МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

КУРСОВАЯ РАБОТА
по дисциплине
Прогнозно-аналитические системы
Тема курсовой работы:
«Разработка сценария анализа и обработки данных на примере задачи оттока клиентов банка с применением логистической модели»

Выполнил студент группы ИМБО-02-22 Ким Кирилл Сергеевич

Руководитель курсовой работы Юрченков Иван Александрович

Введение

Цель: разработать сценарии анализа и обработать данные для прогнозирования оттока клиентов банка с использованием логистической регрессии.

Задачи:

- Исследовать данные о клиентах банка и выявить факторы, влияющие на отток
- Провести предобработку данных и создать новые признаки
- Построить и оценить модель логистической регрессии
- Проанализировать важность признаков и сделать выводы
- Разработать систему скоринга для сегментации клиентов по риску оттока
- Предложить практические рекомендации для банка

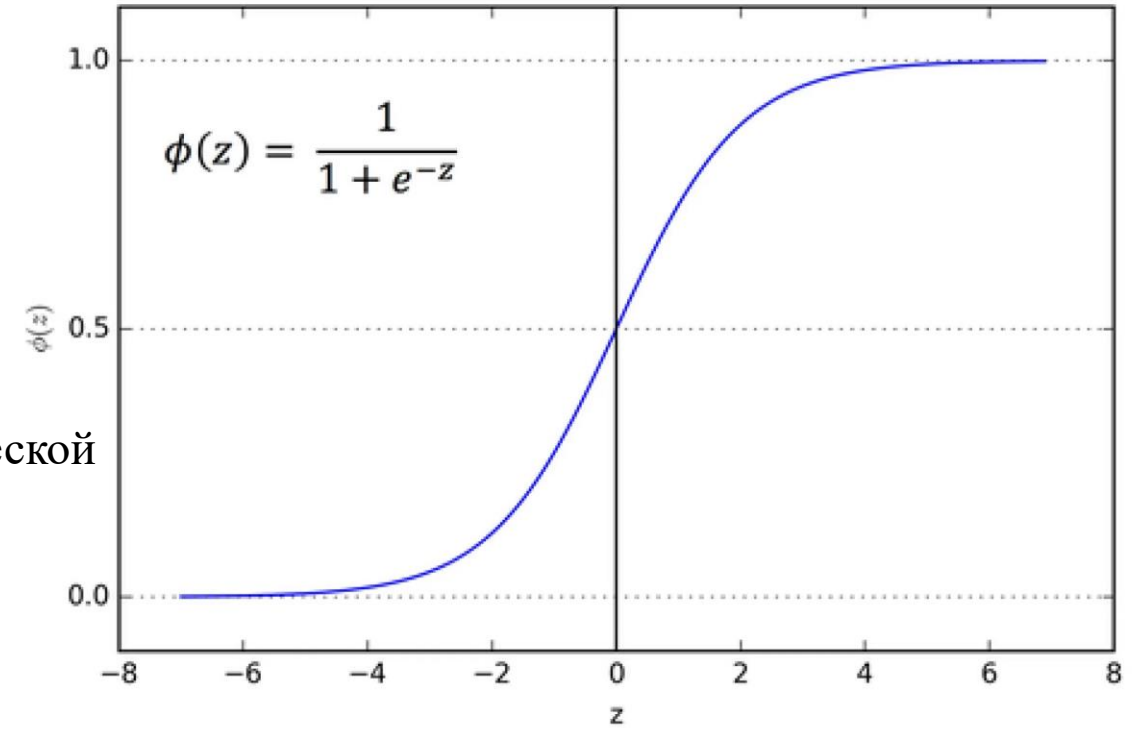
Логистическая регрессия

$$f(z) = \frac{1}{1 + e^{-z}},$$

где $z = \theta^T x$,

θ — вектор-столбец параметров (весов) логистической регрессии,

x — вектор-столбец независимых переменных.



Логистическая функция преобразует линейную комбинацию в вероятность принадлежности к целевому классу, принимающую значения в диапазоне от 0 до 1.

Функция потерь (Log loss)

Формула для правдоподобия:

$$\prod p_i^{y_i}(1-p_i)^{1-y_i} \to max$$

где p_i — вероятность принадлежности к целевому классу i-ого объекта,

y_i — целевое значение i-ого объекта.

