ПРОГНОЗИРОВАНИЕ СТОИМОСТИ ПОЕЗДКИ В ТАКСИ С ПОМОЩЬЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

А. С. Марус

Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь, 11115marus@gmail.com
Научный руководитель — Л. Л. Голубева, кандидат физико-математических наук, доцент

Современный мир ежедневно генерирует огромное количество данных, которые хранятся в цифровых форматах, постоянно анализируются и обрабатываются. В наши дни использование машинного обучения с каждым днем становится более востребованным и актуальным для любой компании, так как оно помогает эффективно использовать ресурсы компании и создавать конкуренцию на рынке. В статье рассматривается интеллектуальный анализ статистических данных методами машинного обучения и прогнозирование стоимости поездок в такси

Ключевые слова: машинное обучение; предсказание стоимости поездок в такси; статистические данные; предобработка данных; интеллектуальный анализ данных; обучение с учителем; линейная регрессия; дерево решений; случайный лес; градиентный бустинг; метод k-ближайших соседей; метод опорных векторов; модель многослойного персептрона; Python.

ВВЕДЕНИЕ

Служба такси является неотъемлемой частью жизни современного общества. А прогнозирование стоимости поездок имеет огромное значение для пассажиров, таксистов и компаний такси. Для прогнозирования стоимости поездок в такси используются данные о поездках «желтого» такси за 2023 год. «Желтые» такси имеют право брать пассажиров по всему городу, но по факту циркулируют на Манхеттене [2].

Данные, используемые в данной статье, представляют собой таблицу, состоящую из 18 критериев (столбцов) и более 24 000 000 поездок (строк). Записи о поездках в такси включают поля, в которых указаны даты, время и зоны посадки и высадки пассажиров, расстояния поездки, детализированные тарифы, типы тарифов, типы платежей и количество пассажиров, о которых сообщает водитель [1].

ПРЕДОБРАБОТКА И АНАЛИЗ ДАННЫХ

Важнейшим этапом машинного обучения является предобработка данных. Поэтому для дальнейшей работы требуется оптимизировать типы данных, удалить пропуски в таблице, получить новые признаки, удалить выбросы, удалить столбцы, не несущие полезной для анализа стоимости

поездок в такси информации, выполнить кодирование категориальных признаков и нормализацию данных.

В целях сокращения памяти тип данных int64 был приведен к int8, а float64 к float16.

Часто большие массивы данных имеют пропуски. Отсутствующие данные объектов можно заменить медианными значениями с помощью функции median().

Для качественного анализа стоимости поездок необходимо выделить или синтезировать признаки, которые существенно влияют на стоимость поездки, например, поездки в часы пик, длительность поездки. Поэтому были найдены четыре новых признака: длительность поездки, час начала поездки, день недели начала поездки и месяц начала поездки.

Важный этап предобработки данных – удаление выбросов. Выброс – это элемент/объект данных, который значительно отличается от остальных объектов. Для очистки данных был выбран метод межквартильного диапазона (IQR). Данный метод заключается в нахождении разницы между первым (25-м процентилем) и третьим (75-м процентилем) квартилем набора данных.

Для облегчения работы требуется удалить столбцы, не несущие полезной информации для анализа и прогнозирования стоимости поездок такси. Такие, как данные, указывающие на то, сохранилась ли запись о поездке, тип оплаты и чаевые. Также были выявлены столбцы, не содержащие никакой информации, поэтому они были удалены.

Обрабатываемые данные имеют один столбец с категориальными признаками, RatecodeID (тип поездки). Для кодирования данного столбца был выбран способ One-Hot Encoding.

Для обрабатываемых данных был выбран метод стандартизации StandardScaler изменения размера распределения значений так, чтобы среднее значение наблюдаемых значений стало равно 0, а стандартное отклонение равно 1.

МАШИННОЕ ОБУЧЕНИЕ

В ходе работы исследовано десять различных методов решения задачи регрессии, связанные с построением моделей машинного обучения: линейная регрессия, гребневая регрессия, регрессия Лассо, стохастический градиентный спуск, дерево решений, случайный лес, градиентный бустинг, метод опорных векторов, метод k-ближайших соседей и модель многослойного персептрона [3].

