Применение методов машинного обучения для предсказания поведения абонентов оператора сотовой связи

Выполнил: Корыстов М.А. 444 группа
Научный руководитель: Терехов А.Н. д.ф.-м.н., проф.
Рецензент: Невоструев К.Н.

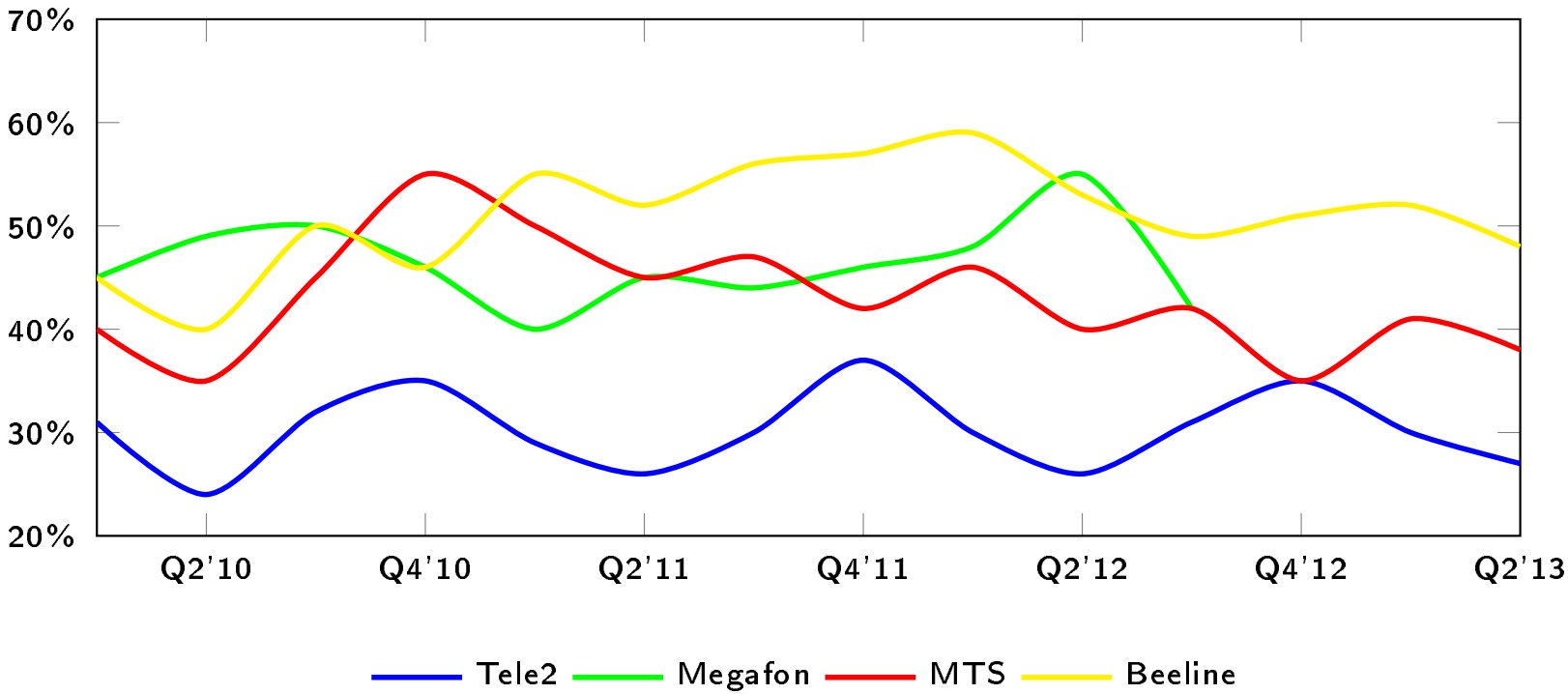
СПбГУ

июнь 2015 г.

Введение

Отток абонентов в российской телекоммуникационной отрасли

- Отток 25-50% в год
- Привлечение нового абонента в несколько раз дороже удержания старого



Постановка задачи

Исходные данные

Выборка активности по 10 показателям 150 000 абонентов за 15 последовательных месяцев.

