Разработка программной системы машинного обучения для прогнозирования оттока клиентов

Студент: Москалев Д.И.

Руководитель: Тракимус Ю.В., к.т.н., доцент кафедры ПМт

ЦЕЛИ И ЗАДАЧИ РАБОТЫ

Цели:

- Создание моделей прогнозирования оттока клиентов в банковской сфере, используя методы классификации
- Проведение сравнительного анализа точности полученных результатов с помощью визуализации метрик качества

Задачи:

- Поиск и обработка наборов больших данных (Big Data)
- Определение возможной лояльности клиентов по анализу входных данных
- Адаптация алгоритмов и построение моделей
- Сравнительный анализ полученных результатов
- Интерпретация результатов с помощью визуального представления метрик полученных моделей

ЗАЧЕМ НУЖНО ПРОГНОЗИРОВАНИЕ?

Прогнозирование позволяет:

- Анализировать текущее состояние организации
- Получать рекомендации по сегментации клиентов и продуктов банка

Соотношение клиентов в банках

ПОСТАНОВКА ЗАДАЧИ

- Формальная постановка задачи классификации представляет собой неизвестную целевую зависимость

$$y^*: X \to Y, \quad (1)$$

где X – множество описаний объектов, Y – конечное множество номеров классов.

При этом значения отображения (1) известны только на объектах обучающей выборки:

$$X^n = \{(x_1, y_1), \dots, (x_n, y_n)\},$$

где n – количество строк объектов.

В задаче бинарной классификации допустимое множество номеров классов $Y = \{f_1, f_2\}$. Обычно $f_1 = 0, f_2 = 1$.

ВЫБРАННЫЕ МЕТОДЫ РЕШЕНИЯ

Все выбранные методы основаны на деревьях принятия решений. Деревья представляют собой набор узлов, которые можно разделить на два типа:

- Узлы решения – признаки, по которым строится дерево.
- Вероятностные (замыкающие) узлы – листья деревьев, в которых вычисляются промежуточные или окончательные значения признаков.

1. Бустинг (Boosting) – технология последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов.

- Методы: *XGBoost (Extreme Gradient Boosting), CatBoost (Categorical Boosting).*

2. Бэггинг (Bagging) - технология классификации, где все элементарные классификаторы вычисляются параллельно перед построением деревьев решений.

Метод: *Random Forest (Случайный Лес).*

ВХОДНЫЕ ДАННЫЕ

Входные данные представлены матрицей $N \times M$, где N – клиенты банка, а M – их признаковое описание.

Банк A: 45211 клиентов.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Банк B: 10000 клиентов.

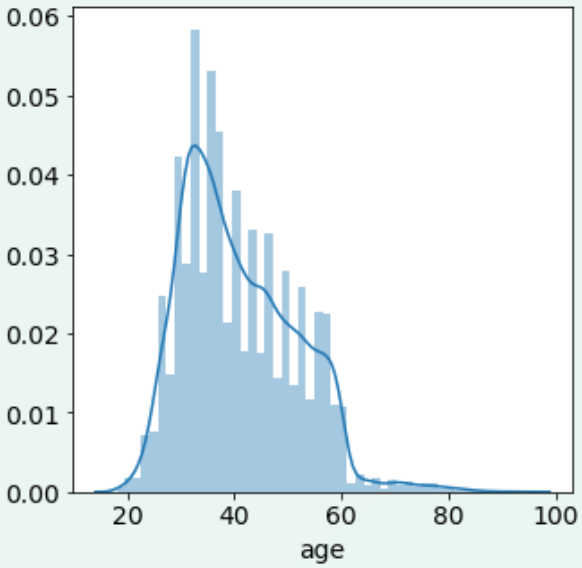
CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

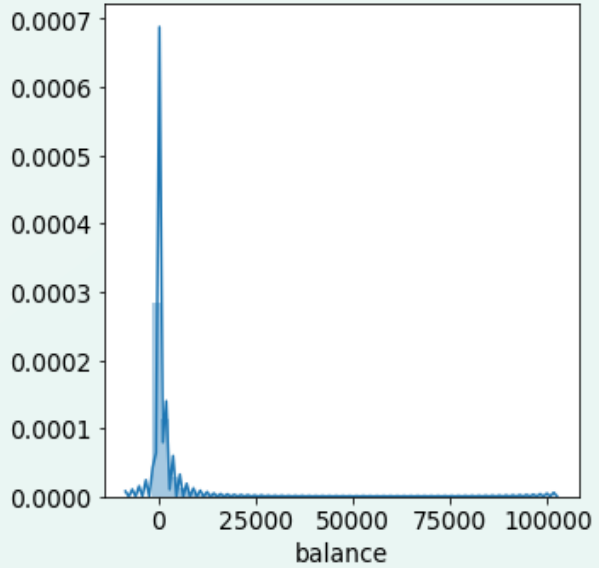
Признаки представляют собой данные числового и категориального типов. Целевым признаком для банка A является столбец *“y”*, для банка B – *“Exited”*.

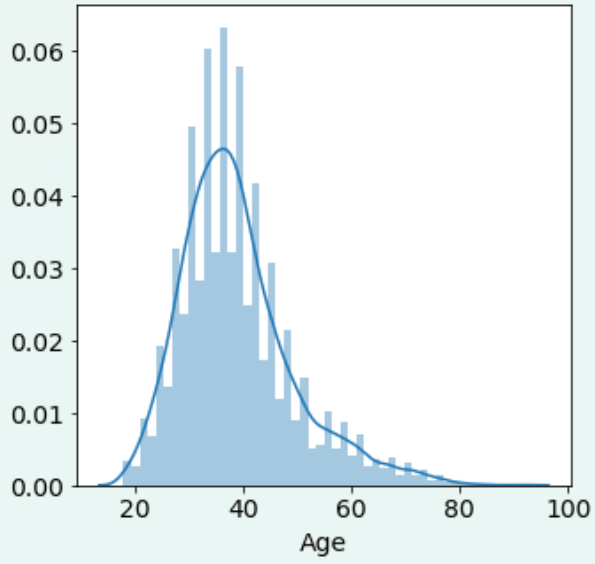
РАСПРЕДЕЛЕНИЕ ПРИЗНАКОВ

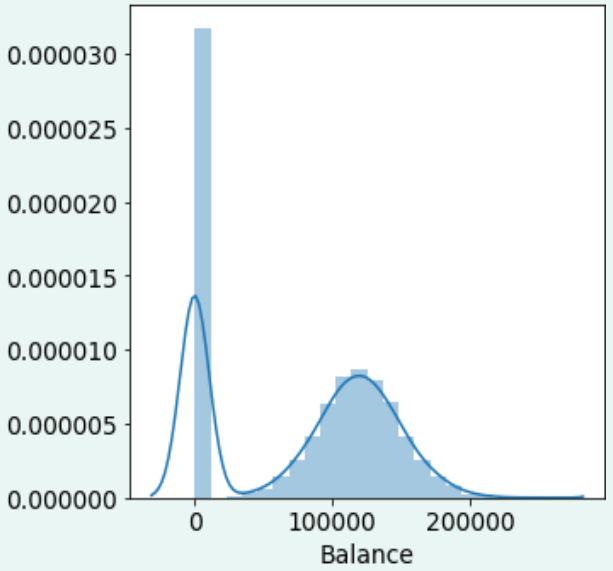
Распределение признаков показывает долю каждого уникального значения признака в наборах данных и является визуальным показателем сбалансированности.
По оси **OX** откладывается признак, по оси **OY** – распределение признака в выборке.
$\sum_{i=1}^{N} Feature_p = 1, p \in \overline{1, M}.$

Признаки:
Age – возраст клиента,
Balance – баланс.