Log loss:

$$-\frac{1}{n}\sum(y_i * \log(p_i) + (1-y_i) * \log(1-p_i)) \to min$$

где p_i — вероятность принадлежности к целевому классу i-ого объекта,

y_i — целевое значение i-ого объекта,

n — количество объектов.

Метрики качества

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

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

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

$$F1 = 2 * \frac{precision*recall}{precision+recall}.$$

Кривая ROC — это график, который иллюстрирует производительность классификационной модели при всех возможных порогах классификации. Ось X данного графика представляет собой FPR, т.е. ложноположительную частоту, а ось Y — TRP.

$$TPR = \frac{TP}{TP+FN},$$

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

Чтобы агрегировать эту кривую в скаляр вычисляется площадь под этой кривой.

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

TP — количество действительно положительных объектов;

TN — количество действительно отрицательных объектов;

FP — количество ложноположительных объектов;

FN — количество ложноотрицательных объектов.

Описание признаков

Тип данных	Поле	Описание поля
целочисленный	**RowNumber, CustomerId, Surname**	Идентификация
целочисленный	**CreditScore**	Кредитный рейтинг
строковый	**Geography**	Страна клиента
строковый	**Gender**	Пол
целочисленный	**Age**	Возраст
целочисленный	**Tenure**	Сколько лет клиент с банком сотрудничает
вещественный	**Balance**	Баланс
целочисленный	**IsActiveMember**	Активность (0/1)
целочисленный	**Exited**	Целевая переменная (0 — остался, 1 — ушёл).

Данные

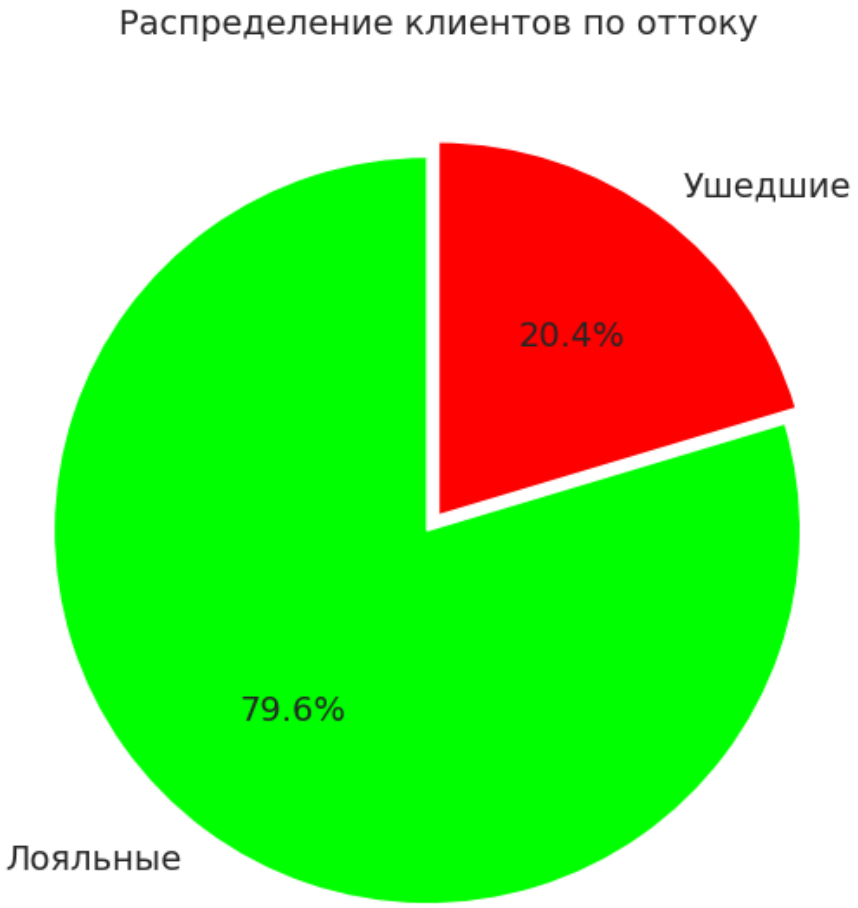
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

