Домашнее задание: Анализ данных с использованием методов машинного обучения в R

Тема: Использование методов машинного обучения для анализа данных в R.

Цель: Освоить основные подходы к анализу данных с помощью линейной и нелинейной регрессии, нейронных сетей, и моделей случайного леса.

Задача: Найти подходящую базу данных (из открытых источников или своей профессиональной области) и выполнить анализ по следующим этапам.

Этапы выполнения задания

1. Описание выборки и предварительный анализ данных

1. Описание выборки: краткая характеристика данных (размер, структура, ключевые переменные).
2. Проверка числовых переменных на нормальность (например, с использованием теста Шапиро-Уилка и графиков распределения).
3. Построение корреляционной матрицы, визуализация корреляций (heatmap или scatterplot matrix).

2. Линейная регрессия

1. Построить простую линейную регрессию для анализа зависимости одной переменной от другой.
2. Проверить адекватность модели (значимость коэффициентов, проверка на автокорреляцию ошибок, нормальность остатков).
3. Рассчитать коэффициент детерминации R^2 и интерпретировать его.
4. Сделать выводы о модели и ее практическом применении.

3. Множественная регрессия

1. Построить множественную линейную регрессию (с несколькими предикторами).
2. Проверить значимость предикторов, исключить незначимые при возможности.
3. Рассчитать R^2 и его скорректированную версию Adjusted R^2.
4. Сделать вывод о значимости каждого коэффициента и определить, какие переменные наиболее важны для модели.
5. Построить график влияния предикторов (например, через парциальные зависимости).

4. Нелинейные модели регрессии

1. Построить квадратичную, кубическую и экспоненциальную регрессию.
2. Сравнить модели по критериям качества (AIC, BIC, R^2).
3. Обосновать выбор наилучшей модели и интерпретировать результаты.

5. Модель искусственных нейронных сетей

1. Построить модель искусственной нейронной сети с использованием библиотеки `neuralnet` или аналогичной.
2. Проверить адекватность модели: качество предсказаний, визуализация кривой обучения, сравнение с предыдущими методами.
3. Построить график, показывающий результаты работы сети.

6. Модель случайного леса

1. Построить модель случайного леса с использованием библиотеки `randomForest` или аналогичной.
2. Оценить важность признаков и построить график значимости переменных.
3. Сравнить качество предсказаний случайного леса с предыдущими моделями (например, с использованием MSE или RMSE).

Все модели должны сопровождаться уравнениями, а также графиками (за исключением множественной регресии).

Советы по выбору данных

Подходящие темы для исходных данных:

- **Рынок недвижимости**: данные о ценах на квартиры, влияние характеристик (площадь, район, этажность) на стоимость.
- **Продажи на маркетплейсах**: анализ продаж продуктов, зависимости выручки от скидок, сезона или рейтинга товаров.
- **Технические задачи**: производственные данные, анализ факторов, влияющих на производительность оборудования.
- **Экология**: данные о загрязнении воздуха и воды, влияние на здоровье.
- **Финансы**: анализ акций, прогнозирование прибыли компании.

Критерии оценки работы

1. Полнота описания выборки и предварительного анализа.
2. Качество построения и интерпретации моделей.
3. Наличие графиков, таблиц и уравнений регрессии.
4. Корректность выводов и аргументация выбора подходящей модели.
5. Творческий подход к выбору данных и визуализации.