Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка алгоритма повышения качества данных розничной торговли

Выполнила: Лопатина Анастасия Сергеевна, гр. 6381
Руководитель: Кирьянчиков Владимир Андреевич, к.т.н., доцент

Санкт-Петербург, 2020

Проблема и актуальность

Актуальность: В современных условиях конкурентного сосуществования компаний розничной торговли остро стоит вопрос о качестве анализируемых данных. Работа с качественными данными позволяет фирмам быть уверенными в результатах, которых они достигают.

Проблема: Влияние качества данных на их дальнейший анализ и конкурентоспособность компании

Цель и задачи

Цель: Разработать алгоритмы, обрабатывающие пропущенные и аномальные значения в данных о продажах

Задачи:

- Исследовать предметную область
- Сравнить существующие подходы
- Разработать алгоритм, устраняющий пропуски в данных
- Разработать алгоритм, обнаруживающий аномальные значения в данных
- Выполнить анализ полученных результатов

Исследование предметной области

Качество данных – это оценка пригодности данных для выполнения предполагаемой цели их использования

Аспекты качества данных:

- Полнота
- Согласованность
- Точность
- Своевременность
- Действительность
- Уникальность

Этапы проработки аспектов:

- Профилирование
- Стандартизация
- Очистка
- Обогащение
- Дедупликация

Сравнение подходов устранения пропущенных значений

Название подхода	Критерии		
	Тип пропусков	Уменьшение количества данных	Искажение распределения
Удаление строк	MCAR	+	-
Методы с заполнением	MAR	-	+
Методы, основанные на моделировании	MAR, MCAR	-	+
Методы семейства Zet (Wanga)	MAR	-	-
Методы взвешивания	MAR	+	-

Разработка алгоритма, устраняющего пропуски: EM-алгоритм

«Expectation»:

$$\mathrm{Q}\left(\theta;\theta^{(\mathrm{m})}\right)=\mathrm{E}_{\theta^{(m)}}(\log L(\theta;Y_O,Y_M)|Y_O=y_O),$$

где Y_O – наблюдаемые значения, Y_M – пропущенные значения, $\theta^{(\mathrm{m})}$ – оценка параметров на итерации m, $\mathrm{Q}\left(\theta;\theta^{(\mathrm{m})}\right)$ – ожидаемый логарифм правдоподобия

«Maximization»:

$$\theta^{(\mathrm{m}+1)}=\arg\max_{\theta}\mathrm{Q}(\theta;\theta^{(\mathrm{m})})$$

Анализ результатов устранения пропусков

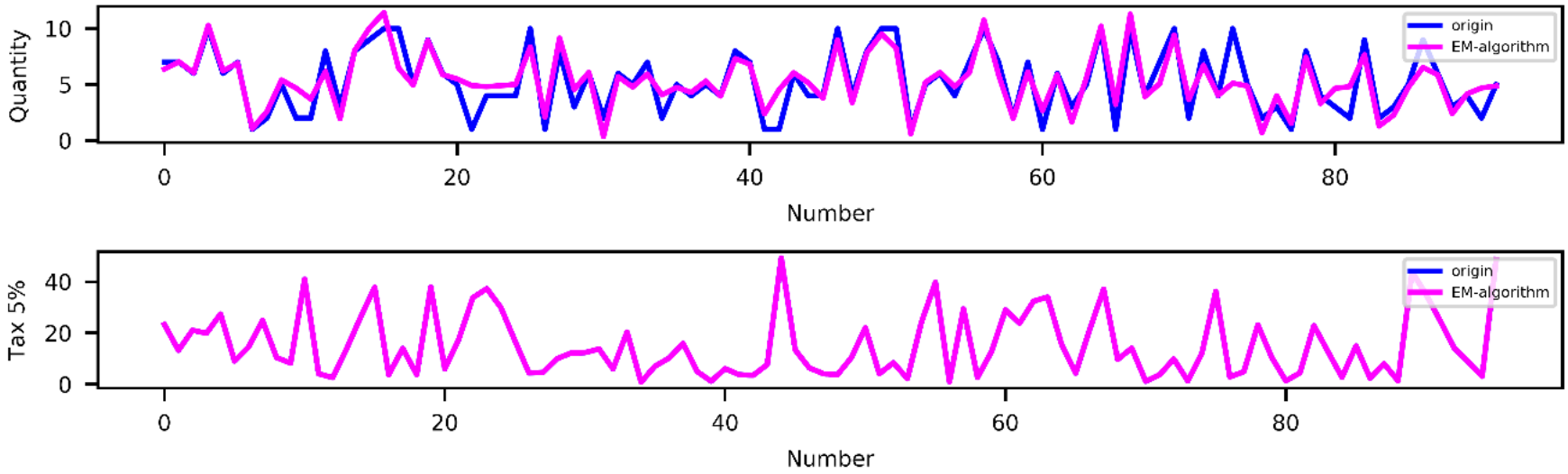

Рисунок 1 – Восстановление пропусков в признаках «Quantity» и «Tax 5%»

