Практическая работа 2

По предмету «Технологии интеллектуального анализа данных мониторинга безопасности»

Выполнил: Воронцов С. А.

Проверил: Латыпова О. В.

Выявление мошенничества с помощью с помощью платформы Knime

Цели и задачи работы

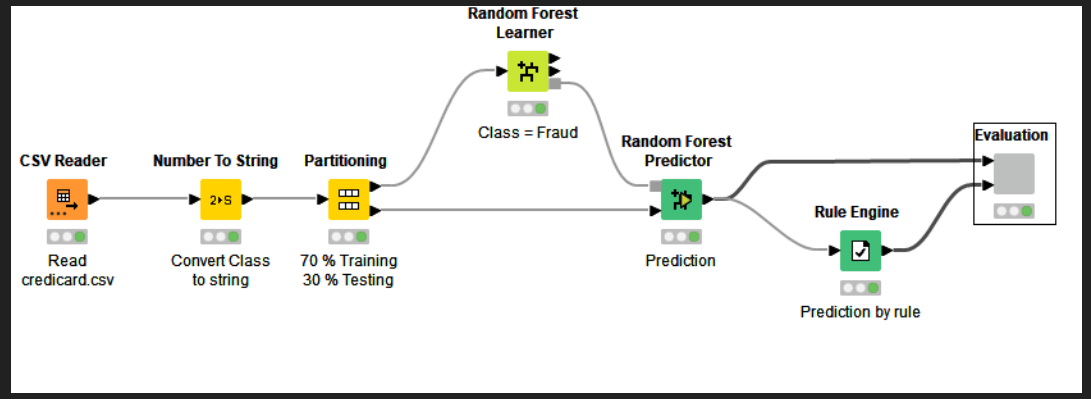
- Изучить данную выборку данных и модель машинного обучения;
- Модифицировать модель машинного обучения и посмотреть на изменения;
- Модифицировать модель машинного обучения.

Выборка

- Выборка данных крайне несбалансированная;
- Выборка имеет 31 характеристику;
- Характеристики имеют выбросы;
- Выборка была модифицирована с помощью метода главных компонент, чтобы скрыть чувствительную информацию.

Модель машинного обучения №1

- Модель основана на алгоритме Случайного леса;
- Алгоритм настроен на 100 деревьев, глубиной в 10 уровней;
- Результирующий порог классификации был понижен до 0.3;
- Выборка была разделена 30/70 на тестовую и обучающую.



Визуализация модели в Knime

Evalutation Credit Card Fraud Detection Model

Threshold = 0.5

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85289	6	99.99%
1 (Actual)	39	109	73.65%
	99.95%	94.78%	

Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-measure
0	99.99%	99.95%	99.99%	73.65%	99.97%
1	73.65%	94.78%	73.65%	99.99%	82.89%

Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.95%	0.829

Threshold = 0.3

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85279	16	99.98%
1 (Actual)	34	114	77.03%
	99.96%	87.69%	

Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-measure
0	99.98%	99.96%	99.98%	77.03%	99.97%
1	77.03%	87.69%	77.03%	99.98%	82.01%

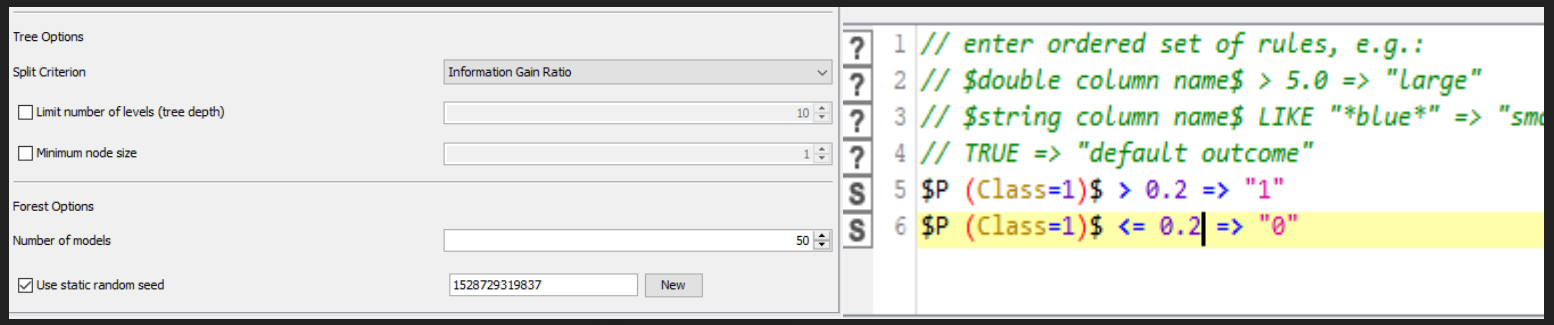
Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.94%	0.820

