Выявление мошеннических транзакций методами машинного обучения при несбалансированной выборке

Студент: Карпов С.М.

Научный руководитель: д.т.н., профессор Беляков С.Л.

Цели и задачи ВКР

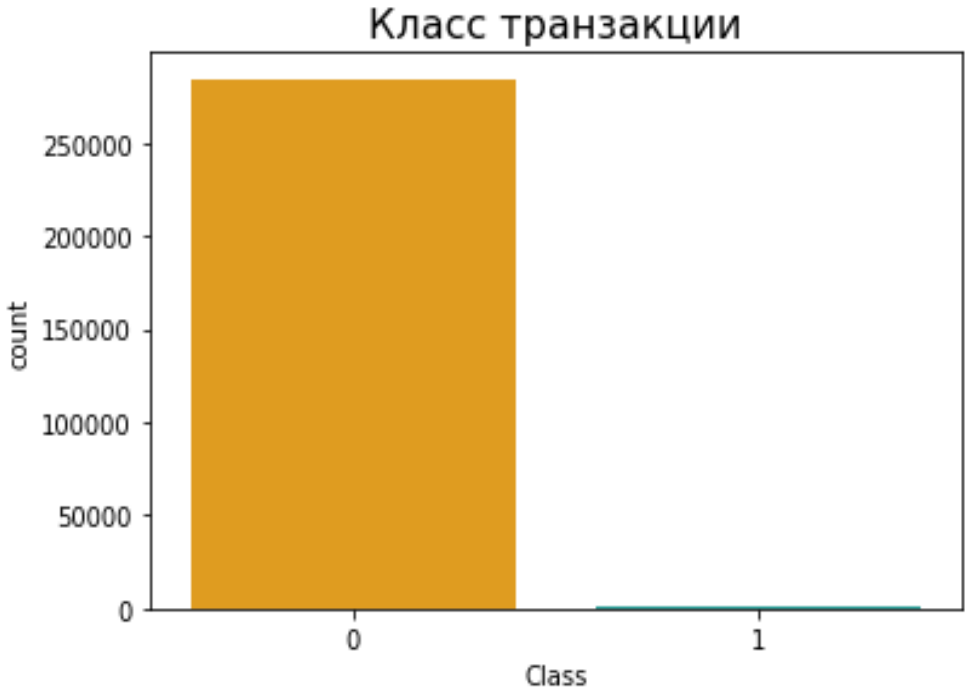
Наименование задачи: Разработка и реализация решения в области выявления мошеннических транзакций на основе несбалансированных данных.

Цель разработки: Эффективное выявление мошеннических транзакций при несбалансированной выборке.

Предъявляемые требования:

- Масштабируемость;
- Эффективная работа в условиях несбалансированных данных;
- Способность точно и оптимально классифицировать как мошеннические, так и честные транзакции.

Исходные данные



31 признак по каждой транзакции:

- 28 признаков со скрытым смыслом;
- Время транзакции;
- Сумма транзакции;
- Класс транзакции.

Соотношение классов транзакций:

Честные транзакции 99.83%

Мошеннические транзакции 0.17%

Матрица путаницы

TP – Мошеннические транзакции которые распознаны как мошеннические.

FN – Мошеннические транзакции которые не распознанные как мошеннические.

TN – Честные транзакции распознанные как честные транзакции.

FP – Честные транзакции распознанные как мошеннические.