	MCAR			**MAR**		
% пропусков	10	30	50	10	30	50
MSE	0,92	7,85	14,88	1,12	6,14	15,7

Сравнение подходов обнаружения аномальных значений

Название подхода	Критерии		
	Режим обнаружения	Результат работы	Временная сложность работы
Статистический анализ	Semisupervised	Вероятность	$O(n)$
Кластеризация	Semisupervised, Unsupervised	Метка	$O(n^2)$
Классификация	Semisupervised, Supervised	Метка	$O(n)$
Методы, основанные на алгоритме ближайшего соседа	Unsupervised	Вероятность	$O(n^2)$
Спектральные методы	Semisupervised, Unsupervised	Метка	$O(n)$
Деревья решений	Supervised, Unsupervised	Метка	$O(tnlogn)$

Разработка алгоритма, обнаруживающего аномалии на основе k Nearest Neighbors

$$d_{max} \leq \left| d(x_i, c_k) - d(p_j, c_k) \right|,$$

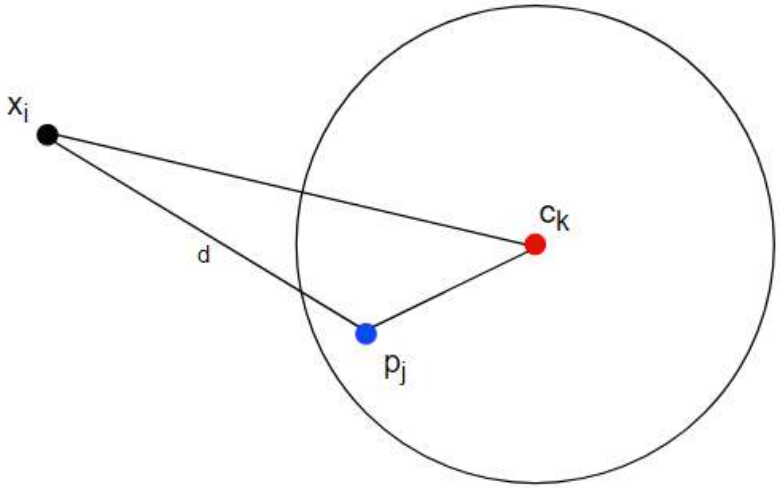

Рисунок 2 – Графическое изображение алгоритма

где c_k – центр k-го кластера, p_j – j-ый объект обучающей выборки, x_i – i-ый объект тестовых данных, $\mathrm{d}(\mathrm{x}_i, p_j)$ – расстояние между x_iи p_j, d_{max}– максимальное расстояние между тестовым объектом и его соседями

Анализ результатов обнаружения аномалий

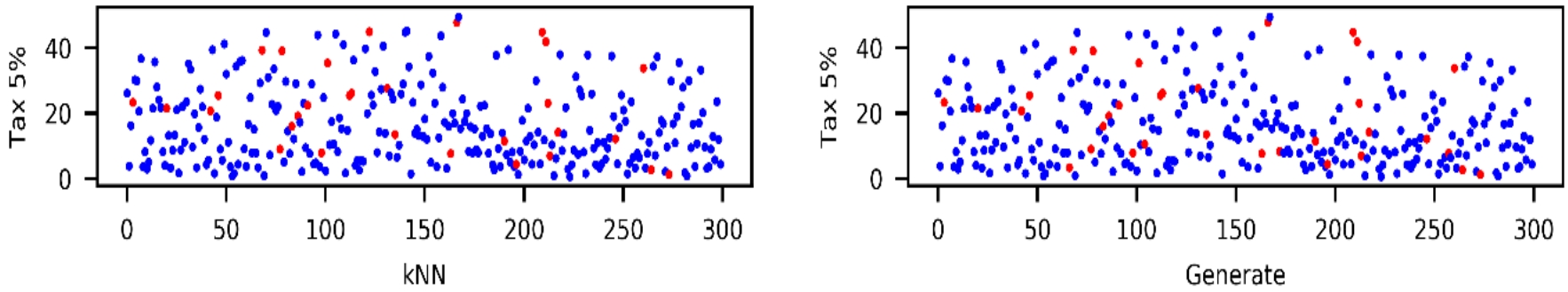

Рисунок 3 – Обнаружение аномалий в признаке «Tax 5%»

