ИССЛЕДОВАНИЕ ПРИМЕНИМОСТИ РЕГРЕССИОННЫХ МЕТОДОВ АНАЛИЗА ДАННЫХ ДЛЯ ЗАДАЧ ПРОГНОЗИРОВАНИЯ ЗАГРУЗКИ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ

Алексеева Н.С., Воронов В.И.
МТУСИ «Московский Технический Университет Связи и Информатики»

Актуальной проблемой, возникающей при развитии и эксплуатации облачной платформы, является эффективная загрузка и планирование вычислительных ресурсов. Объектом исследования в данной работе является процесс машинного обучения. Предмет исследования в данной работе является облачная платформа коммерческой компании. Цель работы заключается в исследовании применимости регрессионных методов для оценки загрузки облачной платформы (ОП).

Целью машинного обучения является частичная или полная автоматизация решения профессиональных задач. [1] В основе современных методологий управления ИТ лежит стремление навести порядок путём организации системы менеджмента качества ИТ-услуг.[2] Управление мощностями - процесс, отвечающий за своевременное и эффективное по затратам соответствие мощности услуг и инфраструктуры требованиям согласованных целевых показателей уровня услуг.[3] Основная задача процесса управления мощностями, отвечает за обеспечение соответствующего баланса в ИТ-организации. Один из технологических элементов, который набирает популярность, это «обнаружение знаний в базах данных» Knowledge Discovery in Databases (KDD). [4] Этот процесс состоит из множества шагов, одним из которых является интеллектуальный анализ данных Data Mining (DM). Основной особенностью Data Mining является использование методов классификации, кластеризации и прогнозирования, применяемых для выявления неявных закономерностей и свойств, которые присутствуют в собранных данных. Прогнозирование является задачей предсказания будущего, на основании накопленных знаний. Анализируя данные прошлых периодов, можно построить с некоторой точностью прогноз на будущее. [5]

Рассмотрим линейную регресию. Задача данного анализа описать имеющиеся данные уравнением прямой. Мы знаем точки и знаем координаты, так же знаем что X влияет на Y, и мы наблюдаем определенные значения Y при конктерном X.

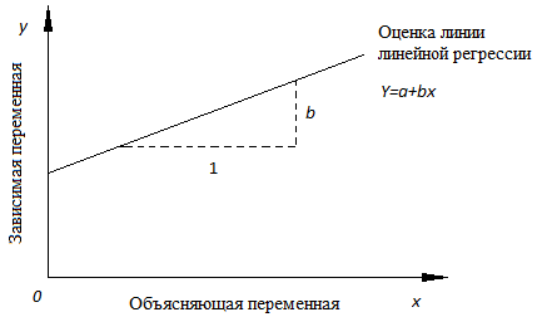

Рис.1 Линия линейной регресии

Происходит подбор коэффициентов прямой, чтобы минимизировать ошибку описания данных. Для оценки параметров регрессионного уравнения часто используют метод наименьших квадратов. Сущность данного метода в нахождении параметров модели, при которых

минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии [6] Корень из этой ошибки: RMSE – Root Mean Squared Error. Средняя Абсолютная Ошибка- это средний модуль отклонения (MAE – Mean Absolute Error). Как MAE, так и показатель RMSE выражают среднюю ошибку прогнозирования модели в единицах интересующей переменной. Обе метрики могут находиться в диапазоне от 0 до ∞ и безразличны к направлению ошибок. Это отрицательно ориентированные оценки, что означает, что чем ниже значение, тем лучше.

Для моделирования программного обеспечения был выбран язык программирования Python. Исторические данные, были выгружены из системы управления облачной платформы, за период начиная с марта 2013 года по сентябрь 2020 года в формате exl. Рассмотрим линейную регрессию для среднесрочного прогнозирования .