Критерии результата

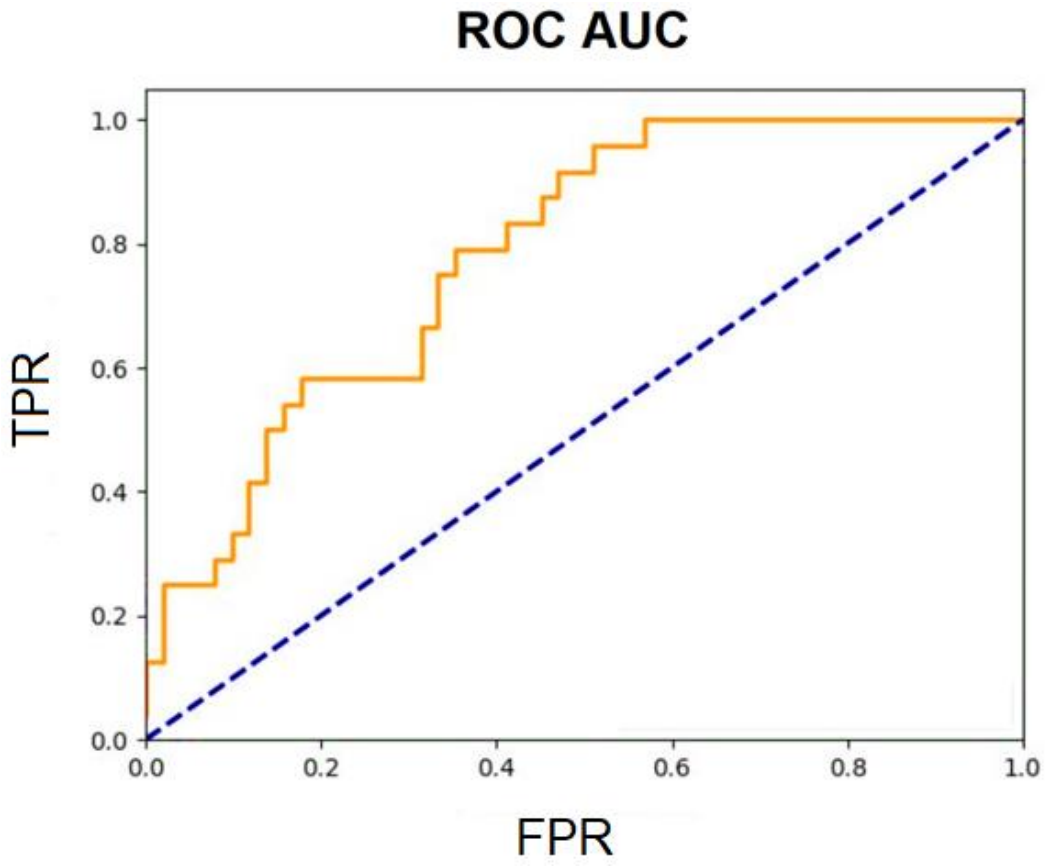
$$recall = \frac{TP}{TP + FN}$$

Доля **истинных** мошеннических транзакций, распознанных моделью **из числа всех истинных** мошеннических транзакций.

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

Доля **истинных** мошеннических транзакций от числа всех транзакций **предсказанных** моделью **как мошеннические**