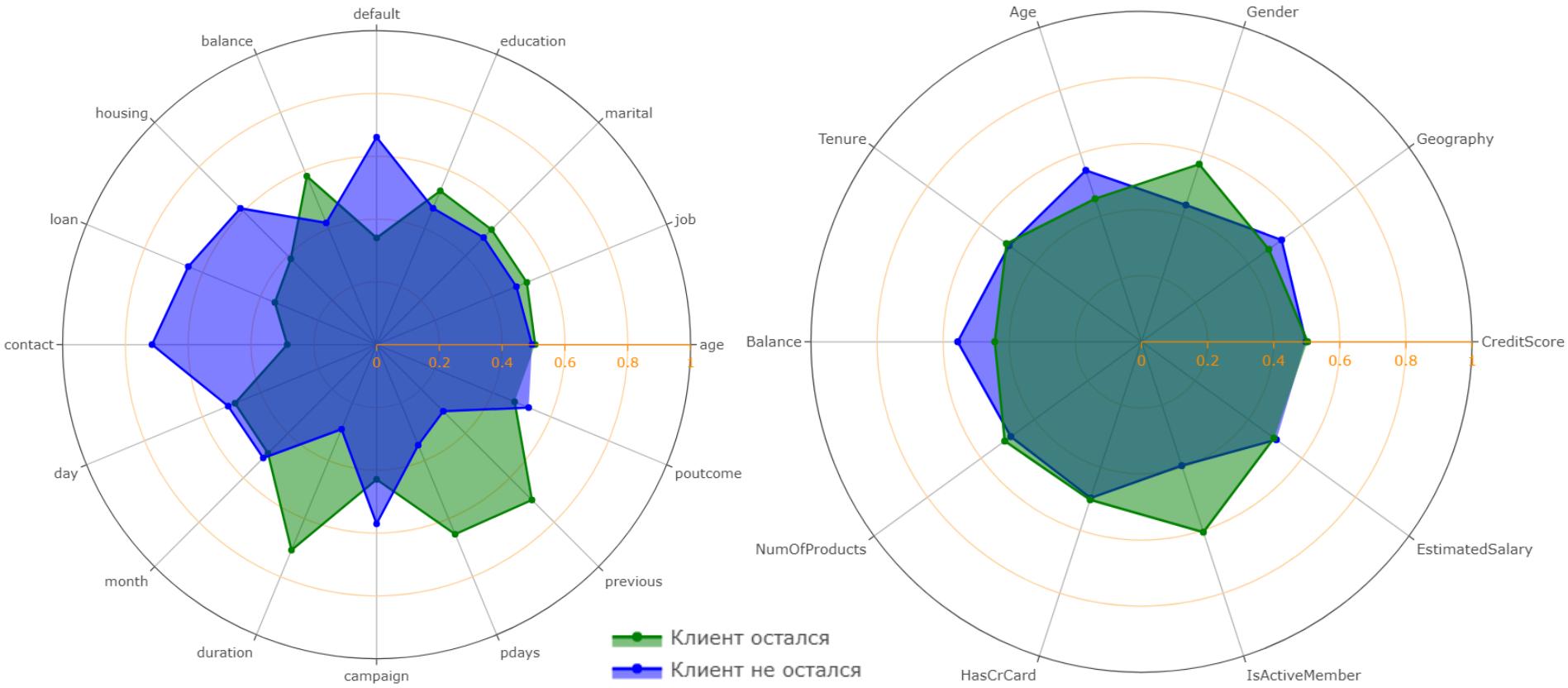








СРЕДНИЕ ЗНАЧЕНИЯ ПРИЗНАКОВ

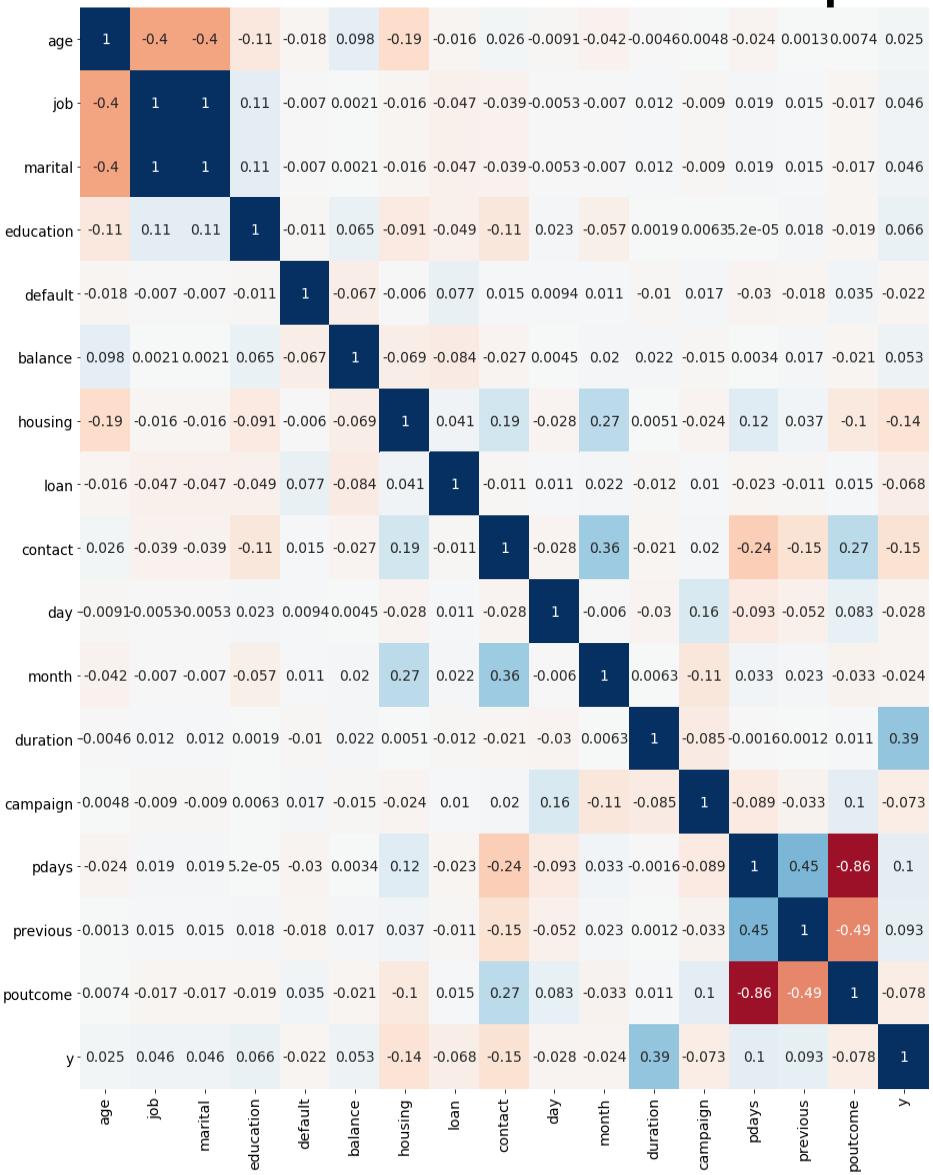


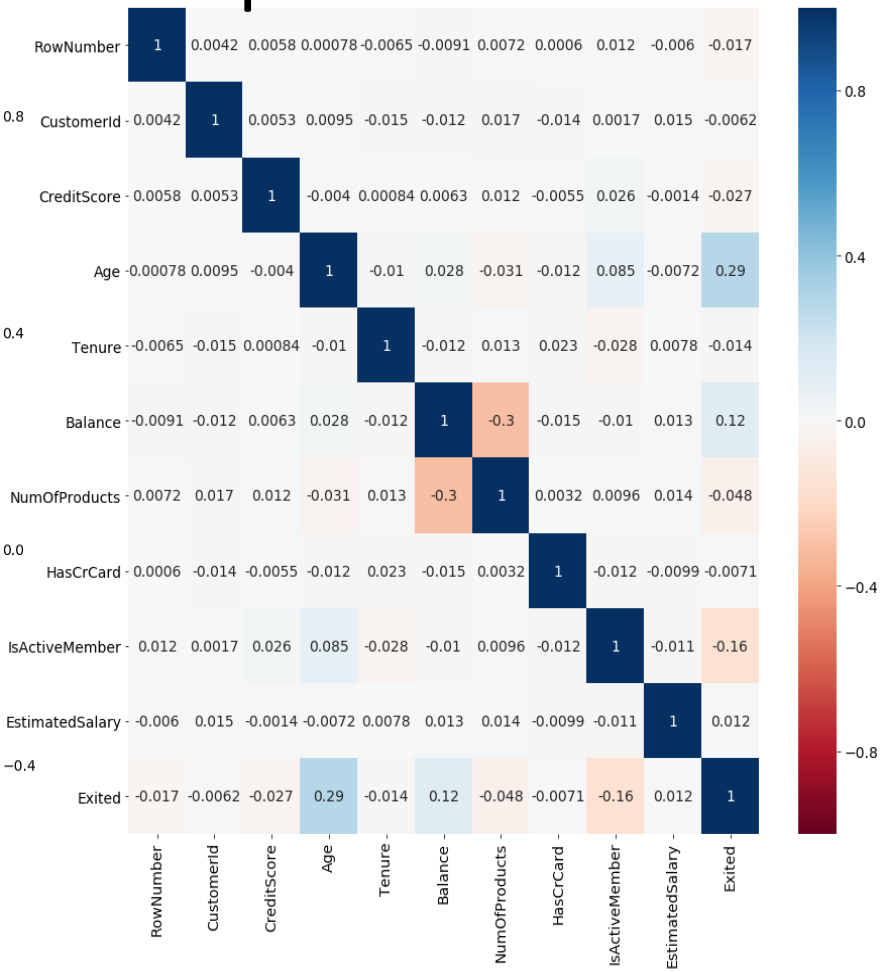
Значения признаков можно рассчитать по формуле:

$$Feature_{mean_p} = \frac{Feature_{mean_i}}{Feature_{mean_0} + Feature_{mean_1}}, i = \{0,1\}, p = \overline{1, M}.$$

Из диаграмм видно, что больше различаются признаки в банке А и для него легче выбрать порог отсечения.