LinearRegression (Линейная регрессия) используется для моделирования линейной зависимости между независимыми переменными и зависимой переменной.

Ridge regression (Гребневая регрессии) – метод линейной регрессии с l2-регуляризацией.

SGDRegressor (Стохастический градиентный спуск) находит локальные максимум или минимум функции при помощи градиента.

DecisionTreeRegressor (Дерево решений) – модель машинного обучения, которая строит дерево, разбивая данные на подгруппы и прогнозируя значения целевой переменной для новых наблюдений.

KNeighborsRegressor (Метод k-ближайших соседей) – модель регрессии, которая ищет k ближайших соседей для каждой точки в тестовом наборе данных и усредняет их целевые значения.

RandomForestRegressor (Случайный лес) – алгоритм машинного обучения, основанный на объединении большого количества деревьев решений для получения более точного прогноза.

GradientBoostingRegressor (Градиентный бустинг) – метод машинного обучения, основанный на построении последовательности слабых моделей, каждая из которых исправляет ошибки предыдущей модели.

SVR (Метод опорных векторов) – метод машинного обучения, основанный на оптимизации разделяющей гиперплоскости с учетом ширины зазора, за пределами которого должны находиться точки данных.

MLPRegressor (Многослойный персептрон) – модель нейронной сети, представляет собой многослойный перцептрон, состоящий из нескольких слоев нейронов, включая входной слой, скрытые слои и выходной слой.

Модели RandomForestRegressor, GradientBoostingRegressor, SVR и MLPRegressor показали одни из лучших результатов, поэтому затем были улучшены при помощи настройки гиперпараметров.

Гиперпараметры – это параметры модели, которые настраиваются до начала процесса ее обучения. Они определяют саму структуру модели и способ её обучения, имеют решающее значение для достижения ее высокой производительности и точности. Для подбора гиперпараметров применяется байесовский алгоритм BayesSearchCV.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ

Для оценки результатов прогнозирования моделей регрессии чаще всего используются метрики средней абсолютной ошибки (1), средней квадратичной ошибки (2) и коэффициента детерминации (3):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - y_i^* |, \quad (1)$$

$$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2}, \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - y_i^{**})^2}, \quad (3)$$

где y_i – фактическое значение для -го наблюдения, y_i^* – предсказанное значение для i-го наблюдения, y_i^{**} – среднее арифметическое всех правильных ответов, n – количество наблюдений.

Все модели показали хорошие результаты (таблица). Модель опорных векторов (SVR) показала наименьшую среднюю абсолютную ошибку, модель многослойного персептрона (MLPRegressor) показала наименьшую среднюю квадратическую ошибку, а коэффициент детерминации оказался лучшим у моделей RandomForestRegressor, SVR и MLPRegressor.

Сравнительный анализ моделей

Модель/Метрика	*MAE*	*MSE*	R^2
LinearRegression	0.415	0.483	0.966
Ridge	0.413	0.468	0.966
Lasso	1.045	1.972	0.858
SGDRegressor	0.414	0.470	0.966
DecisionTreeRegressor	0.358	0.720	0.948
RandomForestRegressor	0.270	0.338	**0.976**
GradientBoostingRegressor	0.261	0.343	0.975
SVR	**0.260**	0.340	**0.976**
KNeighborsRegressor	0.358	0.434	0.969
MLPRegressor	0.269	**0.331**	**0.976**

Библиографические ссылки

1. TLC Trip Record Data [Electronic resource]. URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page (date of access: 23.04.2024).

2. Sripathi Mohanasundaram. Newyork Yellow Taxi Trip Data [Electronic resource]. URL: https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019 (date of access: 24.04.2024).

3. Scikit-learn. Машинное обучение в Python [Электронный ресурс]. URL: https://scikit-learn.org/stable/index.html (date of access: 08.05.2024).