Критерии результата

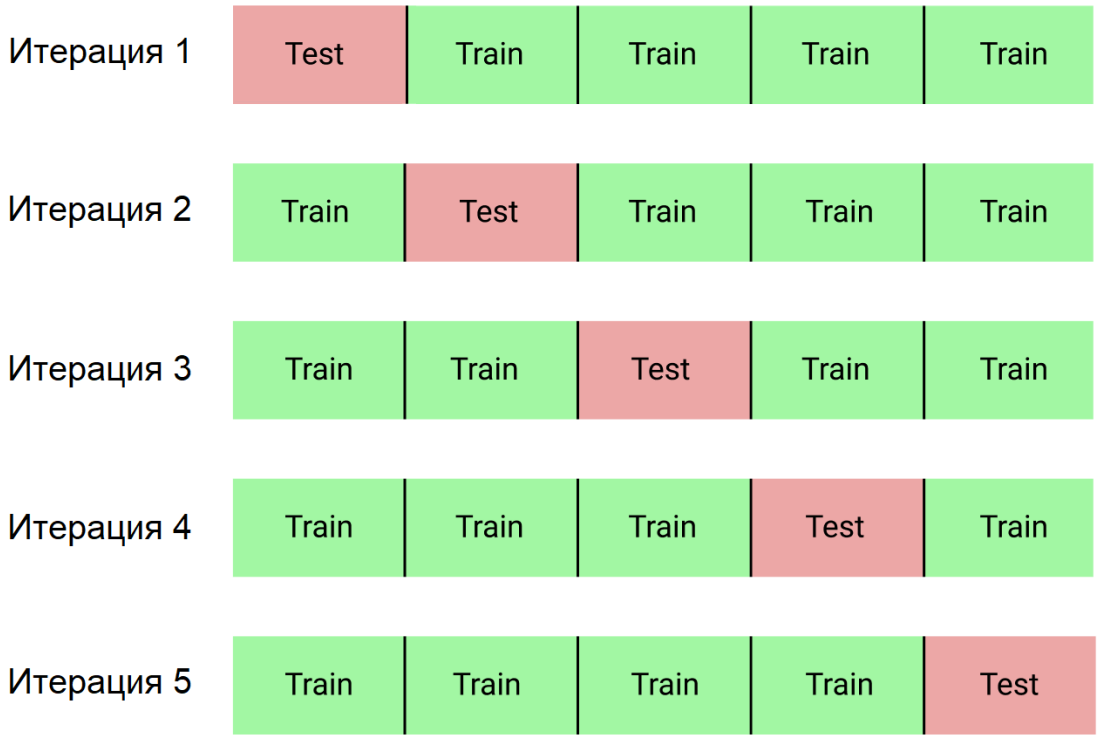


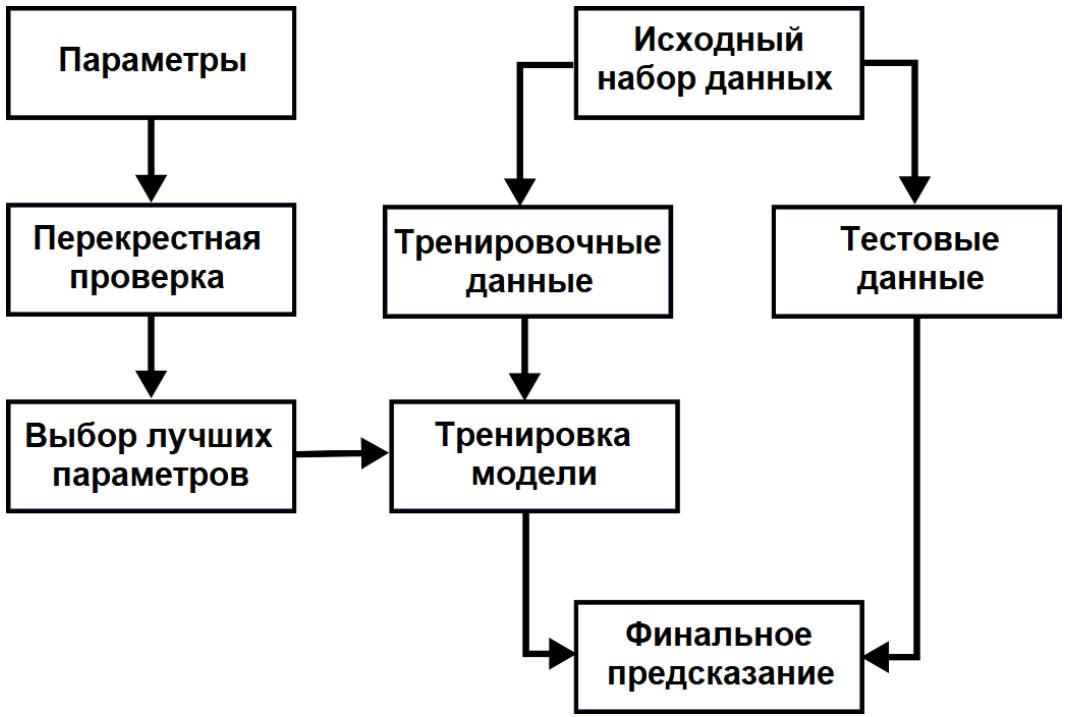
$$\boldsymbol{TRP} = \frac{TP}{TP + FN} = \boldsymbol{recall}$$

$$\boldsymbol{FPR} = \frac{FP}{TN + FP}$$

FPR отражает долю честных транзакций предсказанных неверно.