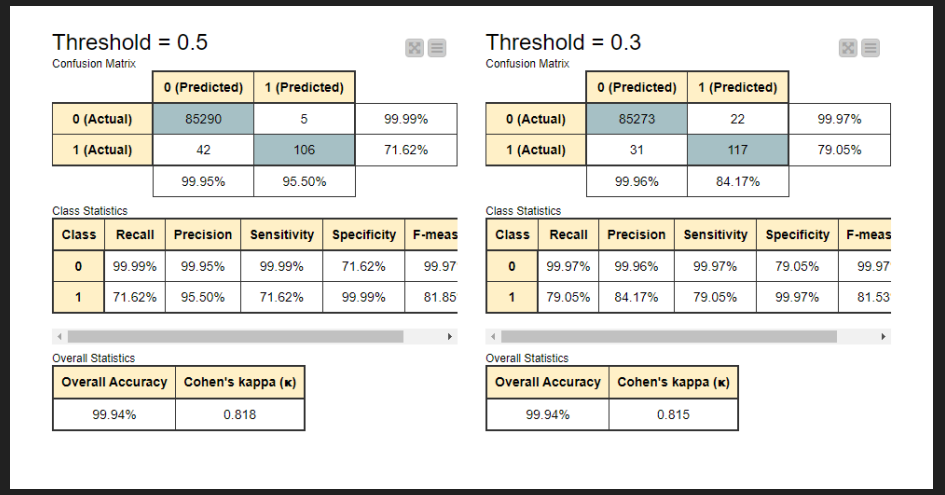
Результаты обучения модели

Модификация модели

- Количество деревьев было уменьшено до 50;
- Обучающая и тестовая выборка была поделена в соотношении 50/50;
- Порог был уменьшен до 0.2.



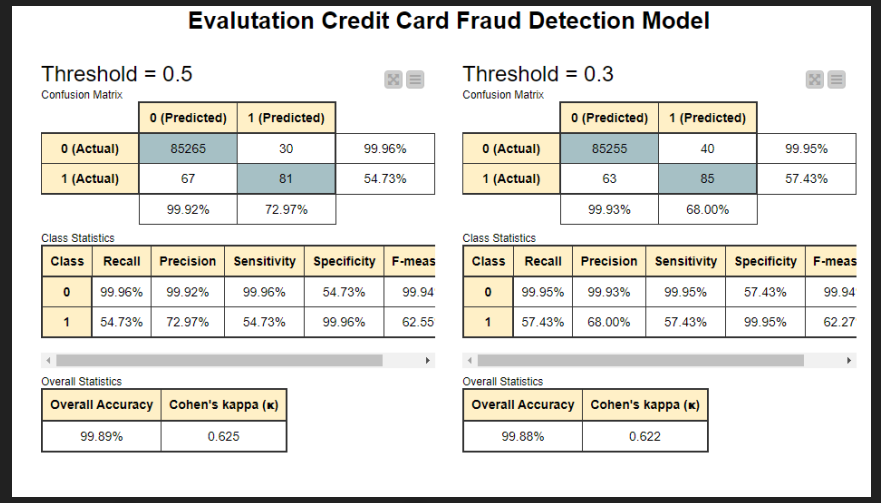
Настройки модели



Результаты обучения модели

Модель машинного обучения №2

- Модель основана на наивном алгоритме Байеса;
- Параметры алгоритма были подобраны итеративно и отображены ниже;
- Результирующий порог классификации был понижен до 0.3;
- Выборка была разделена 30/70 на тестовую и обучающую.