Для начала работы необходимо добавить библиотеки и импортировать в программу

```
import matplotlib.pyplot as plt
import datetime as dt
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.preprocessing import PolynomialFeatures
from sklearn import preprocessing
```

Пи работе с датасетом были выявлены пустые ячейки со сроком заказа, по умолчанию установлено 30 календарных дней, коммерческие заказы установлены до текущей даты. Далее выполняется подсчет загрузки вычислительных ресурсов из датасета:

```
# Подсчет загрузки
def calc_load(dataset, firstTestDate, lastTestDate, firstComercDate,
lastComercDate):
    start_date = firstComercDate if firstTestDate >= firstComercDate else
firstTestDate
    end_date = lastComercDate if lastTestDate <= lastComercDate else lastTestDate
```

Так же необходимо произвести подсчёт загрузки по vCPU, vRAM, vHDD

```
# Производим общий подсчет загрузки по всем датам
resource_load_dates, loading_resource_cpu, loading_resource_ram,
loading_resource_hdd = calc_load(data,
```

Максимальные значения исследуемых параметров:

```
max_cpu = 89481
max_ram = 252710
max_hdd = 3000000
```

В полном наборе данных было выделено 80% для тестирования и 20% обучение за период 2013-2020год. Создадим модель для линейной регрессии и вычислим оптимальное значение весов для каждого параметра. Результат вычислений представлен на рисунках 2,3,4