Перекрестная проверка





Проблема переобучения

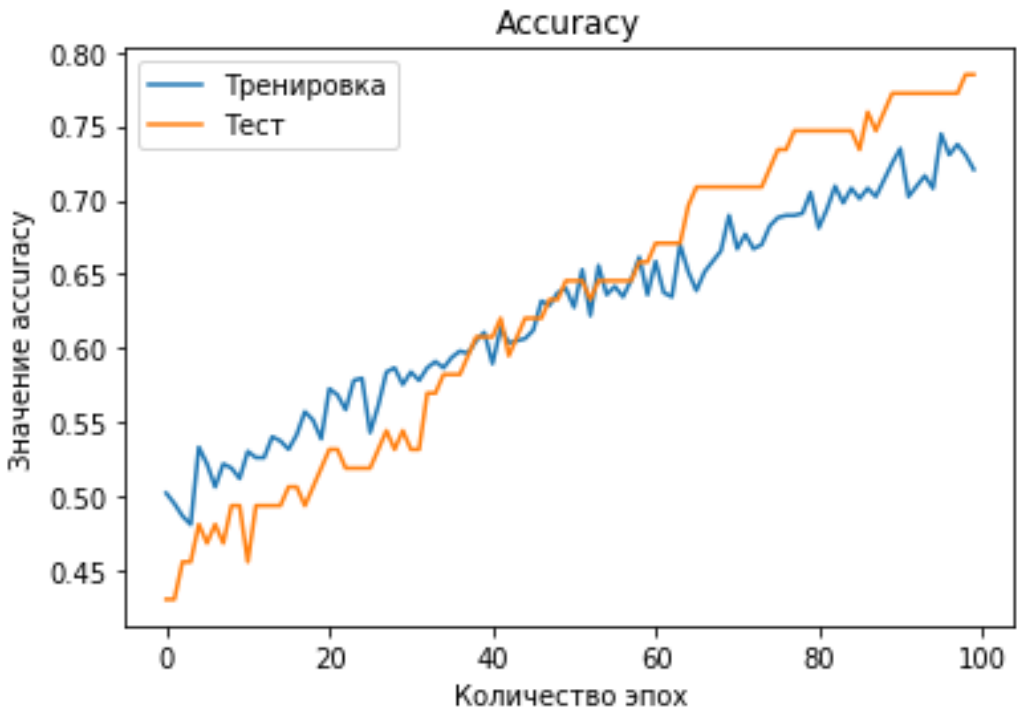


График кривой точности

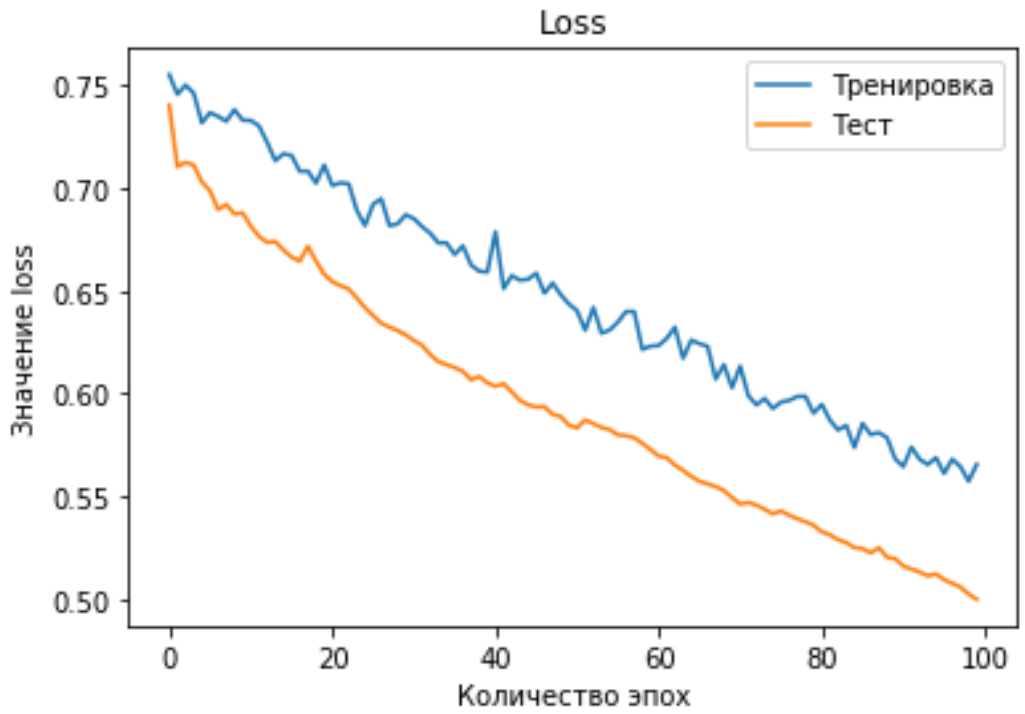


График кривой функции потерь

Используемые модели машинного обучения

- Метод опорных векторов;
- Алгоритм k-ближайших соседей;
- Случайный лес;
- Логистическая регрессия;
- Нейронная сеть типа перцептрон.

Используемые методы передискретизации

Исходная выборка Итоговая выборка

Алгоритмы избыточной выборки:

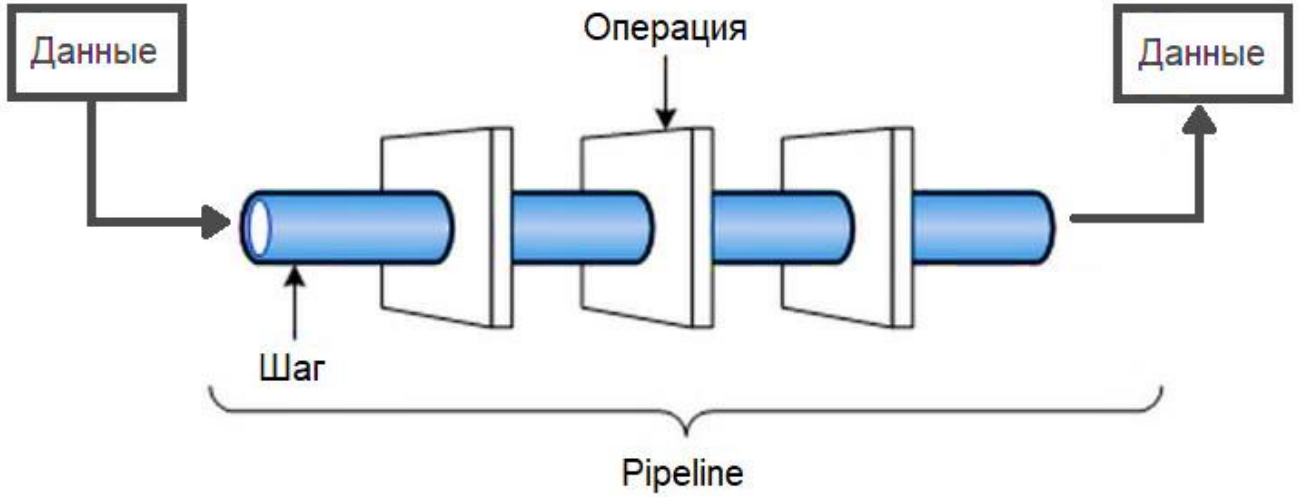
- ADASYN;
- SMOTE.

Исходная выборка Итоговая выборка

Алгоритмы недостаточной выборки:

- NearMiss-1;
- RandomUnderSampling.

