

Прогнозирование стоимости подержанных автомобилей с использованием линейной регрессии на Python

Ран Максим Александрович

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

В статье рассматривается применение метода линейной регрессии для прогнозирования рыночной стоимости подержанных автомобилей на основе таких параметров, как год выпуска, пробег и количество владельцев. На примере реального датасета демонстрируется процесс подготовки данных, обучения модели и интерпретации результатов с использованием библиотек `scikit-learn` и `numpy`. Проведённый анализ показывает, как даже простая линейная модель может быть полезна для оценки цен, несмотря на ограничения, связанные с малым объемом данных и линейными предположениями.

Ключевые слова: Линейная регрессия, машинное обучение, прогнозирование цен, подержанные автомобили, `scikit-learn`, анализ данных

Predicting the cost of used cars using linear regression in Python

Ran Maksim Aleksandrovich

Sholom-Aleichem Priamursky State University

Student

Abstract

The article discusses the use of the linear regression method to predict the market value of used cars based on parameters such as year of manufacture, mileage, and number of owners. Using the example of a real dataset, the process of data preparation, model training, and interpretation of results using the `scikit-learn` and `numpy` libraries is demonstrated. The analysis shows how even a simple linear model can be useful for estimating prices, despite the limitations associated with a small amount of data and linear assumptions.

Keywords: Linear regression, machine learning, price forecasting, used cars, `scikit-learn`, data analysis

1 Введение

1.1 Актуальность

Рынок подержанных автомобилей требует прозрачных и объективных методов оценки стоимости. Традиционные подходы часто субъективны, тогда как методы машинного обучения позволяют автоматизировать процесс, учитывая ключевые факторы ценообразования. Линейная регрессия, как один

из самых интерпретируемых алгоритмов, даёт возможность не только предсказывать цену, но и анализировать влияние каждого параметра на итоговую стоимость.

1.2 Обзор исследований

И.И. Лихенко в исследовании «Ошибки использования линейной регрессии в оценке стоимости: предпосылки, методы обнаружения и программный пакет на Python» анализирует типичные проблемы, возникающие при применении линейной регрессии для прогнозирования цен, включая нарушение предпосылок метода (линейность, гомоскедастичность, нормальность ошибок). Автор предлагает методы диагностики этих ошибок и реализует их в виде Python-библиотеки, что делает работу полезной для практикующих аналитиков [1]. М.Н. Уркумбаев, Л.С. Маркина, А.Е. Циркунов и А.А. Стародуб в работе «Реализация решения упрощённой задачи линейной регрессии на языке Python» рассматривают базовые аспекты построения регрессионной модели, включая подготовку данных и интерпретацию коэффициентов. Особое внимание уделяется практической реализации на Python с использованием библиотек `pumpru` и `scikit-learn`, что делает материал доступным для начинающих специалистов [2]. А.А. Миролубова и О.Л. Ксенофонтова в статье «Оценка развития рынка платных услуг региона с помощью линейной регрессии на Python» демонстрируют применение линейной регрессии для экономического прогнозирования. Авторы подчёркивают важность учёта региональных особенностей и предлагают методику отбора значимых факторов, что расширяет область применения метода за пределы технических задач [3]. Д.Ю. Алюнов в исследовании «Реализация аналитического решения задачи линейной регрессии на Python» фокусируется на математических основах метода, включая вывод нормальных уравнений и их численное решение. Работа дополнена примерами кода для ручного расчёта коэффициентов, что полезно для углублённого понимания алгоритма [4].

1.3 Цель исследования

Цель исследования – разработать модель линейной регрессии для прогнозирования стоимости подержанных автомобилей, оценить её точность на небольшом датасете и выявить основные закономерности, влияющие на цену.

2 Результаты и обсуждения

Линейная регрессия – один из базовых методов машинного обучения и статистики, предназначенный для моделирования линейной зависимости между независимыми переменными (признаками) и зависимой переменной (целевым значением). Её ключевая задача — найти коэффициенты уравнения вида:

$$y = \omega_1 x_1 + \omega_2 x_2 + \dots + \omega_n x_n + b$$

где

- y — прогнозируемое значение (например, цена автомобиля),
- $x_1, x_2, \dots, x_n$ — признаки (год выпуска, пробег и т.д.),
- $w_1, w_2, \dots, w_n$ — весовые коэффициенты,
- b — смещение (intercept).