Цели

Построить классификатор, предсказывающий уход абонентов на исходных данных.

Задачи

1. Обработать исходные данные
2. Проанализировать характеристики представленной выборки
3. Провести эксперименты сравнения моделей
4. Выбрать модель машинного обучения с лучшим результатом

Обработка данных

Получение образцов из временных рядов с помощью скользящего окна.

Дополнительные характеристики

- Арифметические отношения
- Геометрические отношения
- Центральные моменты
- Данные об устройстве связи

Итоговая выборка состоит из 96460 образцов с 104 характеристиками.

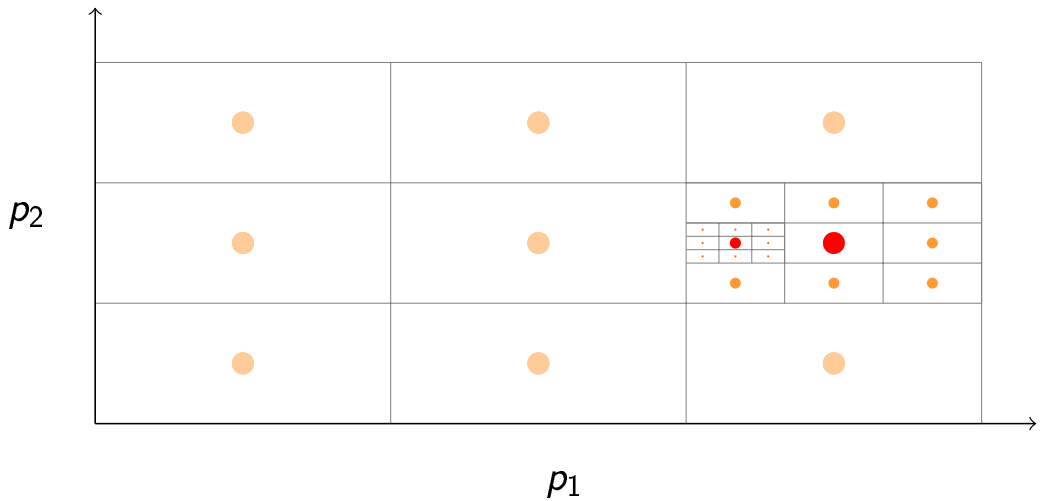
Обучение

Перспективные модели

- Логистическая регрессия
- Нейронные сети
- Random forest
- Градиентный бустинг над решающими деревьями

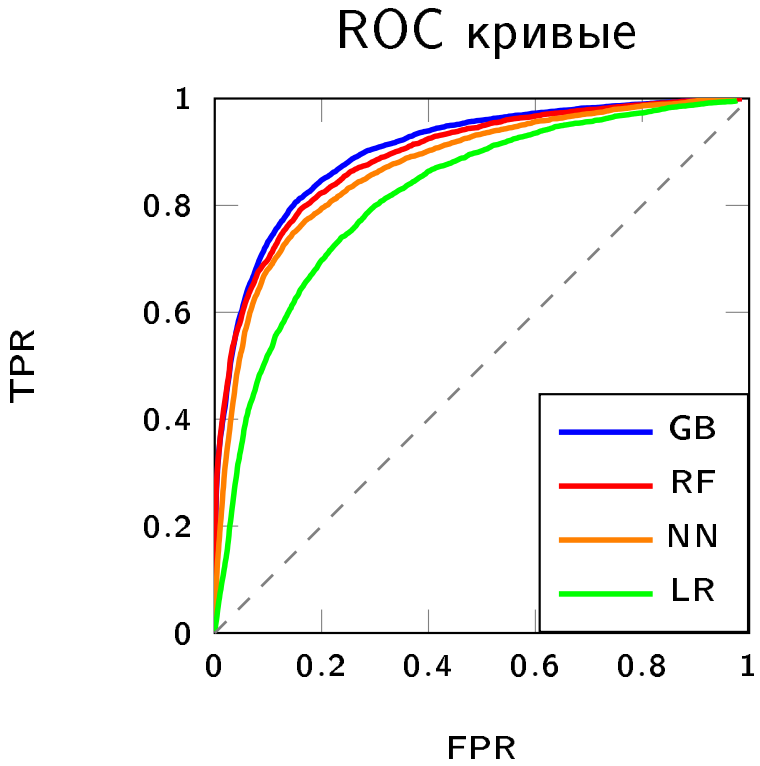
Подбор параметров

Перебор параметров модели с помощью алгоритма поиска в сетке.