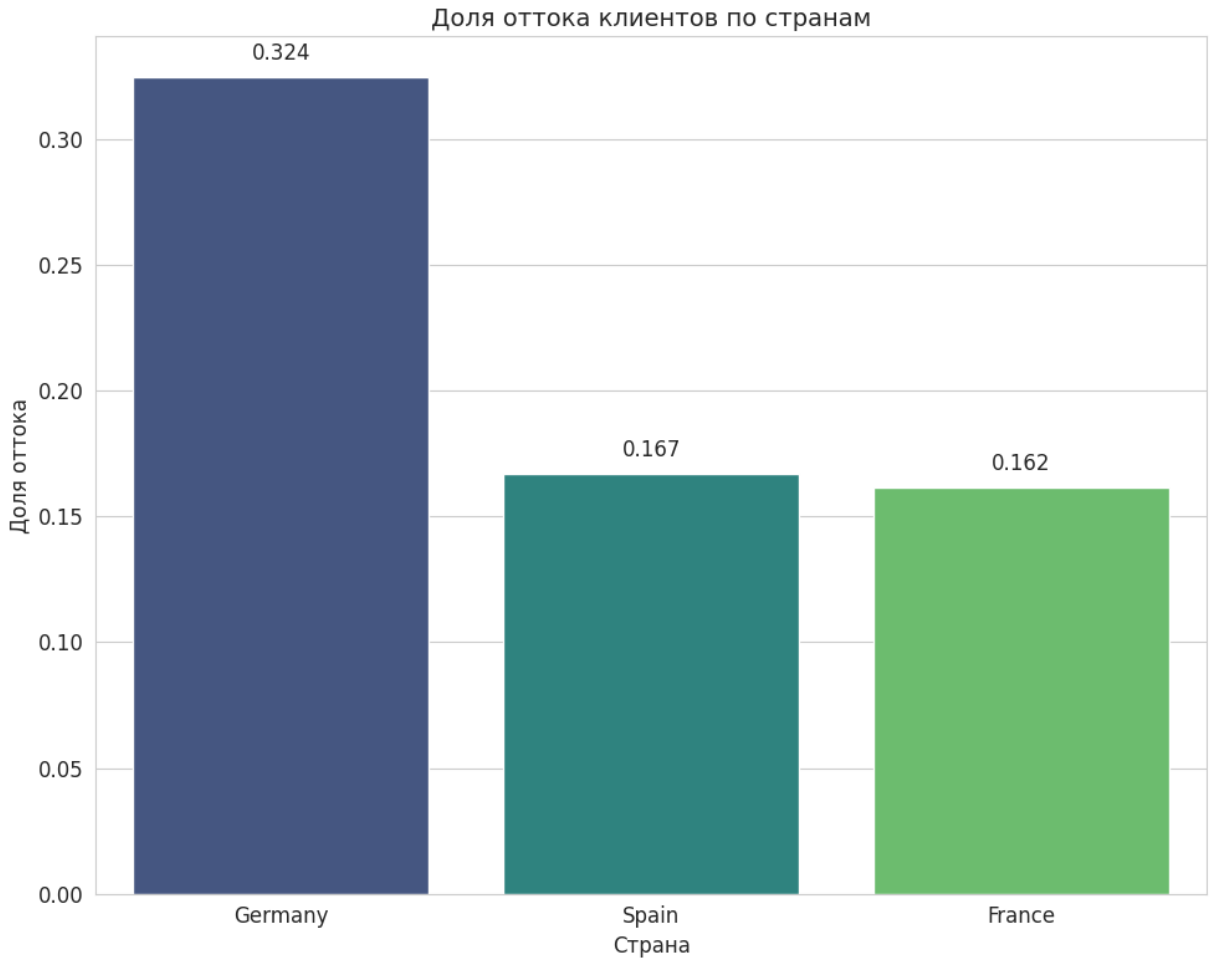
Данные

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
7	376	0	29	4	115046.74	4	1	0	119346.88	1
15	616	1	45	3	143129.41	2	0	1	64327.26	0
16	653	1	58	1	132602.88	1	1	0	5097.67	1
26	756	1	36	2	136815.64	1	1	1	170041.95	0
28	574	0	43	3	141349.43	1	1	1	100187.43	0

