Сумленинова Полина Денисовна АДПУР 7.1

Задание 1. Реализация линейной множественной регрессии

Цель работы: практическое применение линейной множественной регрессии для факторного анализа и оценки влияния ключевых переменных на объём продаж косметических товаров. Необходимо построить математическую модель зависимости числа проданных единиц продукции Y от выбранных факторов; определение значимых факторов, оказывающих статистически обоснованное влияние на продажи; сравнение моделей с полным и сокращённым набором факторов по качественным метрикам (R^2, MSE, системный эффект); проверка соответствия финальной модели условиям Гаусса–Маркова для обеспечения корректности оценок.

Описание датасета

Датасет Supply Chain Analysis был взят с сайта Kaggle (https://www.kaggle.com/datasets/harshsingh2209/supply-chain-analysis/data), он описывает цепочку поставок и продаж косметической продукции для стартапа Fashion&Beauty. В исходном датасете представлены данные:

Product Type, SKU, Price, Availability, Number of products sold, Revenue generated, Customer demographics, Stock levels, Lead times, Order quantities, Shipping times, Shipping costs, Supplier name, Location, Manufacturing lead time, Production volumes, Manufacturing costs, Inspection results, Defect rates, Transportation modes, Routes, Costs.

Для модели регрессии были отобраны четыре ключевые числовые переменные. Для зависимой переменной (Y) был

выбран фактор ***Number of products sold*** (количество проданных единиц товара), а для независимых факторов (X) ***Price*** (цена за единицу продукции, X1), ***Order quantities*** (объём заказанных товаров поставщикам, X2), ***Production volumes*** (объём произведённой продукции, X3).

	Product type	SKU	Price	Availability	Number of products sold	\
0	haircare	SKU0	69.808006	55	802	
1	skincare	SKU1	14.843523	95	736	
2	haircare	SKU2	11.319683	34	8	
3	skincare	SKU3	61.163343	68	83	
4	skincare	SKU4	4.805496	26	871	

	Revenue generated	Customer demographics	Stock levels	Lead times	\
0	8661.996792	Non-binary	58	7	
1	7460.900065	Female	53	30	
2	9577.749626	Unknown	1	10	
3	7766.836426	Non-binary	23	13	
4	2686.505152	Non-binary	5	3	

	Order quantities	...	Location Lead time	Production volumes	\
0	96	...	Mumbai 29	215	
1	37	...	Mumbai 23	517	
2	88	...	Mumbai 12	971	
3	59	...	Kolkata 24	937	
4	56	...	Delhi 5	414	

	Manufacturing lead time	Manufacturing costs	Inspection results	\
0	29	46.279879	Pending	
1	30	33.616769	Pending	
2	27	30.688019	Pending	
3	18	35.624741	Fail	
4	3	92.065161	Fail	

	Defect rates	Transportation modes	Routes	Costs
0	0.226410	Road	Route B	187.752075
1	4.854068	Road	Route B	503.065579
2	4.580593	Air	Route C	141.920282
3	4.746649	Rail	Route A	254.776159
4	3.145580	Air	Route A	923.440632

Описание обработки данных

Анализ данных был произведен в Google Colab на Python, для импорта всех данных были загружены библиотеки pandas (чтение табличных данных формата csv), matplotlib.pyplot и seaborn для визуализации данных ,

numpy (работа с массивами), scikit-learn (LinearRegression, mean_squared_error, r2_score) построение модели OLS и расчет основных метрик (R^2, MSE), statsmodels.api расширенное статистическое моделирование (OLS с подробным выводом, добавление константного столбца), statsmodels.stats.outliers_influence (variance_inflation_factor) для расчета меры мультиколлинеарности, scipy.stats для статистических тестов (тест нормальности, t-тест, F-критерии, z-тест поворотных точек), statsmodels.stats.stattools (durbin_watson) критерий Дарбина–Уотсона.

Для упрощения работы колонки были приведены к коротким меткам (Number of products sold → Y; Price → X1_Price, Order quantities → X2_OrdQty, Production volumes → X3_Production). Далее, были очищены пропуски в столбцах, а также были удалены данные с некорректным форматом.

Коэффициенты первой линейной регрессии и значения указанных показателей

Одним из первых шагов анализа была построена модель первой линейной регрессии, и были найдены значения коэффициента детерминации, среднеквадратичную ошибку, системный эффект факторов.

Коэффициент детерминации (R^2) = 0.0371 (близок к 0) показывает, что вероятно основная часть вариации продаж обусловлена другими причинами (вероятно другими факторами), то есть связь между количеством проданных единиц товара и ценой, объемом продукции и количеством заказов довольно слабая.

Среднеквадратичная ошибка MSE = 87971.4570 показывает средний квадрат отклонения предсказанных значений от фактических, этот фактор большой, так как предсказания

модели могут существенно отклоняться от реальных показателей.

