Регрессионный анализ

Задача.

Используя данные датасета, размещенного по адресу <https://www.kaggle.com/datasets/khalidryder777/500k-chatgpt-tweets-jan-mar-2023>, реализовать задачу нахождения уровня влияния факторов на успешность аккаунта. Критерием успешности является количество лайков.

Необходимо подготовить данные используя функции библиотеки pandas - dropna() или fillna(),перевести все тестовые данные в единый регистр для однообразия данных и числовой формат, убрать выбросы, используя межквартильный размах или Z-оценку. Поделить на тестовую и обучающие выборки.

Используя матрицу корреляции (pandas — corr()) составить таблицу отбора факторов для включения в модель. Реализовать тепловую карту корреляций.

Выбрать один из методов - множественная линейная регрессия или полиномиальная регрессия

Оценить качество полученной модели - средняя абсолютная ошибка (MAE), средняя абсолютная ошибка в процентах (MAPE), коэффициент детерминации R-квадрат.

Оформить выводы.

Теоретические материалы:

Регрессионный анализ позволяет найти функциональную зависимость между переменными, он основан на предположении что зависимая переменная определяется функцией от независимых переменных (факторов) и параметров. Его целью является получение такой зависимости, которая бы наилучшим образом описывала исходные данные.

Выбросы — это значения, которые сильно отличаются от остальных значений в столбце и могут исказить результаты анализа. Они могут возникать из-за ошибок при сборе данных или быть результатом редких событий. Чтобы избежать искажений в данных, выбросы нужно удалить или заменить на более типичные значения. Очистка данных от выбросов позволяет получить более точные результаты анализа и принимать более обоснованные решения на основе этих данных. Есть два распространённых способа идентифицировать и удалить выбросы из датасета: межквартальный диапазон и использование Z значений.

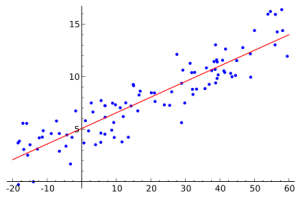
Межквартильный размах — представляет собой разницу между 75-м и 25-м процентилями в наборе данных и измеряет разброс средних 50% значений. Если значение превышает межквартильный размах более чем в 1,5 раза за пределами третьего квартиля (Q3) или первого квартиля (Q1), то оно может быть определено как выброс.

Z-оценка показывает, на сколько стандартных отклонений данное значение отклоняется от среднего значения. Для расчета используется формула . Очистка данных от выбросов позволяет получить более точные результаты анализа и принимать обоснованные решения на основе этих данных.

Множественная регрессия — это статистический метод, который используется для анализа отношений между зависимой переменной и набором независимых переменных. Она позволяет определить, совокупное влияние факторов на результат.

Множественная линейная регрессия — модель, которая описывает линейную зависимость между зависимой переменной и набором независимых переменных. На вход данной модели передается линейная комбинация факторов. Формула множественной линейной регрессии

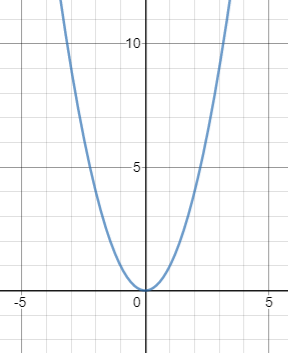
График парной линейной регрессии



Полиномиальная регрессия — модель, которая описывает нелинейную зависимость между зависимой переменной и набором независимых переменных.

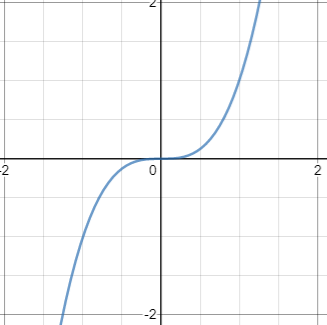
Формула полинома второго порядка для одного фактора

График полинома второго порядка для одного фактора



Формула полиномиальной регрессии третьего порядка для одного фактора имеет вид, представленный в формуле общего вида .

Изображается на графике для одного фактора



[Средняя абсолютная ошибка](https://wiki.loginom.ru/articles/mae.html) (MAE). является линейной оценкой, в которой все ошибки взвешены одинокого. Рассчитывается по формуле

Средняя абсолютная ошибка в процентах (MAPE). Данная ошибка проста в интерпретации и не имеет размерности. Рассчитывается по формуле

[Коэффициент детерминации](https://wiki.loginom.ru/articles/coefficient-of-determination.html)R-квадрат. Показывает долю дисперсии зависимой переменной, объясненной с помощью регрессионной модели. Она позволяет давать оценку качества не значений ошибок, а качества работы модели, что обеспечивает хорошую интерпретацию. Кроме того, большим преимуществом коэффициента детерминации является его инвариантность к масштабу данных. Наиболее общей формулой для вычисления коэффициента детерминации является

Библиотеки Python:

NumPy — библиотека для работы с многомерными массивами и матрицами;

Pandas — библиотека для работы с табличными данными, включая чтение и запись данных из различных форматов;

Matplotlib — библиотека для визуализации данных, включая построение графиков, диаграмм и др.;

Datetime — библиотека, которая предоставляет удобные инструменты для работы с датами и временем;

Seaborn — библиотека, которая предоставляет мощные инструменты для визуализации данных;