367384	7110205	25
2020	февраль	10	92390	1356534	7110205	26
2020	март	12	105238	1522431	7110205	27
2020	апрель	30	101552	914977	7110205	28
2020	май	17	91824	1041553	7110205	29
2020	июнь	10	98700	1340047	7110205	30
2020	июль	9	98930	1511632	7110205	31
2020	август	10	90304	1608996	7110205	32
2020	сентябрь	8	93062	1574094	7110205	33
2020	октябрь	9	94065	1606022	7110205	34
2020	ноябрь	10	95315	1588013	7110205	35

2020	декабрь	8	153648	1854811	7110205	36
2021	январь	16	93059	1476753	7110205	37
2021	февраль	9	104451	1448633	7139468	38
2021	март	9	116355	1622265	7139468	39
2021	апрель	8	117769	1653338	7139468	40
2021	май	16	105247	1678827	7139468	41
2021	июнь	9	109307	1701207	7139468	42
2021	июль	9	108520	1765410	7139468	43
2021	август	9	98974	1877278	7139468	44
2021	сентябрь	8	102233	1849605	7139468	45
2021	октябрь	10	103394	1847684	7139468	46
2021	ноябрь	10	104878	1805998	7139468	47
2021	декабрь	9	172553	2190795	7139468	48
2022	январь	15	103124	1728245	7088432	49
2022	февраль	9	114701	1751505	7088432	50
2022	март	9	146044	1981923	7088432	51
2022	апрель	9	120502	1666472	7088432	52
2022	май	13	113671	1676048	7088432	53
2022	июнь	9	123689	1700349	7088432	54
2022	июль	10	115294	1751405	7088432	55
2022	август	8	109060	1847690	7088432	56
2022	сентябрь	8	113896	1773711	7088432	57
2022	октябрь	10	113163	1789380	7088432	58
2022	ноябрь	9	113723	1806308	7088432	59
2022	декабрь	9	185646	2070194	7088432	60