МАШИННОЕ ОБУЧЕНИЕ В СРЕДЕ СУБД MS SQL SERVER

Абалуев Роман Николаевич
кандидат педагогических наук, доцент
Крумкаченко Артём Андреевич
студент 4 курса Инженерного института

ФГБОУ ВО Мичуринский ГАУ
г. Мичуринск, Россия

Аннотация: В статье проводится обзор современных методов анализа данных посредством процедур машинного обучения интегрированных в систему управления базами данных MS SQL Server 2017.

Ключевые слова: Машинное обучения, системы управления базами данных, информационные системы, хранимая процедура, язык R.

Для автоматизации бизнес-процессов на предприятиях в настоящее время активно используются информационные системы, в процессе функционирования которых накапливаются большие объемы данных. Для предприятий важно иметь возможность прогнозировать результаты своей деятельности используя накопленную информацию [1]. Для решения такого рода задач могут использоваться службы машинного обучения на базе MS SQL Server.

На сегодняшний день одной из наиболее мощных систем работы с базами данных на основе архитектуры "клиент-сервер" является СУБД MS SQL Server, которая в своем составе имеет средства создания баз данных, обработки информации, импорта и экспорта данных, разграничения доступа к информации, резервного копирования и восстановления информации, оптимизации и выполнения запросов к данным, которые сейчас активно используются для информатизации различных отраслей в том числе и сельскохозяйственными предприятиями [2].

Под машинным обучением понимают класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, различные техники работы с данными в цифровой форме.

SQL Server Службы машинного обучения позволяет выполнять скрипты языков Python и R в базе данных. С его помощью можно подготавливать и визуализировать данные, выполнять проектирование признаков, а также обучать, оценивать и развертывать модели машинного обучения в базе данных.

Используя системы машинного обучения с системами хранения данных, можно получить новые данные и генерировать аналитику по хранимым данным без увеличения транзакционной рабочей нагрузки

Службы машинного обучения можно применять для создания и обучения моделей машинного обучения и глубокого обучения в SQL Server, а также для графической интерпретации больших массивов данных.

Для исследования возможностей служб машинного обучения нами была изучена демонстрационная база данных NYCTaxi с сайта компании Microsoft [3]. База содержит набор данных о 1,7 млн поездок в такси.

Первоначально мы скачали и развернули базу NYCTaxi на экземпляр MS SQL Server. Затем проверили, что объекты базы данных существуют и восстановление прошло успешно. Далее мы создали запрос, который сортирует данные по числу пассажиров и сумме тарифов. Запрос выводит данные по количеству пассажиров, все тарифы и средний тариф (Рис.1.).

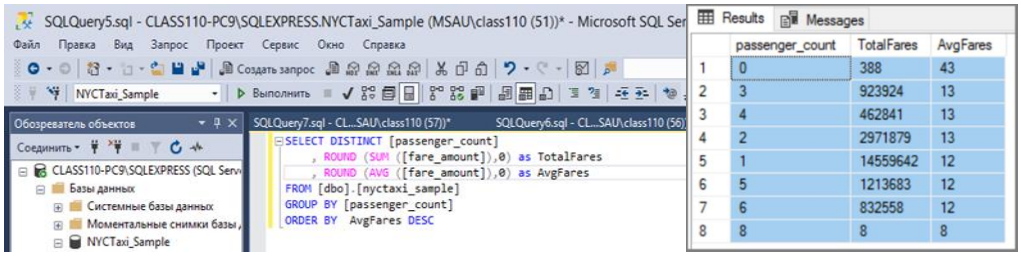

Рисунок 1. Запрос для группировки по количеству пассажиров и расчета тарифов.

Непосредственно для изучения служб машинного обучения нами была интегрирована в наш экземпляр MS SQL Server следующая хранимая процедура:

```
CREATE PROCEDURE [dbo].[RPlotHist]
AS
BEGIN
  SET NOCOUNT ON;
  DECLARE @query nvarchar(max) =
  N'SELECT cast(tipped as int) as tipped, tip_amount, fare_amount FROM [dbo].[nyctaxi_sample]'
  EXECUTE sp_execute_external_script @language = N'R',
  @script = N'
    mainDir <- "D:\\mssql\\hist"
    dir.create(mainDir, recursive = TRUE, showWarnings = FALSE)
    setwd(mainDir);
    print("Creating output plot files:", quote=FALSE)
    dest_filename = tempfile(pattern = "rHistogram_Tipped_", tmpdir = mainDir)
    dest_filename = paste(dest_filename, ".jpg",sep="")
    print(dest_filename, quote=FALSE);
    jpeg(filename=dest_filename);
    hist(InputDataSet$tipped, col = "lightgreen", xlab="Tipped",
       ylab = "Counts", main = "Histogram, Tipped");
     dev.off();
    dest_filename = tempfile(pattern = "rHistograms_Tip_and_Fare_Amount_", tmpdir = mainDir)
    dest_filename = paste(dest_filename, ".pdf",sep="")
    print(dest_filename, quote=FALSE);
    pdf(file=dest_filename, height=4, width=7);
```

```
  par(mfrow=c(1,2));
  hist(InputDataSet$tip_amount, col = "lightgreen",
     xlab="Tip amount ($)",
     ylab = "Counts",
     main = "Histogram, Tip amount", xlim = c(0,40), 100);
  hist(InputDataSet$fare_amount, col = "lightgreen",
     xlab="Fare amount ($)",
     ylab = "Counts",
     main = "Histogram,
     Fare amount",
     xlim = c(0,100), 100);
 dev.off();
  dest_filename = tempfile(pattern = "rXYPlots_Tip_vs_Fare_Amount_", tmpdir = mainDir)
  dest_filename = paste(dest_filename, ".pdf",sep="")
  print(dest_filename, quote=FALSE);
  pdf(file=dest_filename, height=4, width=4);
  plot(tip_amount ~ fare_amount,
     data = InputDataSet[sample(nrow(InputDataSet), 10000), ],
     ylim = c(0,50),
     xlim = c(0,150),
     cex=.5,
     pch=19,
     col="darkgreen",
     main = "Tip amount by Fare amount",
     xlab="Fare Amount ($)",
     ylab = "Tip Amount ($)");
   dev.off();',
@input_data_1 = @query
END
```