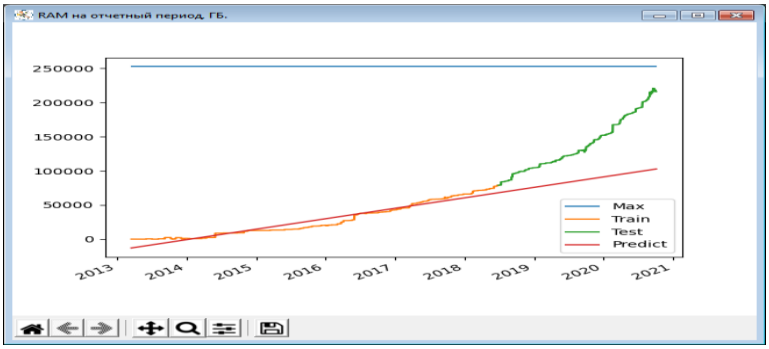

Рис. 2 Показатель RAM

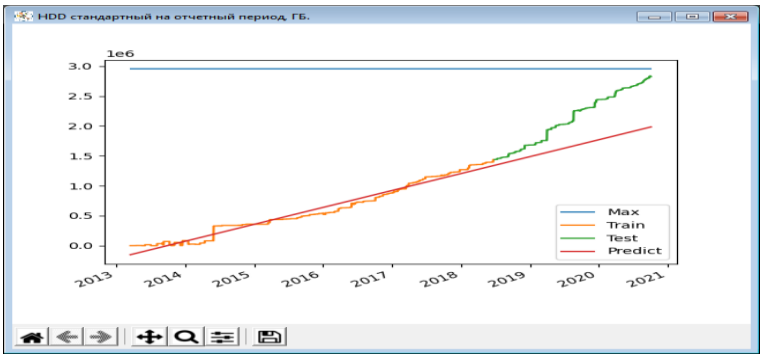

Рис. 3 Показатель HDD

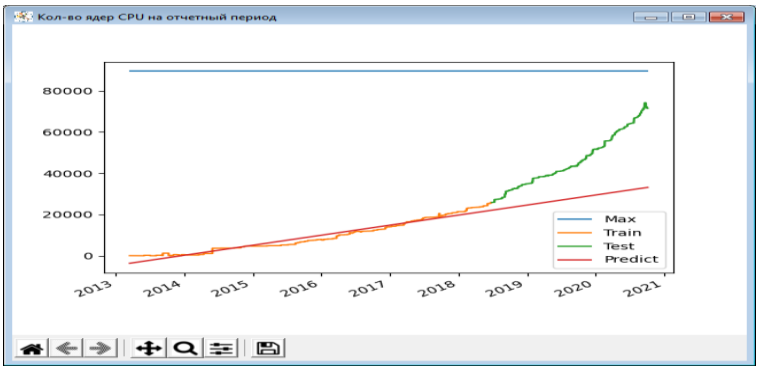

Рис. 4 Показатель CPU

Вычисление RMSE и MAE линейной регрессии дает следующие значения для показателя CPU: MAE 4 155,28, RMSE 5 492,14, для показателя HDD: MAE 129 909,44, RMSE 187 136,19 и для показателя RAM: MAE12 768,25, RMSE17 242,71. На основании полученных результатов можем сделать вывод, что линейная регрессия не так хорошо описывает данные и прогнозирует будущую загрузку ресурсов и не подойдет для среднесрочного планирования 3-5 лет. Чтобы преодолеть несоответствие, нам нужно увеличить сложность модели.

Для создания полиномиальной модели мы будем использовать `PolynomialFeature` класс предоставлен `scikit-learn`. Далее мы обучаем нашу модель. Создаем полиномиальную регрессию для исследуемых параметров:

```
poly_cpu = PolynomialFeatures(poly_degree_cpu)
X_poly_cpu = poly_cpu.fit_transform(x_train_dates_int)

poly_ram = PolynomialFeatures(poly_degree_ram)
X_poly_ram = poly_ram.fit_transform(x_train_dates_int)

poly_hdd = PolynomialFeatures(poly_degree_hdd)
X_poly_hdd = poly_hdd.fit_transform(x_train_dates_int)
```

Выбор степеней полиномиальной регрессии производился вручную, наилучшие результаты были получены при следующих показателях:

```
poly_degree_cpu = 2
poly_degree_ram = 2
poly_degree_hdd = 3
```

Подгонка модели линейной регрессии к преобразованным объектам дает следующие графики, Рисунки 5,6,7

