Практическое задание 3: Одномерный анализ данных для улучшения качества данных

Цель:

В этом задании вы будете применять методы **одномерного анализа данных** для улучшения качества прогнозирования модели машинного обучения посредством работы с набором данных для обучения и тестирования. Сосредоточившись на отдельных переменных, ваша цель — выявить проблемы и реализовать очистку и преобразование данных. Это улучшение качества данных должно привести к лучшей производительности модели машинного обучения при подгонке простой линейной модели (регрессии или классификации). Поэтому рекомендуется добиваться хорошего качества не за счет чрезмерного усложнения самой модели машинного обучения, а скорее за счет качественного исследования данных, в частности, с использованием одномерного анализа данных.

Набор данных:

Вам нужно найти набор данных, содержащий несколько числовых признаков для задач регрессии или классификации. Ваш набор данных должен состоять из:

Зависимая переменная (y), которую вы будете прогнозировать с помощью простой линейной модели (линейной регрессии или логистической регрессии).

Несколько независимых переменных (x1, x2, x3, ..., не менее пяти), которые вы будете использовать в качестве предикторов. Независимые переменные должны быть как непрерывными, так и дискретными.

Набор данных должен содержать такие недостатки, как:

- Отсутствующие значения
- Выбросы
- Скошенные распределения
- Возможные шумы и ошибки ввода данных

Инструкции

Первоначальная эффективность модели (базовый уровень):

Обучить простую линейную модель (линейную регрессию или логистическую регрессию), используя необработанные данные «как есть» (с разделением данных на обучающие и тестовые).

Оцените ключевые показатели эффективности, такие как R-square, RMSE для регрессии и F1, ROC-AUC для классификации.

Эти показатели будут служить вам отправной точкой.

Анализ одномерных данных: выполните методы анализа одномерных данных для каждой отдельной переменной в наборе данных. Используйте следующие шаги для улучшения качества данных:

Описательная статистика:

Для каждой переменной рассчитайте и интерпретируйте ключевые описательные статистики (среднее значение, медиану, стандартное отклонение, размах, квартили, асимметрию и эксцесс).

Визуализация данных:

Создавайте визуализации (гистограммы, диаграммы типа «ящик с усами», графики плотности), чтобы проверить распределение и выявить потенциальные проблемы, такие как выбросы, асимметрия и пропущенные значения.

Обработка пропущенных значений:

Определите пропущенные значения в каждой переменной.

Выберите подходящую стратегию для их обработки (например, заполнение, удаление) и обоснуйте свой выбор. Для этой цели вы также можете использовать значения, взятые из одномерного распределения или подстановку с моментами (среднее, медиана) распределения.

Обнаружение и устранение отклонений:

Определите выбросы, используя такие методы, как межквартильный размах (IQR) или Z-score.

Определите подходящую стратегию работы с выбросами (удаление, преобразование или устранение) и реализуйте ее.

Преобразование переменных:

Если данные демонстрируют асимметрию или ненормальность, примените преобразования (например, логарифмическое, квадратного корня, box-cox) для нормализации распределения там, где это необходимо.

Масштабирование перменных (при необходимости):

Стандартизируйте или нормализуйте переменные, имеющие широкий диапазон значений, чтобы избежать их непропорционального влияния на модель.

Удаление шума:

Выявите и устраните любые очевидные ошибки ввода данных или шум, которые могут ухудшить производительность модели. Для этой цели вы можете использовать одномерные распределения.

Повторное обучение модели:

После завершения одномерного анализа данных и внесения улучшений в набор данных повторно обучите простую линейную модель, используя очищенные и преобразованные данные.

Пересчитайте показатели эффективности и сравните их с исходными результатами.

Отчет: Подготовьте отчет, включающий следующие разделы:

- Введение. Кратко объясните набор данных и проблему, которую вы решаете на основе этого набора данных (классификация или регрессия).
- Одномерный анализ данных. Предоставьте подробную информацию об анализе и процессе очистки для каждой переменной. Включите описательную статистику, визуализацию и обоснования для примененных вами методов.
- Результаты: представить базовую и улучшенную производительность модели, а также обсудить изменения (или их отсутствие).
- Заключение: обобщите свои выводы и влияние одномерного анализа данных на эффективность модели.

Результаты:

Ваш блокнот Jupyter с кодом, использованным для анализа и подгонки модели.

Отчет в формате PDF, обобщающий анализ и результаты.

Критерии оценки:

1. Точность одномерного анализа данных: насколько эффективно вы выявляли и решали такие проблемы, как пропущенные значения, выбросы и асимметрия?
2. Улучшение производительности модели: произошло ли значительное улучшение производительности модели после очистки и преобразования данных?
3. Обоснование выбора: обосновали ли вы свои методы и объяснили, почему каждый шаг очистки/преобразования данных был необходим?
4. Ясность отчета: хорошо ли организован ваш отчет, содержит ли он понятные пояснения и визуализации?