Предобработанные данные

Распределение признаков

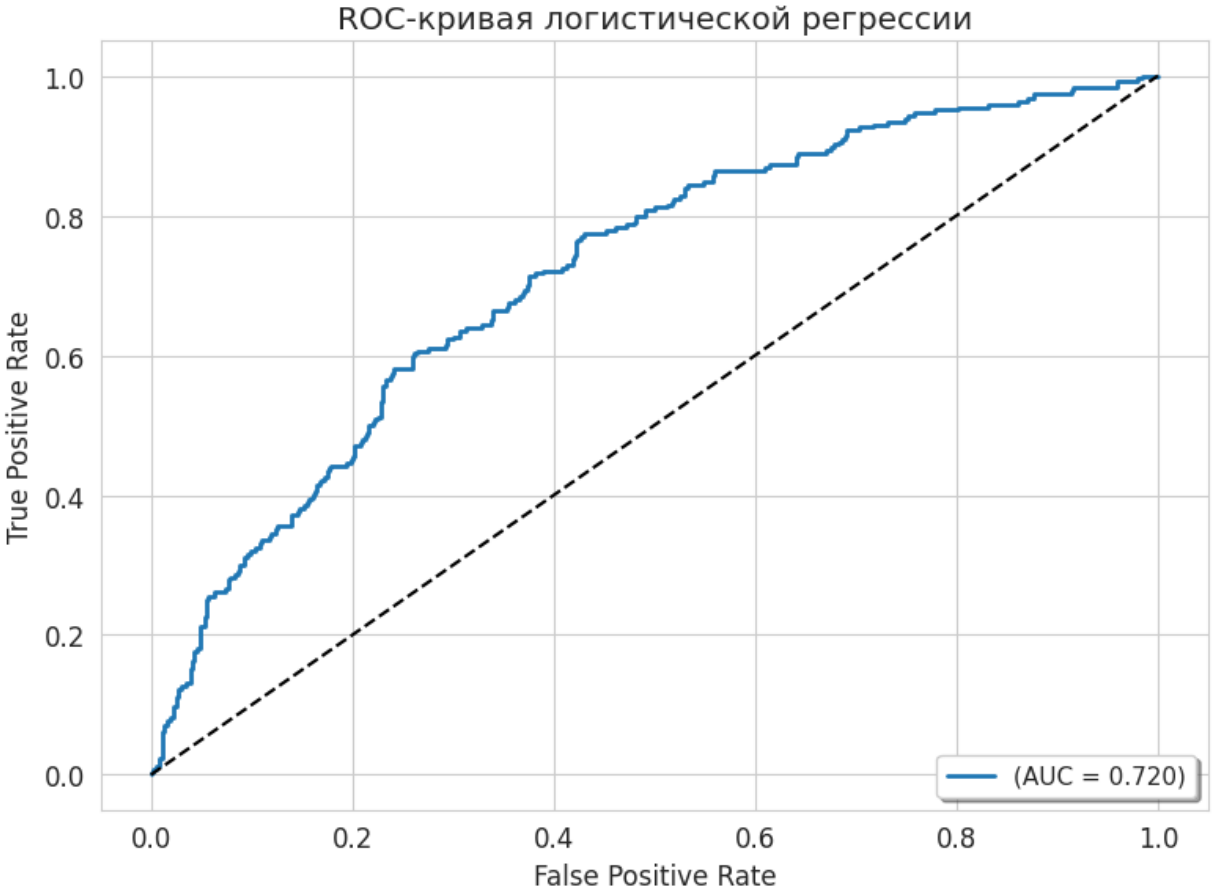