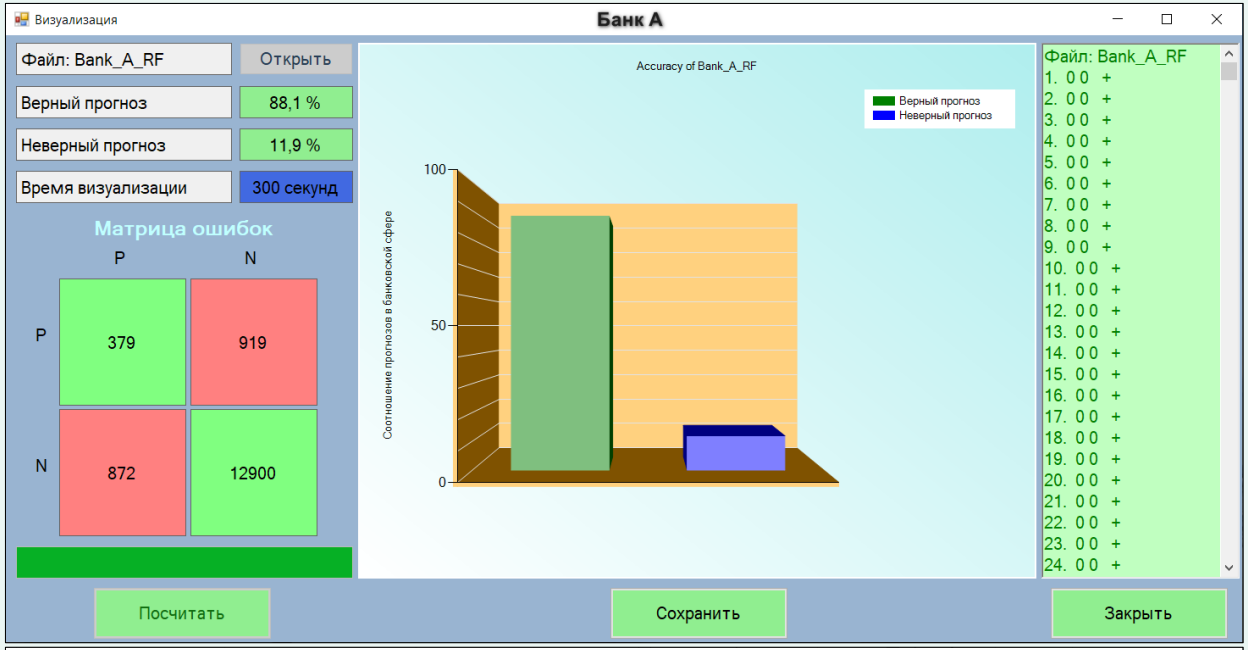
МАТРИЦЫ КОРРЕЛЯЦИИ

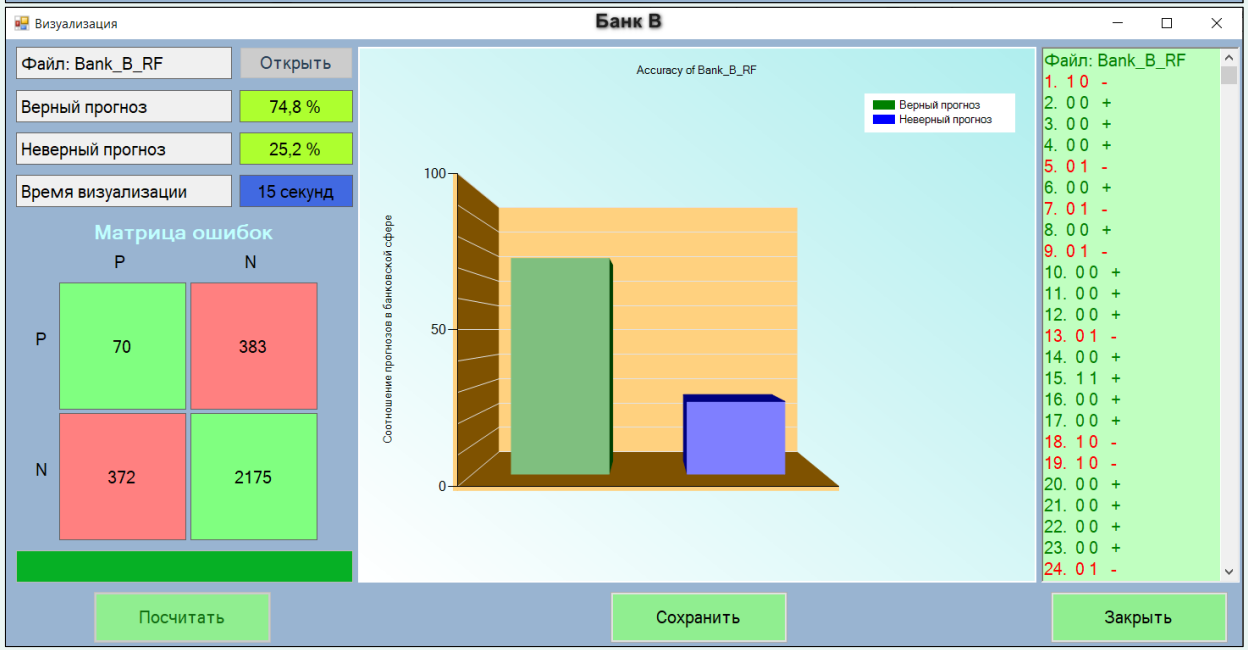




Синий цвет означает строгую прямую связь, белый – отсутствие связи, красный – строгую обратную связь.

ВИЗУАЛИЗАЦИЯ

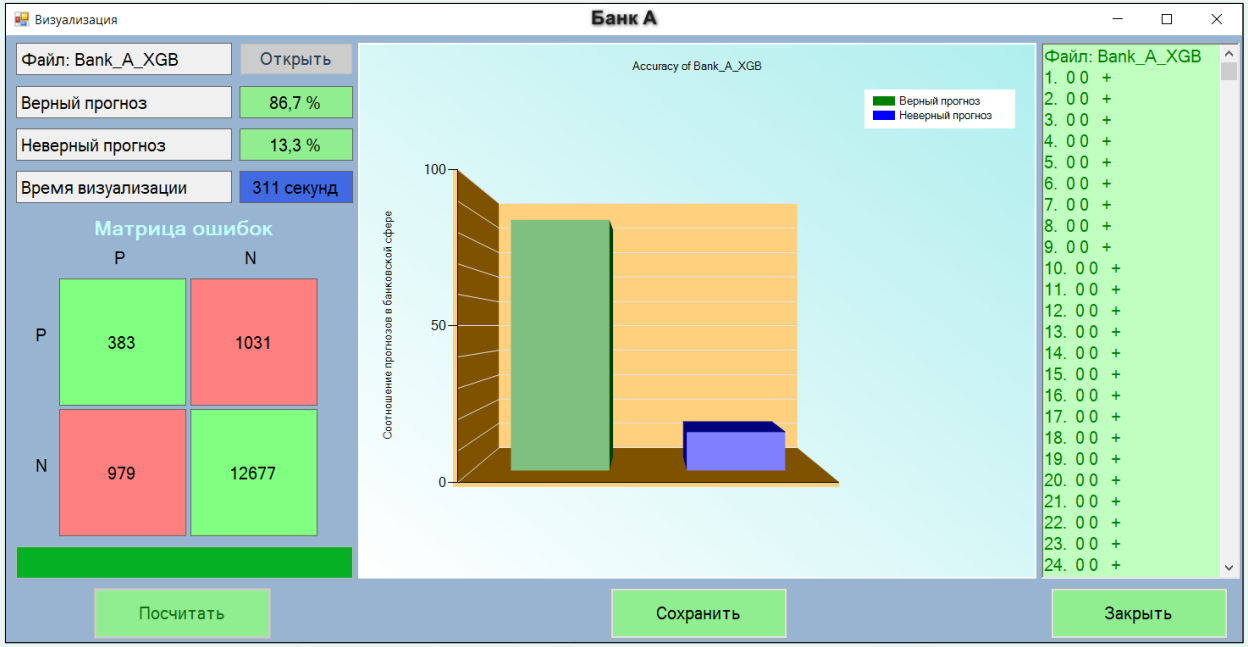