Функция тестирования классификатора



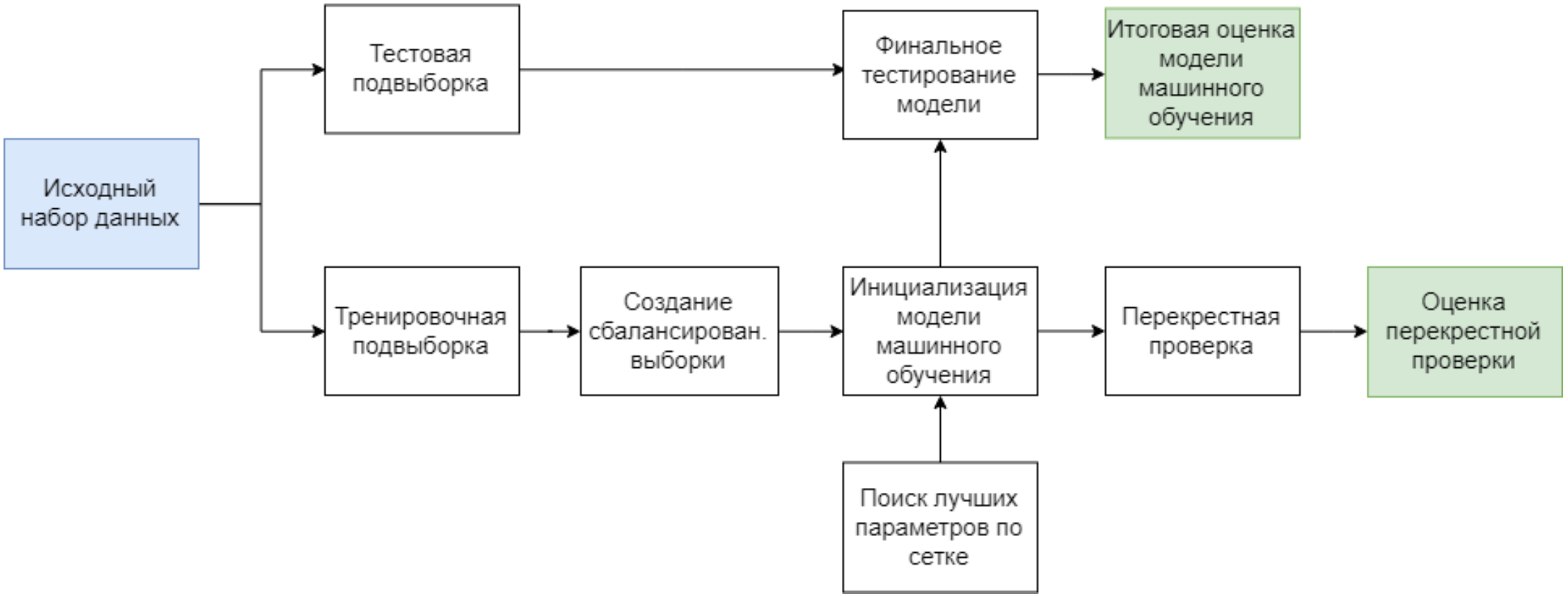
Входные данные:

- Имя классификатора;
- Инициализация классификатора;
- Сетка параметров;
- Инициализация метода создания выборки.

Выходные данные:

- Матрица путаницы;
- Кривая ROC AUC;
- Precision;
- Recall;
- Accuracy;
- F1.

Функциональная схема работы подпрограммы тестирования статистических классификаторов



Функция тестирования нейронной сети

Параметры:

- Многослойный перцептрон;
- 2 скрытых слоя (функция активации Relu);
- Функция активации выходного слоя – Sigmoid;
- Количество эпох – 100;
- Данные валидации – 20% от всего тренировочного набора.

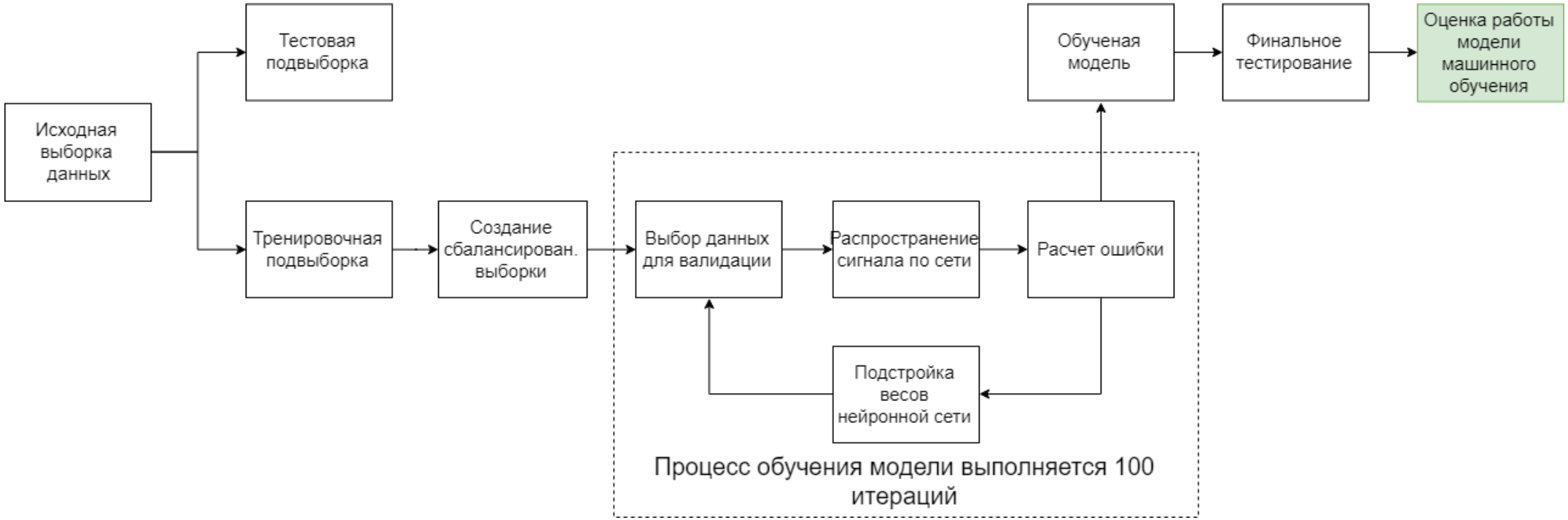
Входные данные:

