Slide2:
Linear regression(Линейные модели):
Добрый день, дорогие коллеги! Сегодня я хотел бы рассказать вам о линейной регрессии - одном из наиболее распространенных методов анализа данных.

Линейная регрессия - это статистический метод, который используется для предсказания числовой зависимости между двумя переменными. Этот метод очень популярен в машинном обучении, статистике и других областях, где требуется анализировать данные и делать предсказания на их основе.

Slide3:
Принцип работы линейной регрессии:

Теперь давайте поговорим о том, как работает линейная регрессия. Прежде всего, линейная регрессия - это модель, которая основывается на линейной зависимости между двумя переменными. Она использует данные, чтобы определить наилучшую прямую линию, которая наиболее точно предсказывает зависимость между двумя переменными. Эта линия называется линией регрессии.

Slide4:
Формула для линейной регрессии включает в себя независимую переменную (x), зависимую переменную (y), коэффициенты регрессии (β_0, β) и остаточную ошибку (ksi). Коэффициенты регрессии определяются на основе данных и используются для создания линии регрессии.

Intercept– свободный коэффициент; то, чему равна зависимая переменная, если предиктор равен нулю
Slope– угол наклона прямой; показывает, насколько изменяется зависимая переменная при изменении предиктора

Slide5:
Примеры использования линейной регрессии включают в себя предсказание продаж в зависимости от рекламы, определение влияния года выпуска на стоимость автомобиля, предсказание температуры воздуха на основе времени года и т.д

Чтобы использовать линейную регрессию, необходимо иметь данные, которые можно использовать для определения зависимости между двумя переменными. Для линейной регрессии необходимо иметь по крайней мере две переменные, одну независимую и одну зависимую.

Перед использованием данных необходимо их обработать. Это включает в себя удаление выбросов, нормализацию данных и другие методы обработки данных. Обработка данных позволяет улучшить точность и надежность модели линейной регрессии.

Slide 6:

- Cost function (функция потерь) - это функция, которая измеряет, насколько хорошо модель соответствует данным. В линейной регрессии наиболее распространенной функцией стоимости является среднеквадратическая ошибка (MSE), которая вычисляет среднеквадратическую разницу между прогнозами модели и фактическими значениями. Цель оптимизации заключается в том, чтобы минимизировать функцию потерь, что достигается путем настройки параметров модели.

- Gradient decent(Градиентный спуск):
  это метод оптимизации, используемый для нахождения минимума функции потерь в модели линейной регрессии. Он работает путем итеративного обновления коэффициентов модели, чтобы минимизировать ошибку модели. Процесс начинается с начальных значений коэффициентов, а затем происходит последовательное обновление коэффициентов на каждой итерации с учетом градиента функции потерь.

Slide7:
Regularsation:
L1 и L2 регуляризация - это методы добавления штрафа за сложность модели в целевую функцию, чтобы уменьшить переобучение. Они отличаются тем, каким образом штраф применяется к параметрам модели.

L1 регуляризация, также известная как Lasso, добавляет абсолютное значение весов параметров модели к целевой функции. Это приводит к реже выбору признаков, поскольку L1 регуляризация может установить некоторые веса параметров в ноль, что исключает соответствующий признак из модели.

L2 регуляризация, также известная как Ridge, добавляет квадрат весов параметров модели к целевой функции. Это приводит к сглаживанию весов параметров и предотвращению переобучения, что может быть особенно полезно при работе с множеством признаков с высокой корреляцией.
Таким образом, L1 и L2 регуляризация имеют разные эффекты на параметры модели и могут использоваться в зависимости от конкретной задачи и типа данных.

Slide8:
Multiple LR (Модель множественной линейной регрессии):
Multiple Linear Regression (MLR) - это модель линейной регрессии, которая учитывает влияние нескольких независимых переменных на зависимую переменную. MLR используется в случаях, когда зависимость между зависимой переменной и одной независимой переменной недостаточно для описания сложных процессов. В MLR каждая независимая переменная имеет свой собственный коэффициент, который определяет, насколько она влияет на зависимую переменную. Обычно перед использованием MLR данные проходят через предварительную обработку, чтобы убедиться в отсутствии

выбросов, пропущенных значений и мультиколлинеарности.

Slide9:
Для оценки качества модели линейной регрессии используются различные метрики, такие как коэффициент детерминации, коэффициент корреляции, остаточная сумма квадратов и другие. Эти метрики позволяют определить, насколько точно модель линейной регрессии предсказывает зависимость между двумя переменными.

Slide10:

Преимущества:

- Простота: линейная регрессия является простым и понятным методом анализа данных, который может быть использован даже без специальных знаний в области статистики.

- Интерпретируемость: результаты линейной регрессии могут быть легко интерпретированы и объяснены, что является важным для принятия решений.

- Эффективность: линейная регрессия является быстрым и эффективным методом анализа больших объемов данных.

Как и любая модель, линейная регрессия имеет свои ограничения и недостатки. Некоторые из них:

1- Линейность: модель линейной регрессии основывается на предположении, что отношения между зависимой и независимыми переменными являются линейными. Если связь между переменными не является линейной, то модель может давать неточные прогнозы.

2- Предположения о распределении: линейная регрессия предполагает, что ошибки распределены нормально и одинаково. Если данные не соответствуют этим предположениям, то модель может быть неточной.

3- Выбросы: линейная регрессия чувствительна к выбросам в данных. Одиночные аномальные значения могут исказить результаты и привести к неточным прогнозам.

4- Мультиколлинеарность: если независимые переменные сильно коррелируют между собой, это может привести к проблемам в моделировании. В таких случаях модель может быть неустойчивой и приводить к неточным результатам.

5- Ограниченность в предсказаниях: линейная регрессия может быть неэффективной в предсказании значений вне диапазона значений, используемых для обучения модели(outliers).

Тем не менее, несмотря на эти ограничения, линейная регрессия остается одной из самых широко используемых моделей в машинном обучении и научных исследованиях благодаря своей простоте, понятности и способности обобщать данные.

Slide11:
Заключение:
В итоге, линейная регрессия - это мощный инструмент для анализа и прогнозирования данных. Огромное количество обычных бизнес-задач хорошо решается линейными моделями. В частности, банки и страховые организации активно ими пользуются уже много-много лет. Линейная регрессия позволяет нам определить, как связаны различные переменные и как изменения в одной переменной влияют на другие переменные. Однако у нее есть свои ограничения, например, она не способна обрабатывать нелинейные зависимости между переменными и может быть чувствительна к выбросам и пропущенным данным. Чтобы улучшить производительность линейной регрессии, мы можем использовать методы, такие как регуляризация и градиентный спуск. Они помогают нам улучшить качество модели и избежать переобучения. Для использования линейной регрессии необходимо иметь достаточно чистые и обработанные данные. Однако, если правильно подготовить данные и настроить параметры модели, линейная регрессия может быть мощным инструментом в анализе данных и предсказании будущих значений.