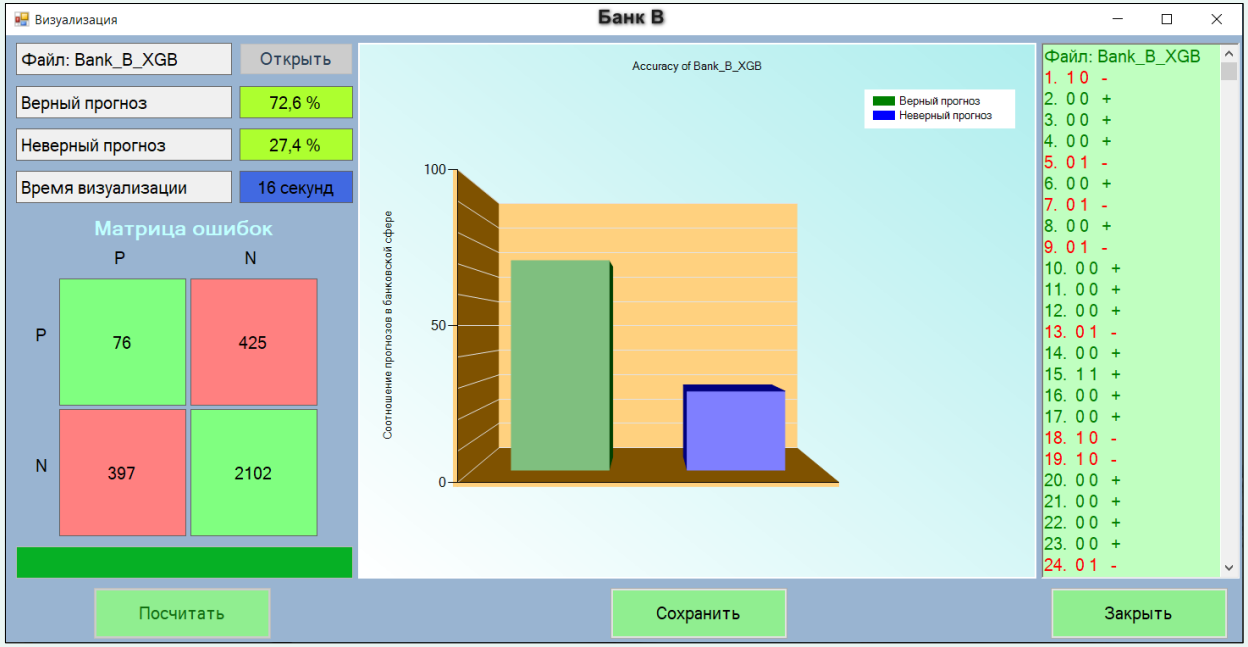


Результаты визуализации моделей, обученных с помощью метода Random Forest (случайный лес). Значение метрики точности Accuracy (верный прогноз): Банк А: 88,1%, Банк В: 74,8%.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \bullet 100\% .$$