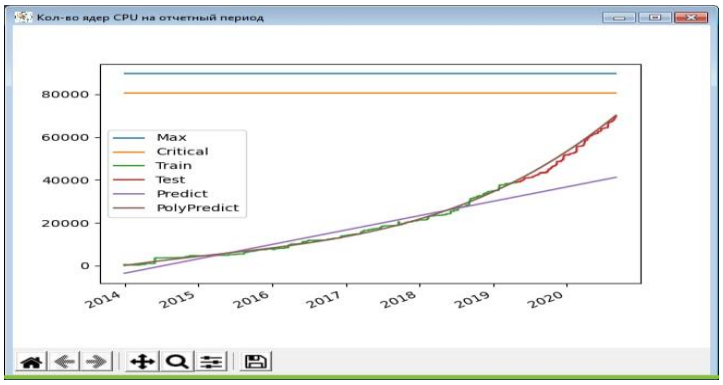

Рис. 5 Показатель CPU

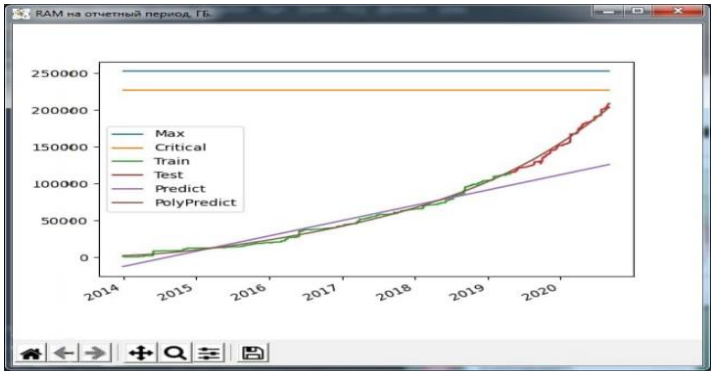

Рис. 6 Показатель RAM

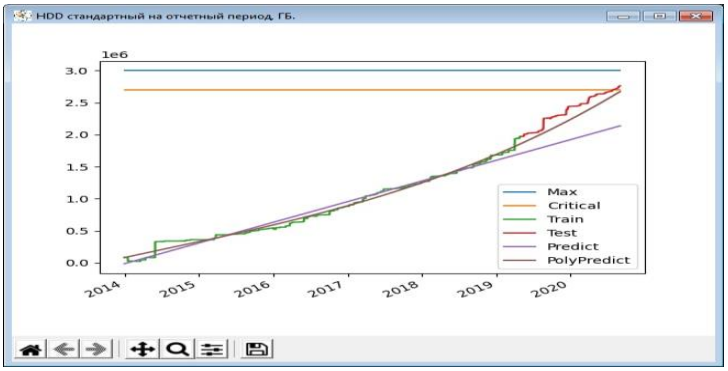

Рис. 7 Показатель HDD

Вычисление RMSE и MAE полиномиальной регрессии дает следующие значения для показателя CPU: MAE 2 128,55, RMSE 3 049,21, для показателя HDD: MAE 63 954,72, RMSE 92 568,09 и для показателя RAM: MAE 6 184,13, RMSE 8 521,35.

Моделирование при более коротком периоды демонстрирует неплохие показатели, особенно интересен показатель HDD. На полученном графике по показателю HDD мы видим, как прогнозная линия и тестовая перешли за «пороговое значение» 90% загрузки по ресурсам, что означает необходимость планирования расширения мощностей.

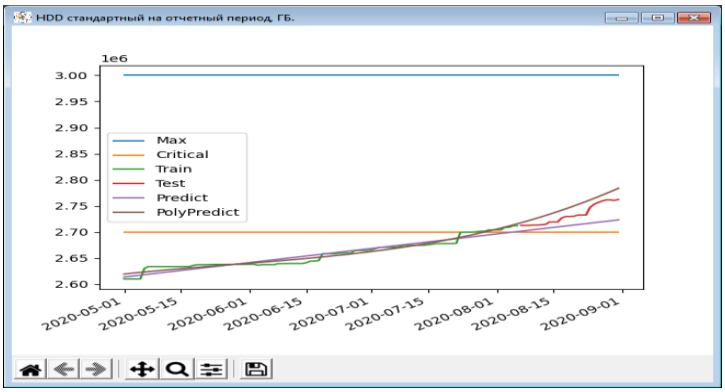

Рис. 8 Показатель HDD

Показатели RMSE и MAE для каждой модели при планировании на краткосрочном периоде -4 месяца

```
Вычисление результатов
полиномиальный регрессии
CPU RMSE: 2.277                    CPU MAE: 0.137
RAM RMSE: 1.037                    RAM MAE: 0.886
HDD RMSE: 4.335                    HDD MAE: 0.197
```

Полиномиальная регрессия оказалась более гибкой и может моделировать сложные взаимосвязи как на длительные, так и на краткосрочный период. Присутствует полный контроль над моделированием переменных объекта (выбор степени) Минусы модели: при неправильном выборе степени, данная модель может быть перенасыщена.

Таким образом, линейная регрессия может быть применима на краткосрочном прогнозе, при регулярном обновлении данных, а так же с небольшой периодичностью прогноза не более чем 1 месяц. Полиномиальная регрессия более точно описывает как краткосрочный, так и долгосрочный прогнозы, поэтому ее применение считается более эффективном. Прогнозирования на основе регрессионных моделей является легко интерпретируемыми и практически применимыми, что является безусловно достоинством.

Список литературы

1. Машинное обучение. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. URL: http://www.machinelearning.ru/
2. Исайченко, Д., Журавлёв, Р.ITSM. Руководство по измерению / Лайвбук, 2015. — 141 с.
3. Д.А. Скрипник ITIL. IT Service Management по стандартам V.3.1: учебное пособие, Электрон.- Москва: Интернет-Университет Информационных Технологий, 2020
4. Mark Gilchrist.Knowledge Discovery in Databases for Competitive Advantage, Journal of Management and Strategy,2012 URL http://www.sciedu.ca/journal/index.php/jms/article/view/966
5. И.С. Светуньков, С.Г. Светуньков Методы социально-экономического прогнозирования в 2 т. Т. 2 модели и методы. 2015-448с.
6. В.В. Домбровский Эконометрика, Томск – 2016 г. URL: http://sun.tsu.ru/mminfo/2016/Dombrovski/book/toc.ht