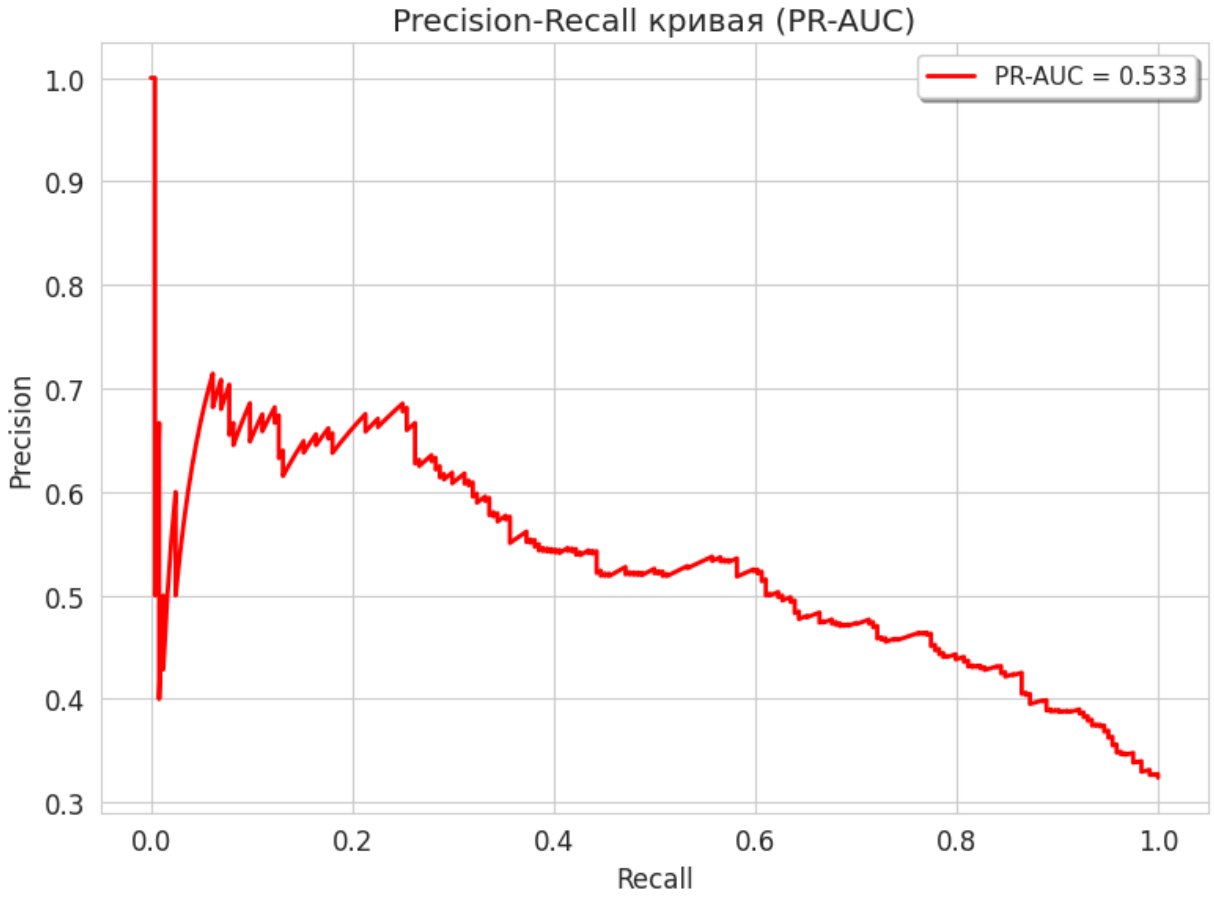


Модель классификации

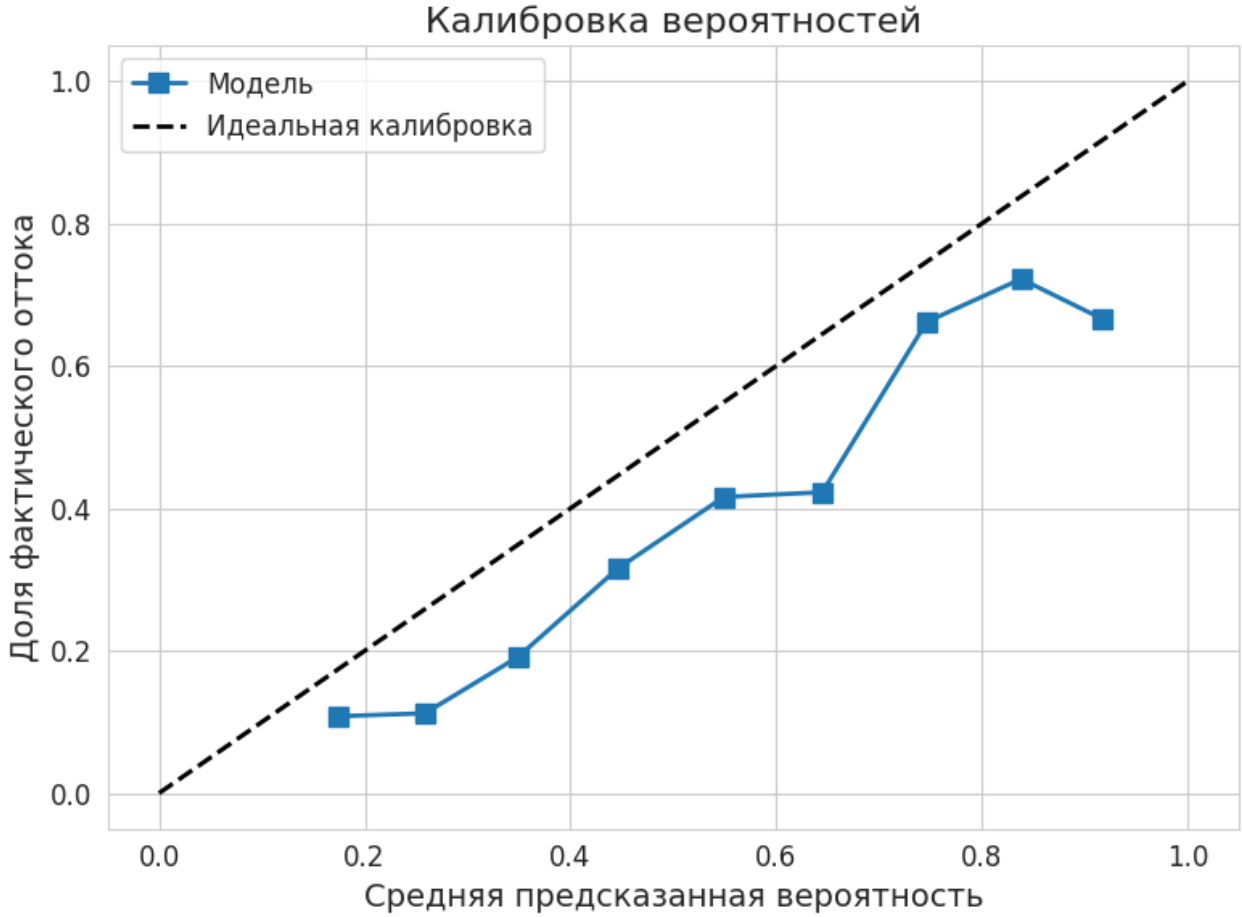
```
Accuracy: 0.6707
Precision: 0.4936
Recall: 0.6352
F1: 0.5556
ROC-AUC: 0.7198
PR-AUC: 0.5330
Логарифмическая функция потерь: 0.6139
```