Сравнение моделей

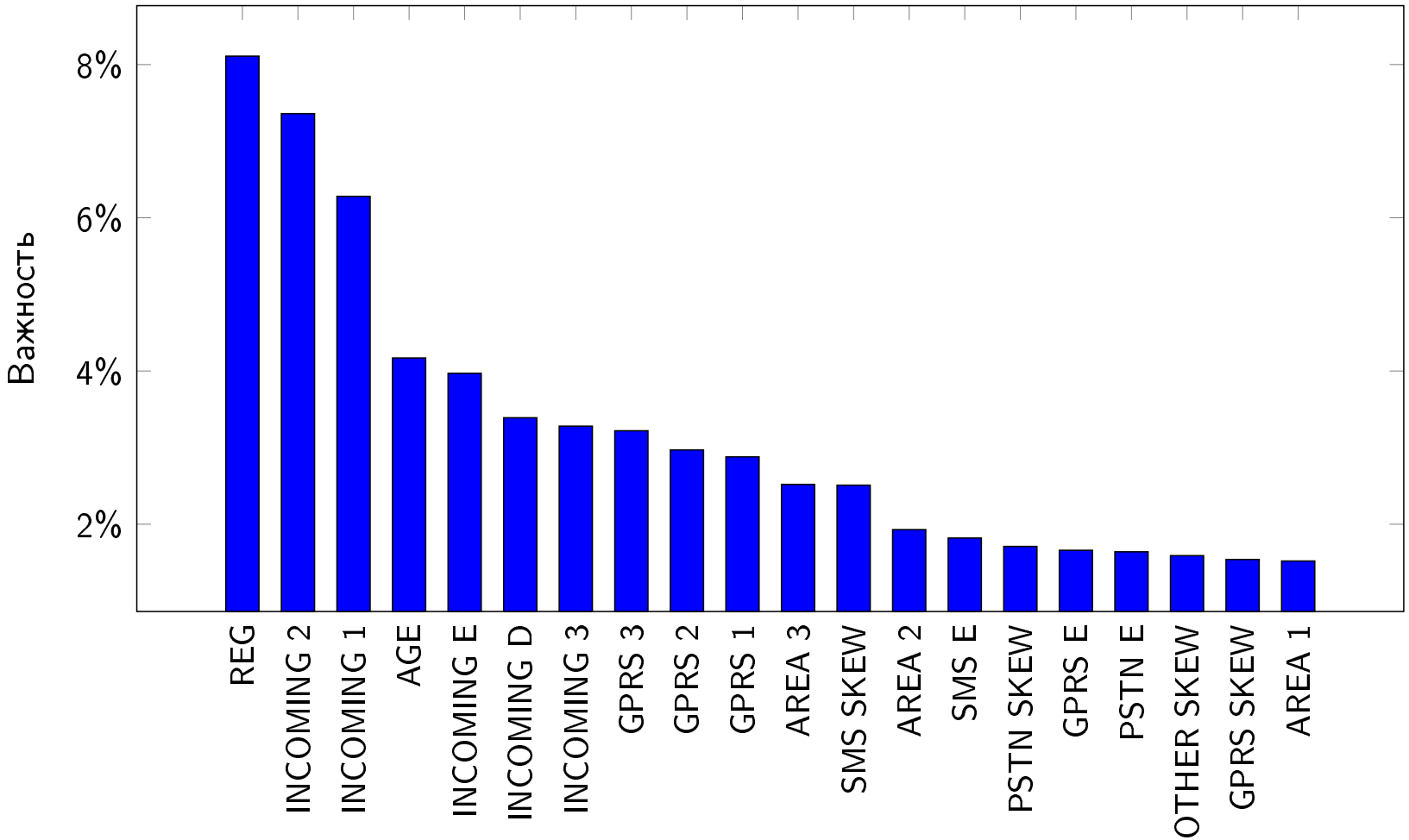
Модель	precision	recall	F1	F0.5	**AUC**
GB	0.70	0.64	0.66	0.68	**0.842**
Random forest	0.72	0.60	0.65	0.69	**0.832**
Нейронные сети	0.69	0.59	0.63	0.66	**0.825**
Логическая регрессия	0.63	0.37	0.46	0.55	**0.842**