ВИЗУАЛИЗАЦИЯ

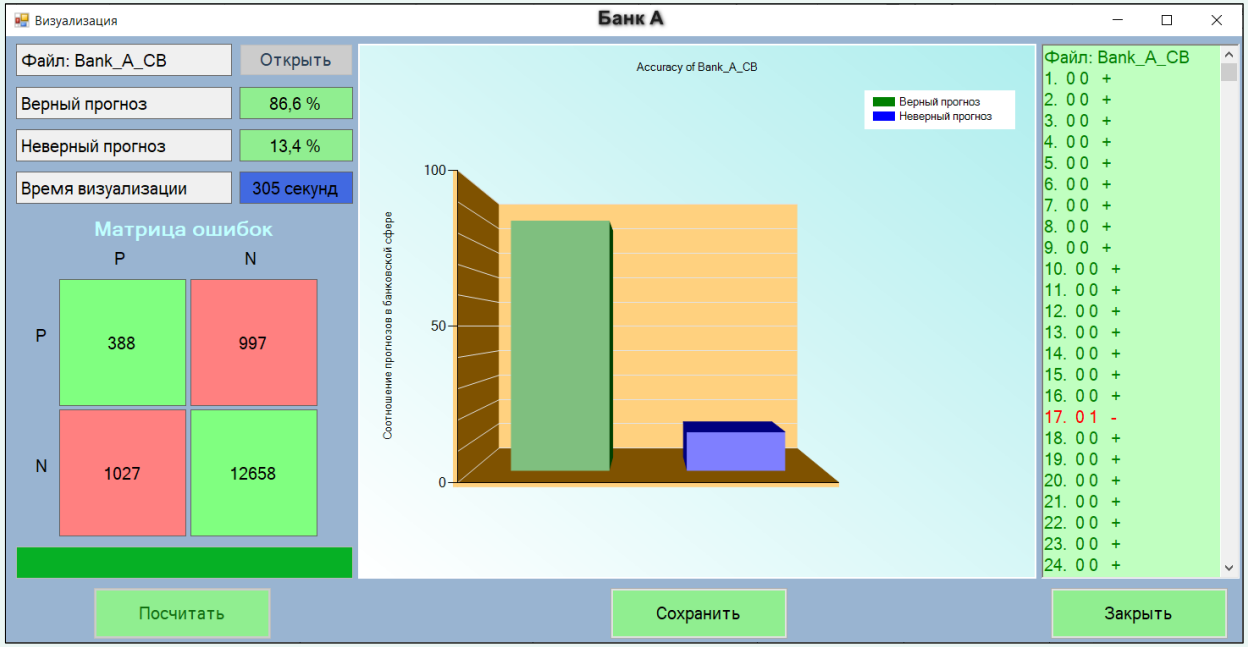


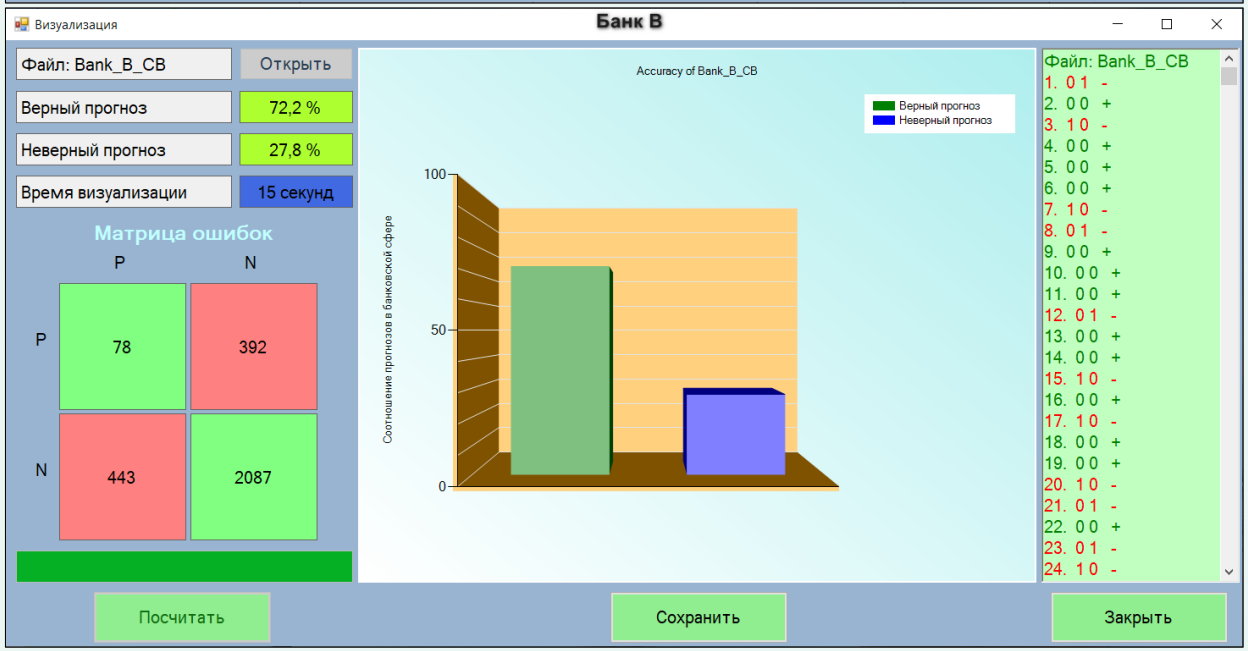


Результаты визуализации моделей, обученных с помощью метода XGBoost (бустинг). Значение метрики точности Accuracy (верный прогноз): Банк А: 86,7%, Банк В: 72,6%.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \bullet 100\% .$$

ВИЗУАЛИЗАЦИЯ





Результаты визуализации моделей, обученных с помощью метода CatBoost (бустинг). Значение метрики точности Accuracy (верный прогноз): Банк А: 86,6%, Банк В: 72,2%.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \bullet 100\% .$$