Линейная регрессия широко применяется в различных областях благодаря своей простоте, интерпретируемости и эффективности. Основное назначение метода заключается в установлении количественной взаимосвязи между независимыми переменными и целевым показателем, что позволяет решать разнообразные практические задачи. В экономике и бизнесе линейная регрессия используется для прогнозирования ключевых показателей - спроса на продукцию, стоимости активов, уровня продаж. Финансовые аналитики применяют этот метод для оценки влияния макроэкономических факторов на доходность ценных бумаг. В маркетинговых исследованиях регрессионный анализ помогает определить степень воздействия рекламных бюджетов на объемы сбыта. Особую ценность линейная регрессия представляет в задачах, где важна не только точность прогноза, но и понимание вклада каждого фактора, так как коэффициенты модели имеют четкую экономическую или физическую интерпретацию. В машинном обучении линейная регрессия часто выступает базовым алгоритмом, с которого начинают изучение методов прогнозирования, а также используют как компонент более сложных ансамблевых моделей. Простота реализации и низкие вычислительные затраты делают этот метод особенно полезным при работе с большими объемами данных, когда применение сложных алгоритмов может быть неоправданно ресурсозатратным.

В данной работе рассматривается применение методов машинного обучения, в частности линейной регрессии, для решения задачи прогнозирования стоимости подержанных автомобилей на языке программирования Python.

Рассмотрим конкретный пример из практики продаж подержанных автомобилей. В нашем распоряжении имеются данные по шести проданным транспортным средствам (рис. 1), где для каждой машины зафиксированы три ключевых параметра: год выпуска, величина пробега в километрах и оценка технического состояния по пятибалльной шкале (где 5 соответствует отличному состоянию). Кроме того, известна окончательная цена продажи каждого автомобиля. Эти данные представлены в виде структурированной таблицы, где каждый экземпляр содержит четыре числовых значения: три характеристики автомобиля и соответствующую цену продажи.

	ГОД	ПРОБЕГ	СОСТОЯНИЕ	ЦЕНА
Машина 1	1996	78000	5	13000
Машина 2	1970	245000	2	2500
Машина 3	1992	30000	3	6000
Машина 4	1980	56000	5	4500
Машина 5	1994	13500	4	9000
Машина 6	1990	64250	3	5000

Рис. 1. Данные о автомобилях

Реализация решения выполнена с использованием популярных библиотек Python. Библиотека NumPy применяется для работы с числовыми массивами и матрицами. Для построения и обучения модели линейной регрессии используется модуль scikit-learn. Подключим данные библиотеки (рис. 2).

```
[ ] import numpy as np
    from sklearn.linear_model import LinearRegression
```

Рис. 2. Импорт необходимых библиотек

Для работы алгоритма необходимо правильно структурировать имеющиеся данные. Каждый элемент списка представляет собой отдельный автомобиль с его характеристиками: год выпуска, пробег (км), оценка технического состояния (по 5-балльной шкале) и фактическая цена продажи (рис. 3). Для удобства дальнейшей обработки мы разделяем эти данные на матрицу признаков и вектор целевых значений.

```
# Данные из таблицы
data = [
    [1996, 78000, 5, 13000],
    [1970, 245000, 2, 2500],
    [1992, 30000, 3, 6000],
    [1980, 56000, 5, 4500],
    [1994, 13500, 4, 9000],
    [1990, 64250, 3, 5000]
```

Рис. 3. Подготовка исходных данных

```
X = np.array([item[:3] for item in data])
y = np.array([item[3] for item in data])
```

Рис. 4. Разделение данных

Здесь X представляет собой матрицу размером 6×3 (6 автомобилей по 3 признака каждый), а y - вектор из 6 значений цен.

Процесс создания и обучения модели линейной регрессии выполняется следующим образом:

1. Инициализация объекта модели линейной регрессии
2. Обучение модели на предоставленных данных (X - признаки, y - целевые значения)
3. Автоматический расчет оптимальных коэффициентов регрессии