- Имя метода передискретизации;
- Инициализация метода передискретизации.

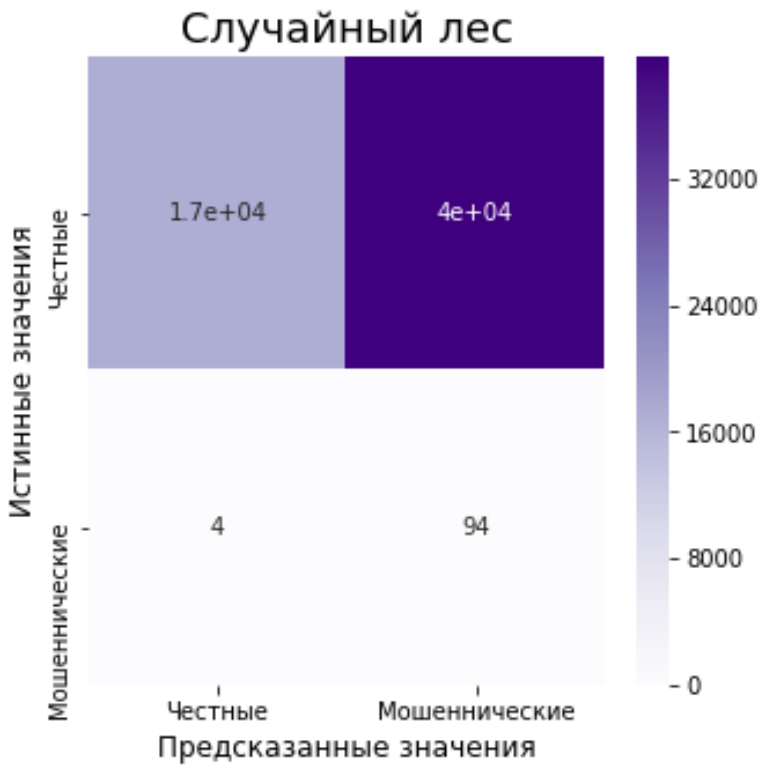
Выходные данные:

- Матрица путаницы;
- График кривой точности;
- График кривой функции потерь.

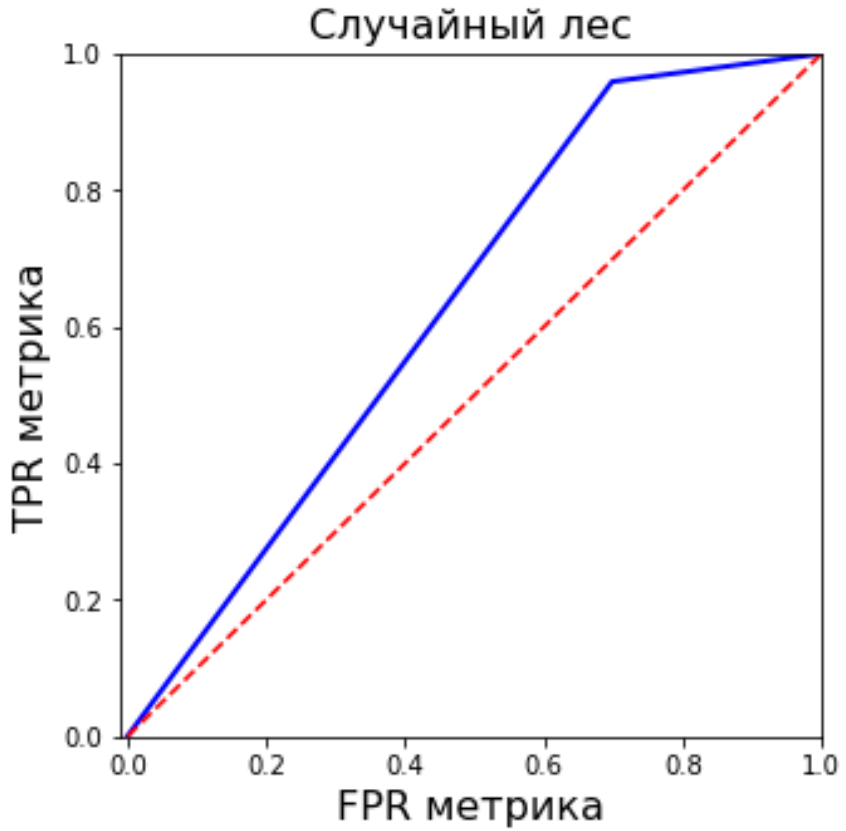
Функциональная схема работы подпрограммы тестирования нейронной сети