ЗАКЛЮЧЕНИЕ

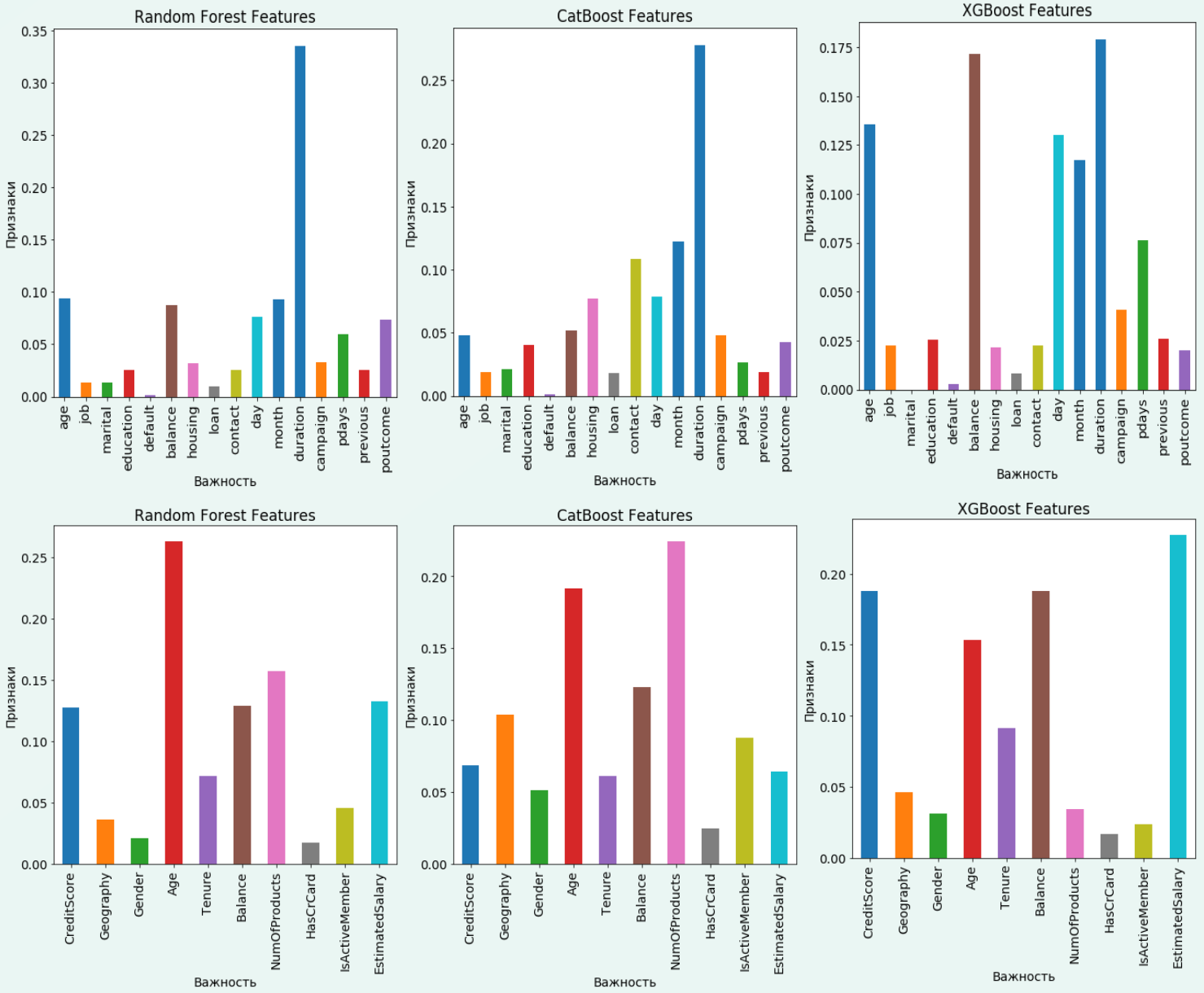
- Сравнительный анализ методов показал, что все классификаторы имеют хорошую точность в пределах от 72,2 до 88,1 процентов, что доказывает их эффективность для решения задачи классификации машинного обучения. Лучшие результаты на тестовой выборке показал метод *Random Forest (88,1%),* на втором месте – *XGBoost* и на третьем – классификатор *CatBoost* от компании Яндекс
- Время работы метода *Random Forest* оказалось наименьшим за счет возможности параллельного построения деревьев решений
- Метод Random Forest (случайный лес) можно применять во всех сферах прогнозирования оттока клиентов

ГРАФИКИ ВАЖНОСТИ ПРИЗНАКОВ

График важности признаков показывает их влияние на построение модели.
По оси **OX** откладываются признаки, по оси **OY** – вероятностное значение признака.

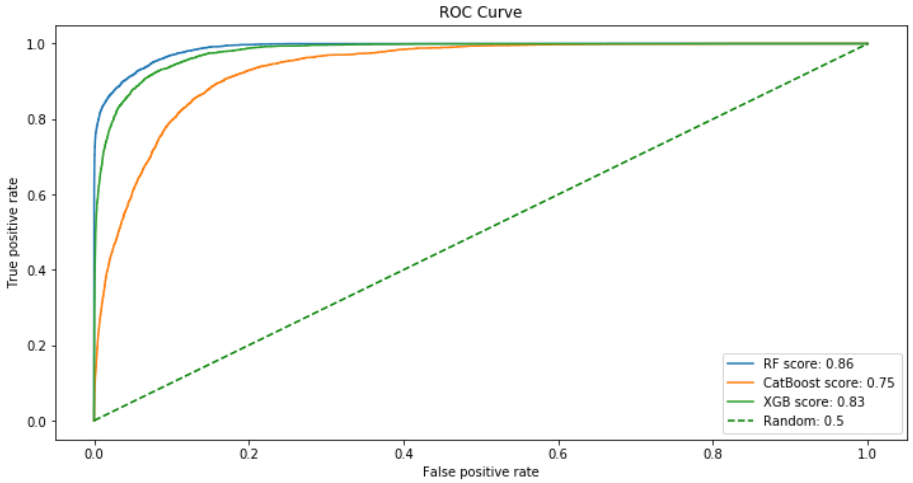
$$\sum_{i=1}^{N} feature_p = 1,$$

где $p \in \overline{1, M}$.