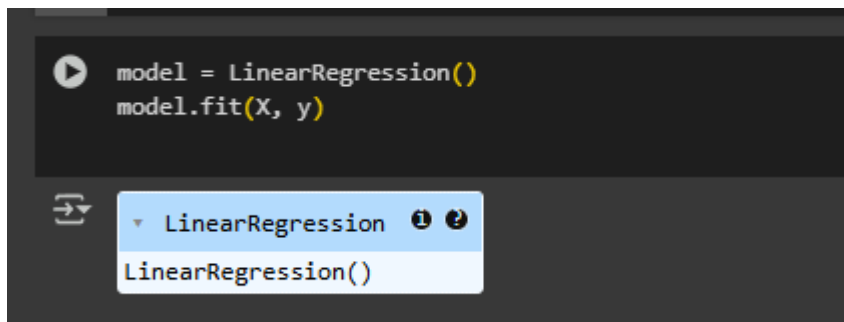


Рис. 5. Создание и обучение модели

Прогнозирование стоимости автомобиля осуществляется путем применения обученной модели линейной регрессии к новым данным. После обучения модели на исторических данных о продажах, она приобретает способность выявлять взаимосвязи между параметрами автомобилей и их конечной стоимостью. Когда мы подаем на вход характеристики нового автомобиля - в данном случае это машина 1990 года выпуска с пробегом 50 000 км и оценкой состояния 4 балла, - модель обрабатывает эти значения с учетом выявленных зависимостей и выдает прогнозную цену (рис. 6).

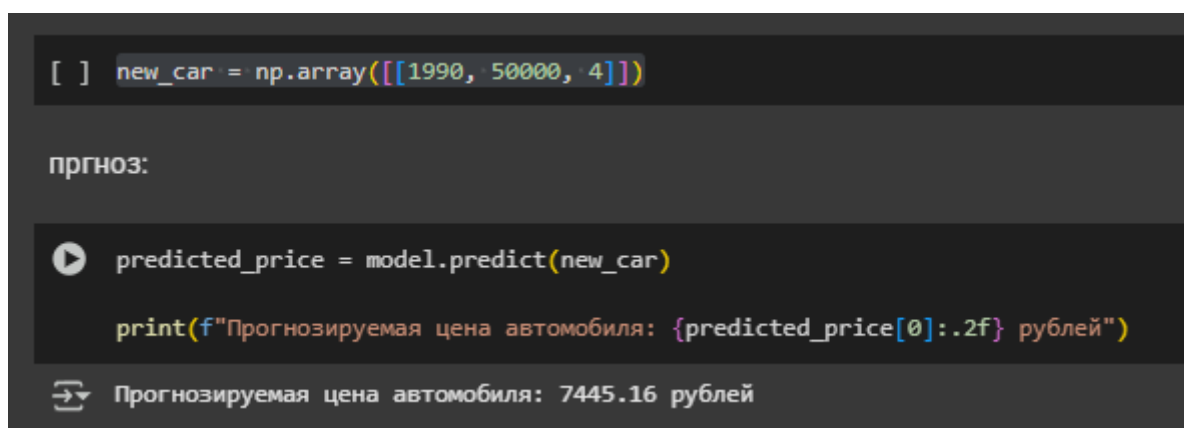


Рис. 6. Прогнозирование стоимости

3 Выводы

Разработанная модель линейной регрессии на Python позволяет прогнозировать стоимость подержанных автомобилей по ключевым параметрам: году выпуска, пробегу и состоянию. Модель демонстрирует

хорошую интерпретируемость. Точность можно повысить добавлением новых параметров и увеличением данных, текущее решение уже полезно для предварительной оценки. Это простой, но эффективный инструмент для автосалонов и частных продавцов, помогающий объективно оценивать автомобили.

Библиографический список

1. Лихенко И.И. Ошибки использования линейной регрессии в оценке стоимости: предпосылки, методы обнаружения и программный пакет на Python // Журнал монетарной экономики и менеджмента. 2024. № 1. С. 77–85.
2. Уркумбаев М.Н., Маркина Л.С., Циркунов А.Е., Стародуб А.А. Реализация решения упрощённой задачи линейной регрессии на языке Python // Перспективные разработки — резидентам территории опережающего социально-экономического развития "Невинномысск". Сборник материалов III Региональной научно-практической конференции. Ставрополь, 2024. С. 175–181.
3. Миролубова А.А., Ксенофонтова О.Л. Оценка развития рынка платных услуг региона с помощью линейной регрессии на Python // Актуальные проблемы экономической деятельности и образования в современных условиях. Сборник научных трудов XVI Международной научно-практической конференции. Волгоград, 2021. С. 102–107.
4. Алюнов Д.Ю. Реализация аналитического решения задачи линейной регрессии на Python // Современные инструментальные системы, информационные технологии и инновации. Сборник научных трудов XVII Международной научно-практической конференции. Курск, 2022. С. 59–62.