Наилучший показатель метрики **recall** среди всех алгоритмов

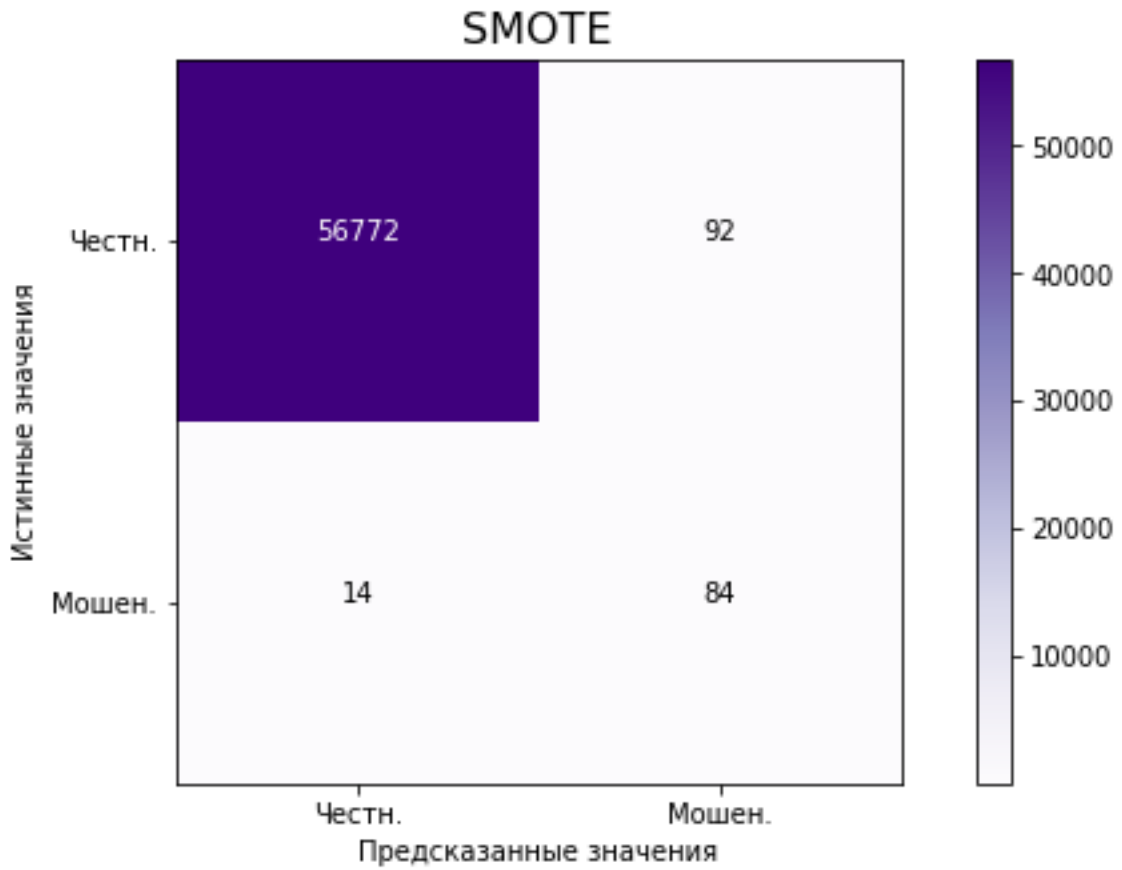



```
_____________________________________
Случайный лес

_____________________________________
Результат кроссвалидации:
accuracy:22.191 %
precision:0.217 %
recall:96.076 %
f1:0.432 %
Roc Auc:59.070 %

_____________________________________

_____________________________________
Результат тестирования:
accuracy:30.280 %
precision:0.236 %
recall:95.918 %
f1:0.471 %
Roc Auc:63.043 %

_____________________________________
```



Recall = 95.92%

Наилучший показатель метрики **precision** среди всех алгоритмов




```
Функция потерь: 0.7617562985137819 %
Точность (accuracy): 83.7997210284049 %
```