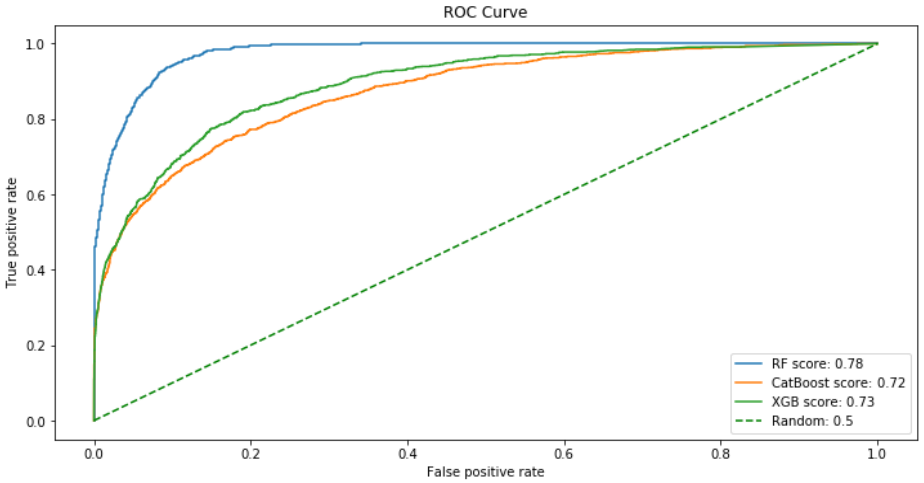


МЕТРИКИ

Банк А



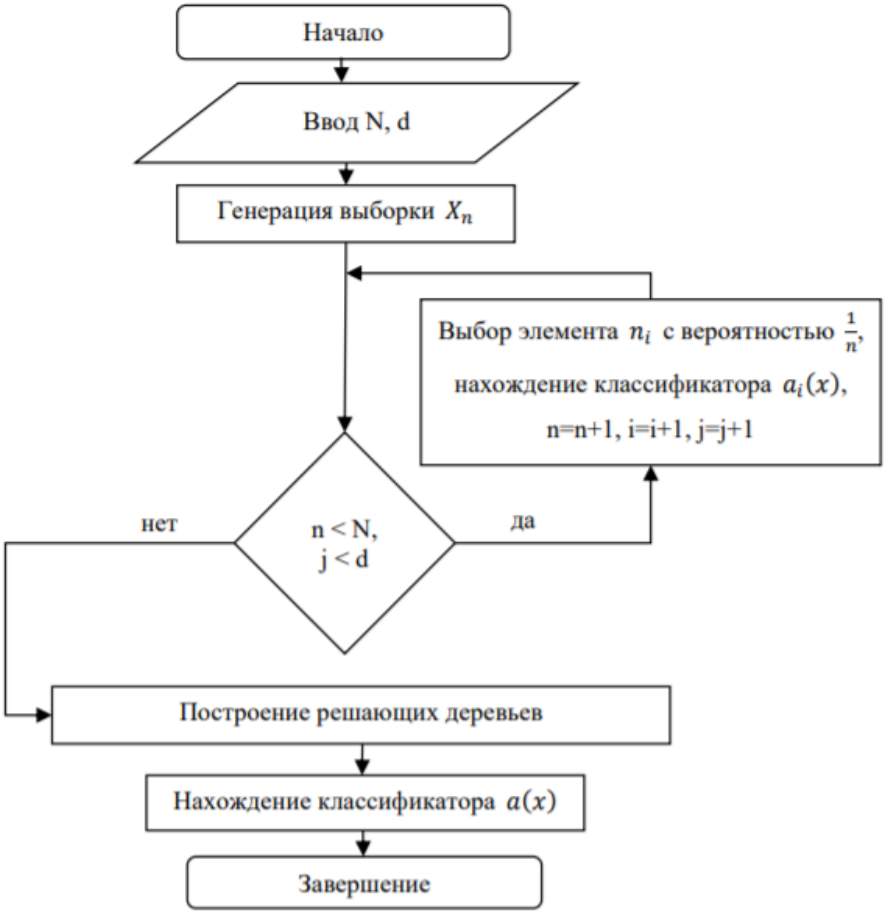
Банк В

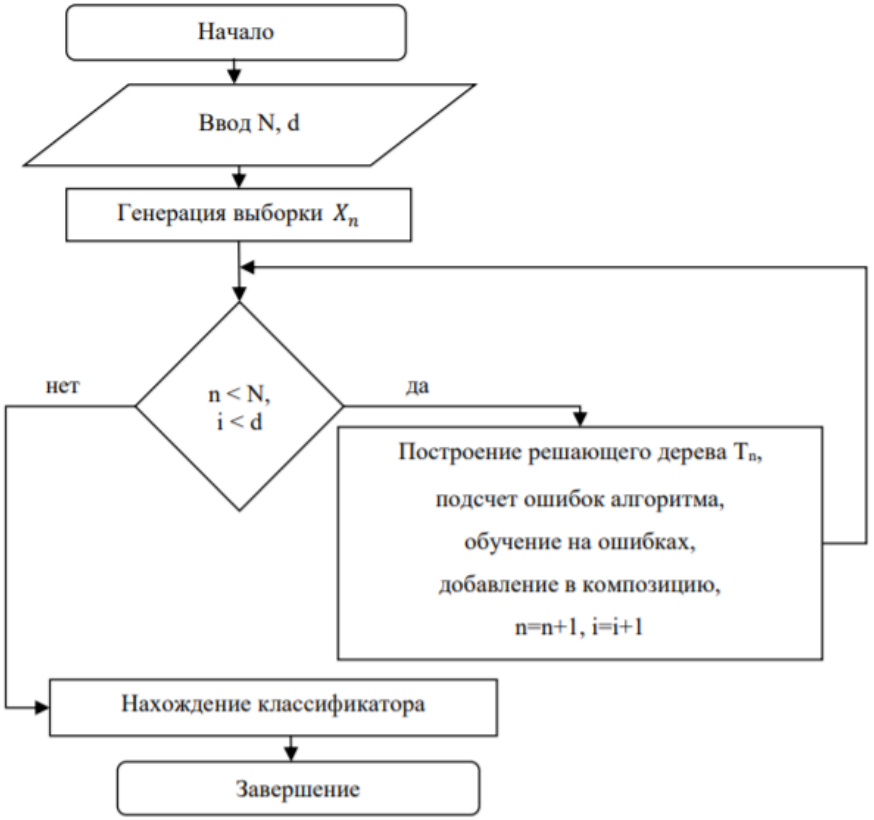


$$FPR = \frac{FP}{N}, \mathrm{TPR} = \frac{TP}{P}$$

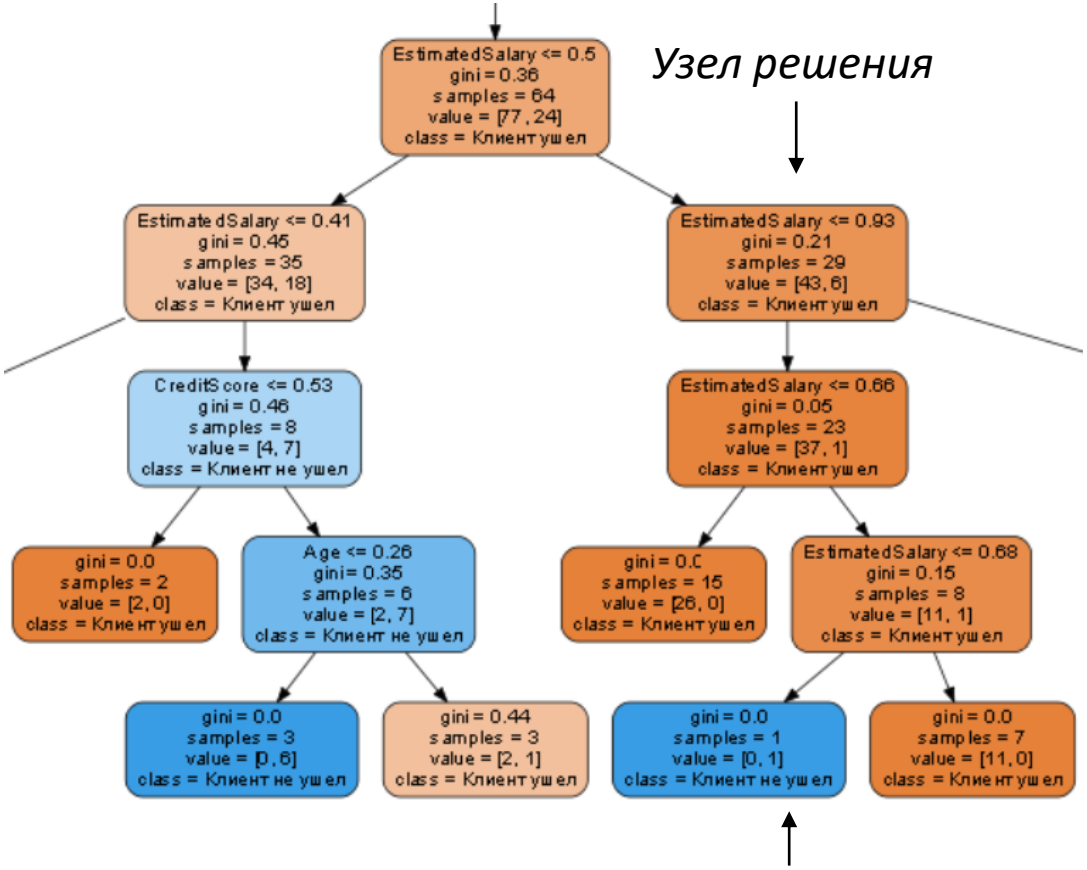
Алгоритм построения ROC-кривой базируется на последовательном соединении точек сетки $m \times n$, где m – число единиц, n – число нулей по следующим условиям:
Если значения целевых признаков совпали (прогноз верный), то происходит сдвиг вверх на одну единицу тестовой выборки, если значения целевых признаков не совпали (неверный прогноз), то происходит сдвиг вправо на одну единицу.
Суммарное выполнение m шагов вверх и n шагов вправо позволяет прийти в точку (1,1).
Цель: $S_{max} \to 1$.
Более гладкие линии на графике банка А обусловлены тем, что тестовая выборка имеет матрицу большей размерности для построения AUC ROC метрики.