% аномалий	Recall	Accuracy	Precision	F
5	0,61	0,96	0,68	0,64
10	0,66	0,94	0,80	0,72
25	0,75	0,88	0,75	0,75
40	0,70	0,77	0,74	0,72

Сравнение времени работы

Датасет	Разработанный kNN, с	Классический kNN, с
Supermarket Sales	3,56	19,62
Sales Data	3,25	11,23
Retail Data Customers Summary	4,16	27,01

Заключение

- Исследование предметной области показало, что существует широкий круг задач для улучшения качества данных, работа с аномалиями и пропусками осуществляется на этапе очистки и позволяет повысить точность и полноту данных
- В результате сравнения существующих подходов для обработки пропущенных значений была выбрана группа методов, основанная на моделировании, для обработки аномальных значений – подход, основанный на алгоритме ближайшего соседа
- Реализован EM-алгоритм для восстановления пропусков, а также kNN алгоритм для обнаружения аномалий
- Анализ результатов показал, что разработанные алгоритмы могут стать альтернативными средствами для решения заявленной проблемы
- В качестве дальнейших исследований предлагается рассмотреть способы оптимизации EM-алгоритма, анализ LOF на данных розничной торговли

Апробация работы

- «Сравнительный анализ подходов для решения задачи повышения качества данных розничной торговли» // Конференция ППС СПбГЭТУ «ЛЭТИ», 2020.
- «Comparative Analysis of Approaches for Solving the Problem of Improving the Quality of Retail Trade Data by Artificial Intelligence» // MECO, 2020.
- Репозиторий проекта https://github.com/Anastasiyalopatina/Graduated_work

Спасибо за внимание!

Метрики качества

При устранении пропусков:

$$E(X) = \sum_{i=1}^{k} x_i p_i$$

$$D(X) = E\big(X - E(X)\big)^2$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

При обнаружении аномалий:

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall}$$

Математическое ожидание признаков с восстановленными значениями

% пропусков	Unit price	Quantity	Tax 5%	Total	COGS	Gross income	Rating
0	55,67	5,51	15,38	322,97	307,59	15,38	6,97
MCAR							
10	55,57	5,51	15,38	322,97	307,59	15,38	6,99
30	56,24	5,48	15,37	322,69	307,33	15,37	7,03
50	56,12	5,50	15,46	324,59	309,14	15,46	6,99
MAR							
10	55,69	5,53	15,38	322,97	307,59	15,38	6,97
30	55,31	5,48	15,41	323,69	308,27	15,41	7,00
50	55,05	5,48	15,39	323,18	307,79	15,39	7,00

Дисперсия признаков с восстановленными значениями

% пропусков	Unit price	Quantity	Tax 5%	Total	COGS	Gross income	Rating
0	701,97	8,55	137,10	60459,60	54838,64	137,10	2,95
MCAR							
10	691,35	8,36	137,10	60459,60	54838,64	137,10	2,64
30	628,88	7,71	136,37	60138,38	54547,28	136,37	2,08
50	581,15	6,41	130,12	56994,40	51811,24	125,98	2,05
MAR							
10	680,51	8,26	137,10	60459,60	54838,64	137,10	2,76
30	634,93	7,83	136,94	60392,52	54777,80	136,94	2,01
50	600,61	7,99	126,74	55892,31	51268,25	128,16	2,11

Таблица корреляций

	Unit price	Quantity	Tax 5%	Total	COGS	Gross income	Rating
Unit price	1,00	0,01	0,63	0,63	0,63	0,63	-0,01
Quantity	0,01	1,00	0,71	0,71	0,71	0,71	-0,02
Tax 5%	0,63	0,71	1,00	1,00	1,00	1,00	-0,04
Total	0,63	0,71	1,00	1,00	1,00	1,00	-0,04
COGS	0,63	0,71	1,00	1,00	1,00	1,00	-0,04
Gross income	0,63	0,71	1,00	1,00	1,00	1,00	-0,04
Rating	-0,01	-0,02	-0,04	-0,04	-0,04	-0,04	1,00