Precision = 47.72%

Recall = 85.71%

Кривые нейронной сети

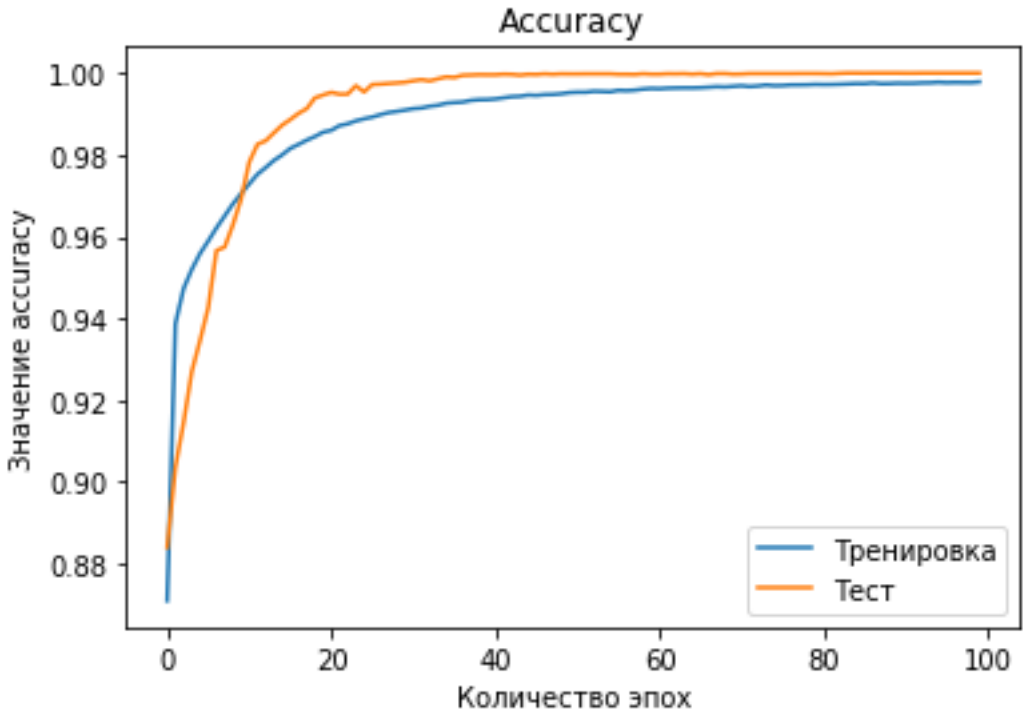


График кривой точности

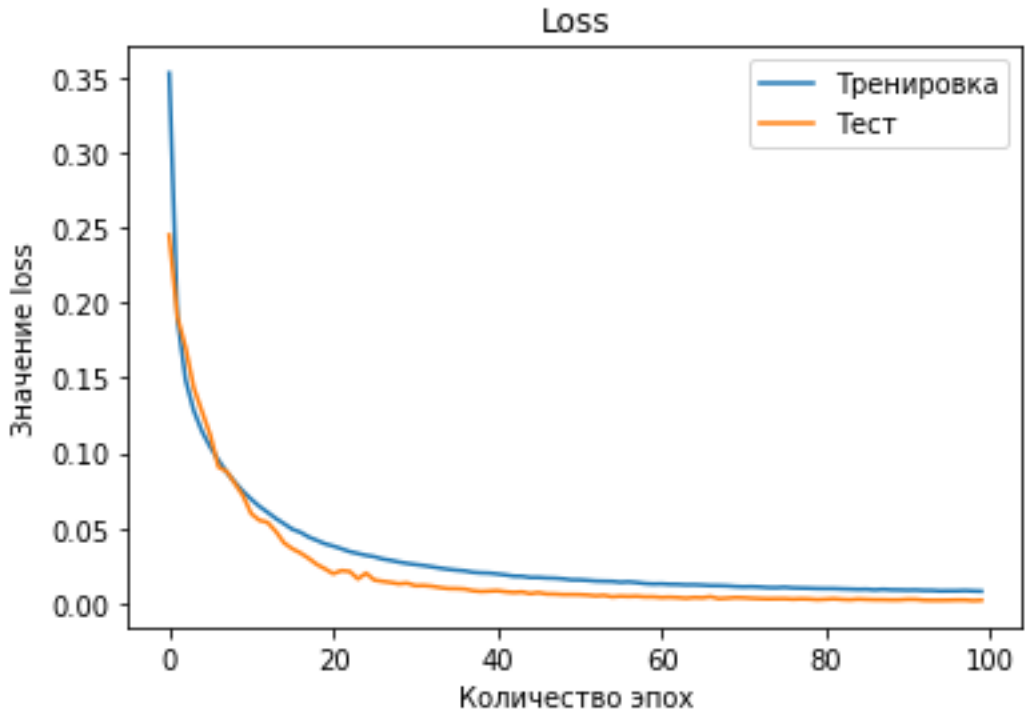


График кривой функции потерь

Итоговая таблица результатов

Наименование модели машинного обучения	Random Under Sampler		NearMiss		SMOTE		ADASYN	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Метод опорных векторов	6.78%	91.84%	0.19%	93.88%	4.64%	85.71%	5.15%	93.88%
К-ближайших соседей	6.66%	88.78%	0.37%	89.79%	4.84%	82.65%	12.08%	90.82%
Случайный лес	6.64%	89.79%	0.24%	**95.92%**	1.89%	87.75%	9.18%	88.78%
Логистическая регрессия	10.17%	87.76%	0.49%	92.86%	6.65%	89.80%	1.74%	93.88%
Нейронная сеть	0.44%	75.51%	0.28%	62.24%	**47.72%**	85.71%	41.43%	88.78%