Evalutation Credit Card Fraud Detection Model

Threshold = 0.5

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85265	30	99.96%
1 (Actual)	67	81	54.73%
	99.92%	72.97%	

Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-meas
0	99.96%	99.92%	99.96%	54.73%	99.94
1	54.73%	72.97%	54.73%	99.96%	62.55

Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.89%	0.625

Threshold = 0.3

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85255	40	99.95%
1 (Actual)	63	85	57.43%
	99.93%	68.00%	

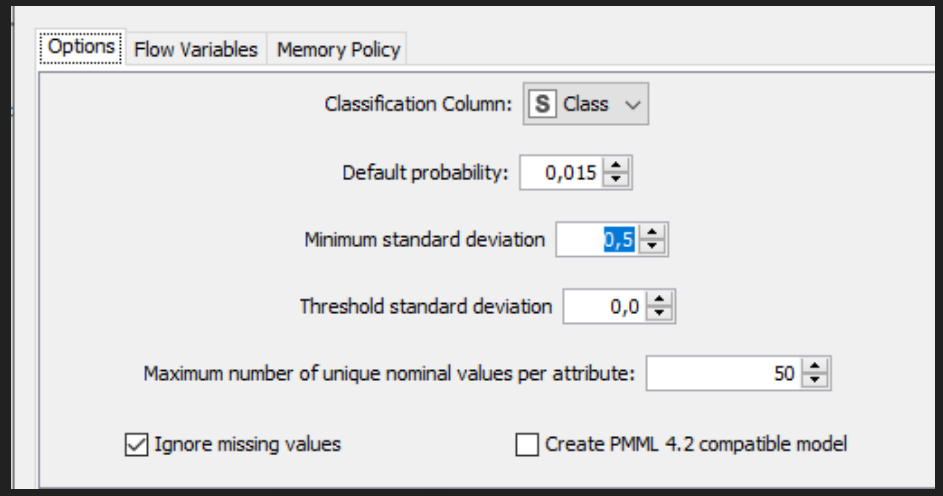
Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-meas
0	99.95%	99.93%	99.95%	57.43%	99.94
1	57.43%	68.00%	57.43%	99.95%	62.27

Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.88%	0.622

Результаты обучения модели



Настройки модели

Модель машинного обучения №3

- Модель основана на наивном алгоритме Деревьев решений;
- Параметры алгоритма представлены ниже;
- Результирующий порог классификации был понижен до 0.3;
- Выборка была разделена 30/70 на тестовую и обучающую.

Evalutation Credit Card Fraud Detection Model

Threshold = 0.5

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85273	22	99.97%
1 (Actual)	51	97	65.54%
	99.94%	81.51%	

Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-meas
0	99.97%	99.94%	99.97%	65.54%	99.96
1	65.54%	81.51%	65.54%	99.97%	72.66

Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.91%	0.726

Threshold = 0.3

Confusion Matrix

	0 (Predicted)	1 (Predicted)	
0 (Actual)	85273	22	99.97%
1 (Actual)	51	97	65.54%
	99.94%	81.51%	

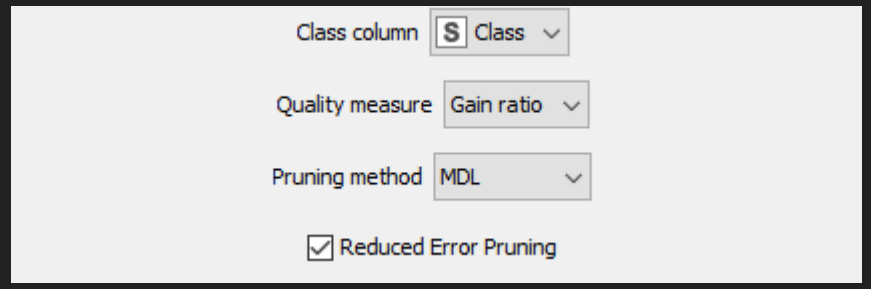
Class Statistics

Class	Recall	Precision	Sensitivity	Specificity	F-meas
0	99.97%	99.94%	99.97%	65.54%	99.96
1	65.54%	81.51%	65.54%	99.97%	72.66

Overall Statistics

Overall Accuracy	Cohen's kappa (κ)
99.91%	0.726

Результаты обучения модели



Настройки модели

Выводы

В результате выполнения данной практической работы был изучен и реализован метод кластерного анализа с применением алгоритма K-средних.