АЛГОРИТМ ПОСТРОЕНИЯ ДЕРЕВЬЕВ РЕШЕНИЙ





Визуализация фрагмента дерева решений



Узел решения

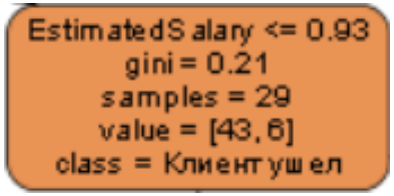
Вероятностный узел

Цель построения деревьев – минимизация критерия вероятности неправильной классификации:

$$Gini = 1 - \sum_{1}^{j} k_j^2 ,$$

где k_j – вероятность класса j.

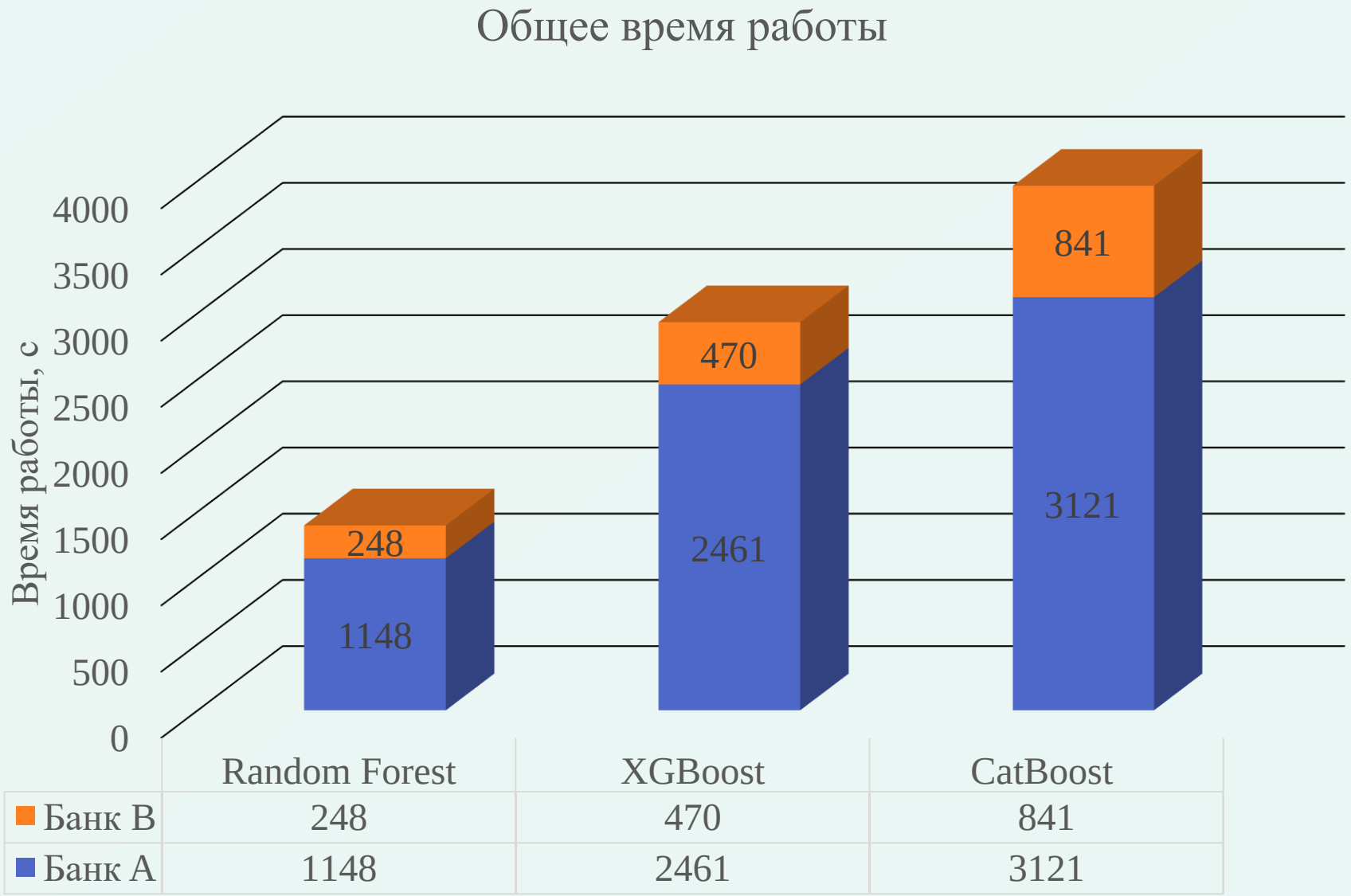
В бинарной классификации существует два класса $\Rightarrow$ $j = 2$.



Пример вычисления Gini для узла решения “$EstimatedSalary$ <= 0.93”:

$$Gini = 1 - (k_1^2 + k_2^2) = 1 - \left(\left(\frac{43}{49}\right)^2 + \left(\frac{6}{49}\right)^2\right) = 1 - \frac{1849 + 36}{2401} \sim 1 - 0{,}785 \sim 0{,}21.$$

Общее время построения и визуализации



	Random Forest	XGBoost	CatBoost
Банк B	248	470	841
Банк A	1148	2461	3121