Визуализация





Визуализация

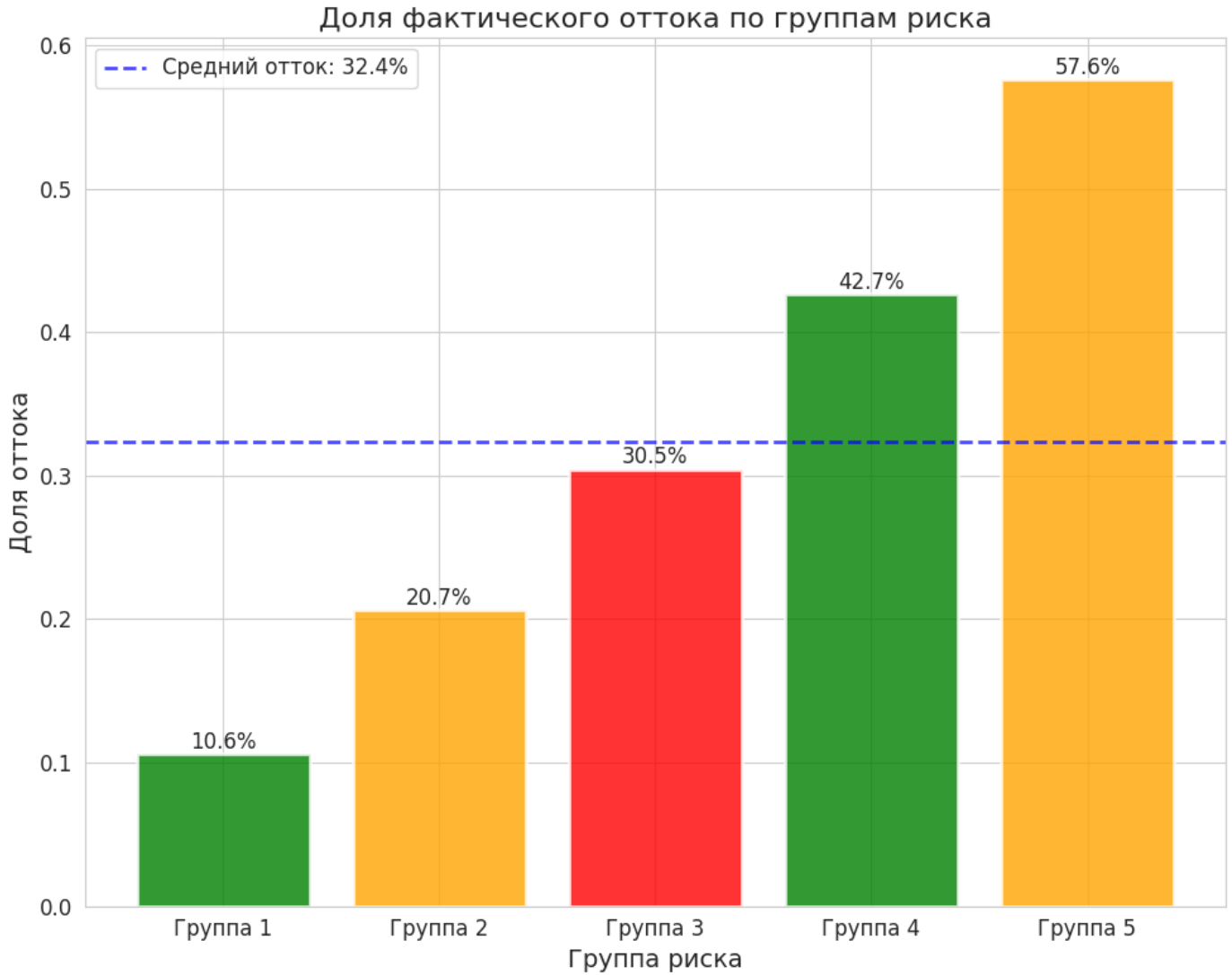


Система скоринга

```
1. Распределение клиентов по группам риска:
Группа 1: 151 клиентов  20.1%
Группа 2: 150 клиентов  19.9%
Группа 3: 151 клиентов  20.1%
Группа 4: 150 клиентов  19.9%
Группа 5: 151 клиентов  20.1%
2. Доля оттока по группам риска:
Группа 1: 10.6% оттока (16 из 151 клиентов)
Группа 2: 20.7% оттока (31 из 150 клиентов)
Группа 3: 30.5% оттока (46 из 151 клиентов)
Группа 4: 42.7% оттока (64 из 150 клиентов)
Группа 5: 57.6% оттока (86 из 151 клиентов)
3. Пороги вероятностей для групп риска:
Группа 1: 0.000 - 0.303
Группа 2: 0.303 - 0.396
Группа 3: 0.396 - 0.512
Группа 4: 0.512 - 0.631
Группа 5: 0.631 - 1.000
```

```
ГРУППА 2 (риск низкий)
ХАРАКТЕРИСТИКИ КЛИЕНТА:
  Пол: Женщина
  Возраст: 36 лет
  Кредитный рейтинг: 702
  Баланс: 105,264.88
  Страна: Франция
  Количество продуктов: 2
  Стаж в банке: 2 лет
  Активный клиент: Да
  Есть кредитная карта: Да
  Оценочная зарплата: 52,909.87

РЕЗУЛЬТАТЫ ПРОГНОЗИРОВАНИЯ:
  Вероятность оттока: 0.359 (35.9%)
  Прогноз модели: Останется
  Фактический результат: Остался
  Результат: Модель правильно спрогнозировала
```



Заключение

Цель данной курсовой работы — разработать и протестировать сценарий анализа и обработки данных для прогнозирования оттока клиентов банка с использованием логистической регрессии — достигнута.

Основные результаты показывают:

- Анализ данных показал, что наибольший отток наблюдается в Германии, а также среди клиентов старшего возраста с высоким балансом.

- Построена и оценена модель логистической регрессии с L2-регуляризацией, показала хорошие результаты (F1 = 0.56, ROC-AUC = 0.72, PR-AUC: 0.53).

- Разработана система скоринга, позволяющая сегментировать клиентов по уровню риска.

- Важные признаки: возраст, баланс, активность клиента, число продуктов.

- Практическая значимость: модель позволяет идентифицировать клиентов с высоким риском оттока и принимать превентивные меры.