Выходные данные запроса SELECT в хранимой процедуре сохраняются в кадре данных R после чего можно вызывать различные функции построения диаграмм R для создания графических файлов. Все файлы сохраняются в локальной папке 'D:\mssql\hist. Папка назначения определяется аргументами, предоставляемыми скрипту R в рамках хранимой процедуры. Для выбора другой папки нужно изменить значение переменной mainDir. В результате выполнения хранимой процедуры мы получили гистограммы «Суммарная выручка», «Стоимость проезда» и график «Суммарная выручка по тарифам» (Рис.2).

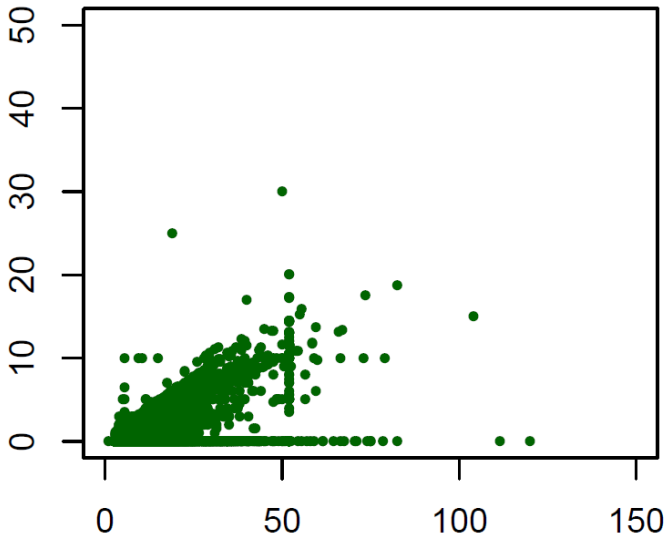

Рисунок 2. Классификация поездок в зависимости от объема суммарной выручки.

В результате исследования демонстрационной базы мы интегрировали хранимую процедуру RPlotHist в рабочий экземпляр MS SQL Server 2017, с помощью которой в дальнейшем можно получать графическую интерпретацию больших массивов информации в базах данных.

Список литературы

1. Абалуев Р.Н. Методика оценки производительности систем управления базами данных автотранспортных предприятий. // Инфокоммуникационные и интеллектуальные технологии на транспорте IITT'2018 [Текст]: материалы I междунар. науч.-практ. конф., 12-13 декабря 2018 г. В 2 т. Т. 1. – Липецк: Изд-во Липецкого государственного технического университета, 2018. – С.171 -174.

2. Абалуев Р.Н., Косенков Д.В. Информационное обеспечение сельского хозяйства. // Научный рецензируемый электронный журнал «Наука и Образование». №2. – 2019.

3. Аналитика данных R для разработчиков SQL. Официальный сайт Microsoft // URL: https://docs.microsoft.com/ru-ru/sql/advanced-analytics/tutorials/sqldev-in-database-r-for-sql-developers?view=sql-server-2017 (дата обращения: 17.09.2019).

4. Криволапов И.П., Щербаков С.Ю., Манаенков К.А. Актуальность подготовки инженерных кадров для обеспечения экологической безопасности

сельскохозяйственного производства // В сборнике: Экологическая педагогика: проблемы и перспективы в свете развития технологий Индустрии 4.0 Материалы Международной научной школы, организованной при финансовой поддержке Администрации Тамбовской области. Под общей редакцией Е.С. Симбирских. 2017. С. 22-24.

MACHINE LEARNING IN THE DBMS ENVIRONMENT MS SQL SERVER

Abaluev Roman Nikolaevich,
Associate Professor,
Krumkachenko Artem Andreevich
4rd year student Engineering Institute
e-mail:abaluevrn@mgau.ru

Michurinsk State Agrarian University,
Michurinsk, Russia.

Annotation: The article reviews modern methods of data analysis through machine learning procedures integrated into the MS SQL Server 2017 database management system.

Keywords: machine learning, database management systems, information systems, stored procedure, language R.