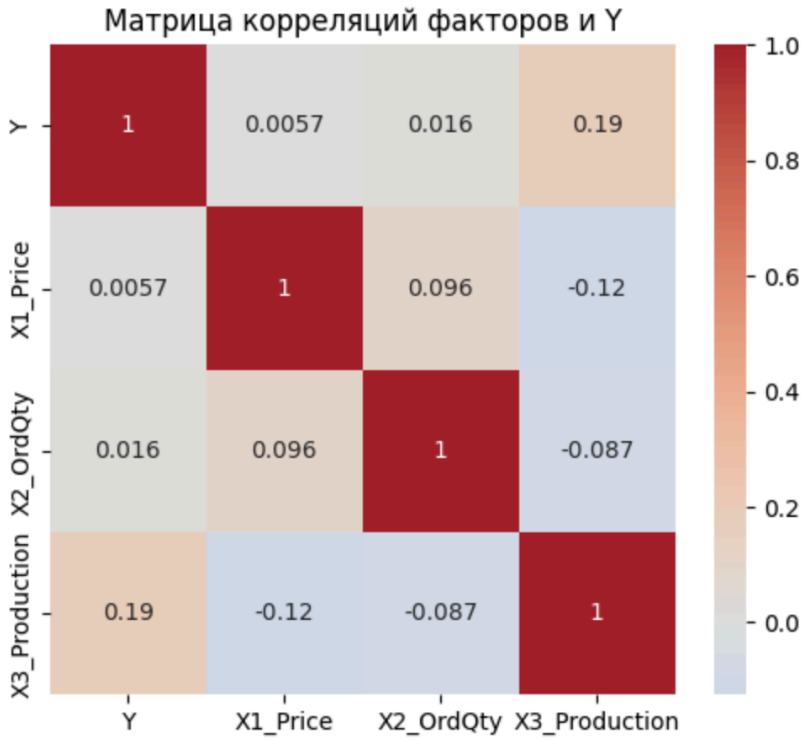
Системный эффект факторов = 0.0385 показывает отношение объяснённой дисперсии к необъяснённой, в датасете это только показывает, что вклад выбранных факторов на количество проданных товаров невелик.

Анализ мультиколлинеарности

Все факторы (X1_Price = 1.02 ; X2_OrdQty = 1.02; X3_Production = 1.02) имеют VIF (мера мультиколлинеарности) < 5, что указывает на отсутствие мультиколлинеарности, то есть факторы практически независимы друг от друга.

Корреляционный анализ

Корреляционная матрица показывает, насколько линейно связаны между собой переменные и зависимая величина.



Корреляция Y с X3_Production (0.1879) выше, чем с остальными факторами, но всё ещё слабая (<0.2). Между независимыми переменными корреляции также низкие (|0.1|), что подтверждает отсутствие сильной мультиколлинеарности.

Далее были найдены p-value для каждого независимого фактора (Xi): p-value (X1_Price) ≈ 0.7904; p-value (X2_OrdQty) ≈0.7654; p-value (X3_Production) ≈ 0.0584

При уровне значимости 0.05 ни один фактор не оказался строго значимым (p<0.05). Тем не менее, Production имеет наименьшее p-value (0.0584), следовательно это наиболее информативный фактор. Вторым по «информативности» идёт OrderQty (p=0.7654 меньше, чем у Price).

Построение новой модели и сравнение показателей

Далее, была написана новая модель линейной регрессии, а также рассчитаны её показатели. Эта модель оставила только два наиболее «значимых» фактора по p-value (X2_OrdQty и X3_Production).

Тогда R^2(sig) = 0.036, MSE(sig)=88036.561, а разница между показателями первой и второй модели: ΔR^2=-0.001, ΔMSE=65.104. По разнице в коэффициенте детерминации можно увидеть, что фактор цены несильно сказывается на общем объеме продаж, а разница среднеквадратичной ошибки указывает также на то, что в контексте датасета средняя ошибка увеличилась не намного, поэтому она мала, и также особо не влияет на прогноз.

После построения новой модели можно увидеть, что существенной разницы между результатными показателями нет, но при этом сам датасет стал чуть компактнее.

F-тест для проверки целесообразности (критерий Фишера)

Необходимо проверить гипотезу о том, что новая модель (без Price) не хуже полной. Для этого были сформулированы нулевая и альтернативная гипотезы:

H(0): первая модель даёт статистически значимо лучшее приближение, чем новая

H(1): новая модель тоже является показательной

Результаты:
F-статистика = 0.0718
F-критическое (α=0.05) = 3.9391
p-value = 0.7893
Исключение факторов целесообразно.

Если F<F-критическое или p>α, то мы не отвергаем H1, соответственно новая модель тоже является показательной. Поскольку 0.0718<3.9391 и p=0.7893>0.05, исключение фактора X1 целесообразно. Его вклад в точность модели статистически незначим, и качество прогноза в новой модели не хуже, чем у первой модели.

Проверка модели на соответствие условиям Гаусса-Маркова

Результат датасета:
Поворотные точки: 60, Z=-1.277, p=0.202
Асимметрия=0.155, Эксцесс=-1.211, тест нормальности p=0.000
Дарбин-Уотсон=2.066
t=-0.0000, p=1.0000

Во-первых, p>0.05 означает, что нет оснований отклонять

гипотезу о случайности последовательности остатков. Распределение данных отличается от нормального, что может повлиять на точность доверительных интервалов и тестов значимости. По критерию Дарбина-Уотсона DW=2.066, где значение близко к 2 и указывает на то, что данные независимы.

Остаточные данные не соответствуют условиям Гаусса–Маркова по независимости и равенству математического ожидания нулю. Однако данные не нормальны, что нарушает одно из дополнительных предположений (нормальности ошибок). Это может усложнить использование точных доверительных интервалов.

Выводы

В ходе работы была проведена множественная линейная регрессия для анализа влияния ценовых, закупочных и производственных факторов на объём продаж косметической продукции. Первая модель с тремя факторами объясняла лишь 3,7% (0,037) дисперсии продаж, что указывает на необходимость поиска дополнительных переменных. Мультиколлинеарность отсутствует, факторы независимы и не искажают оценки коэффициентов. Объем производства и количество заказов оказались более информативными признаками, исключение цены в новой модели оказалось статистически оправданным.

В дальнейшем при рассмотрении датасета можно включить в модель дополнительные релевантные переменные (например, демография, возраст и тп).