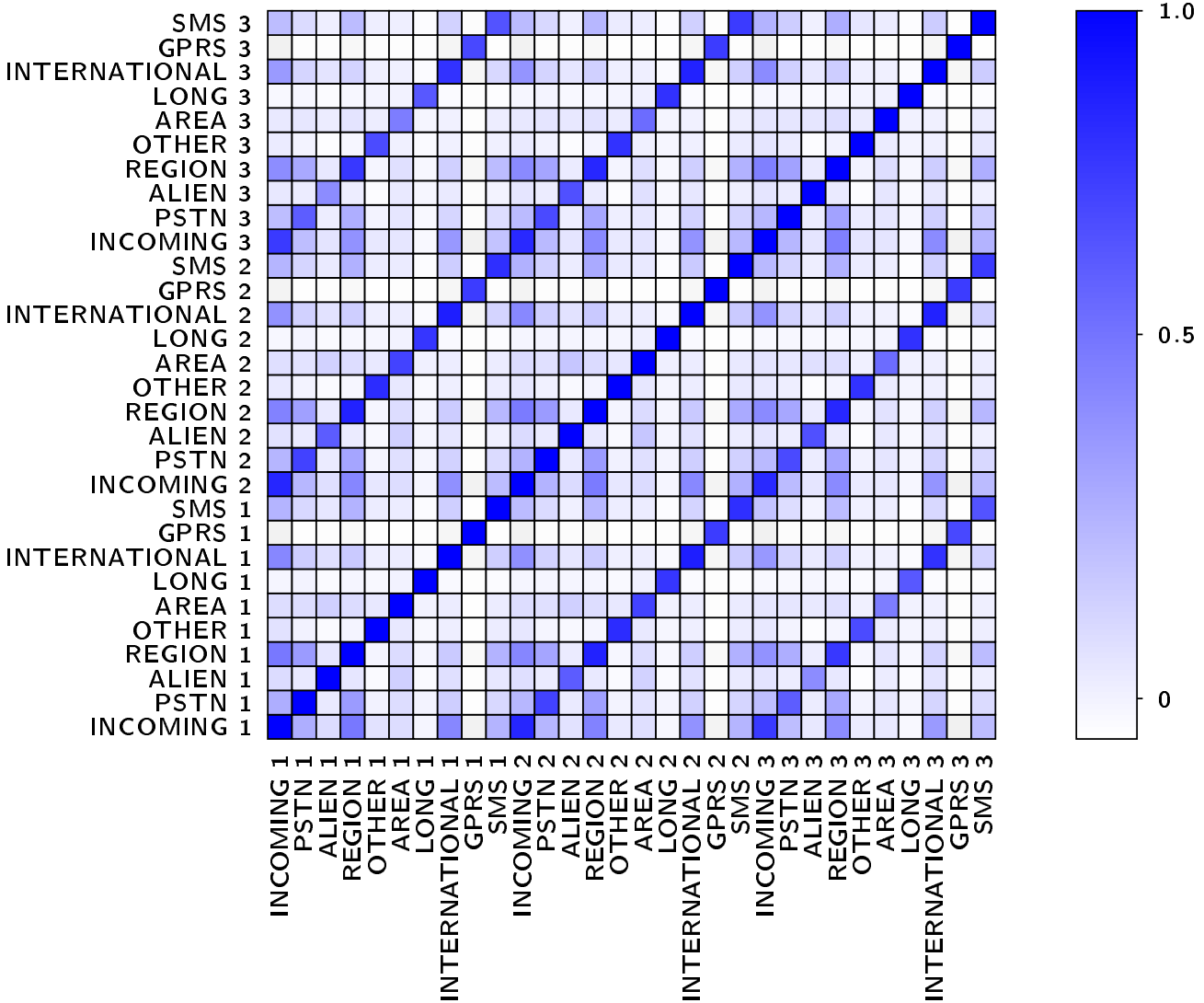
$$TPR = TP/P$$

$$FPR = FP/N$$

Важность характеристик



Важность характеристик

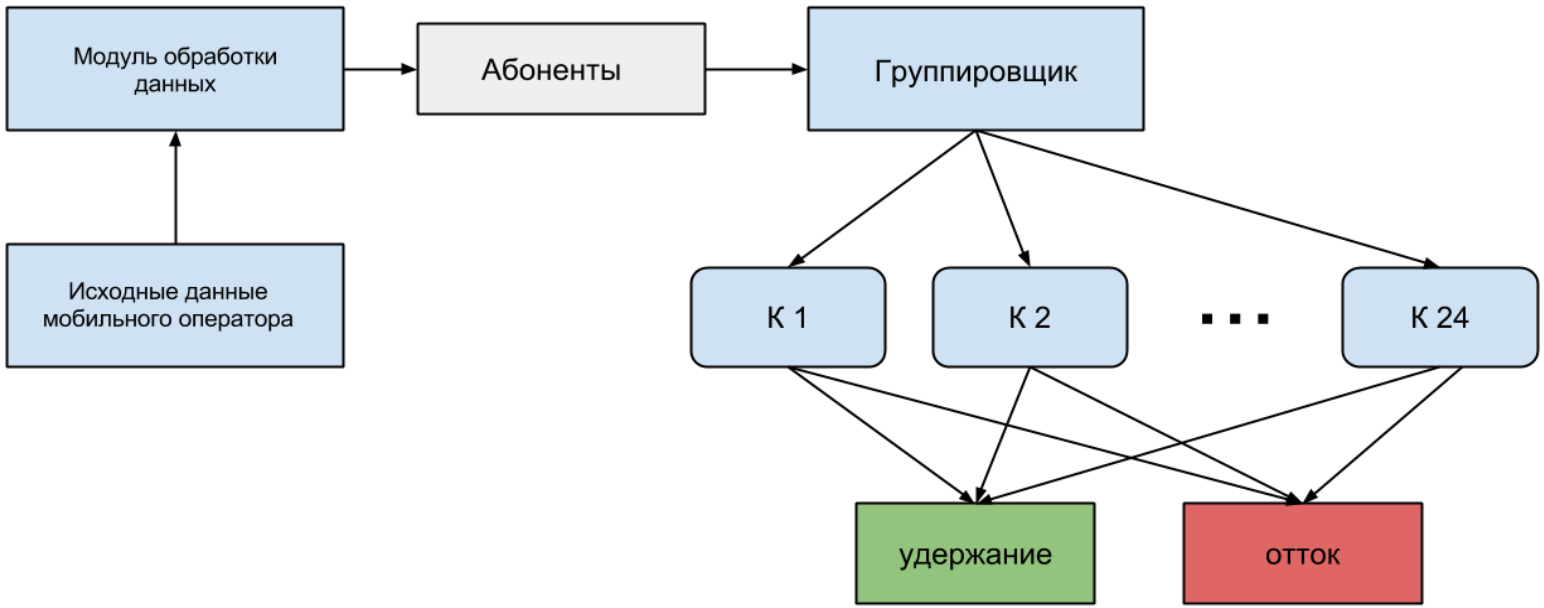


Группирование

Группирование по параметрам:

- Пол (М, Ж, Юр.лицо)
- Дата регистрации ($>$ 2 лет, $>$ 1 года, $>$ 3 месяцев, $<$ 3 месяцев)
- Использует только интернет (телефон, модем)

Обучение 24 независимых классификаторов в каждой группе.



Результаты

Модель	precision	recall	F1	F0.5	**AUC**
Набор классификаторов	0.75	0.66	0.70	0.73	**0.90**
GB	0.72	0.65	0.68	0.70	**0.88**
ADA boost	0.68	0.60	0.63	0.66	**0.86**
Random forest	0.62	0.82	0.70	0.65	**0.85**
SVM	0.64	0.57	0.60	0.62	**0.80**

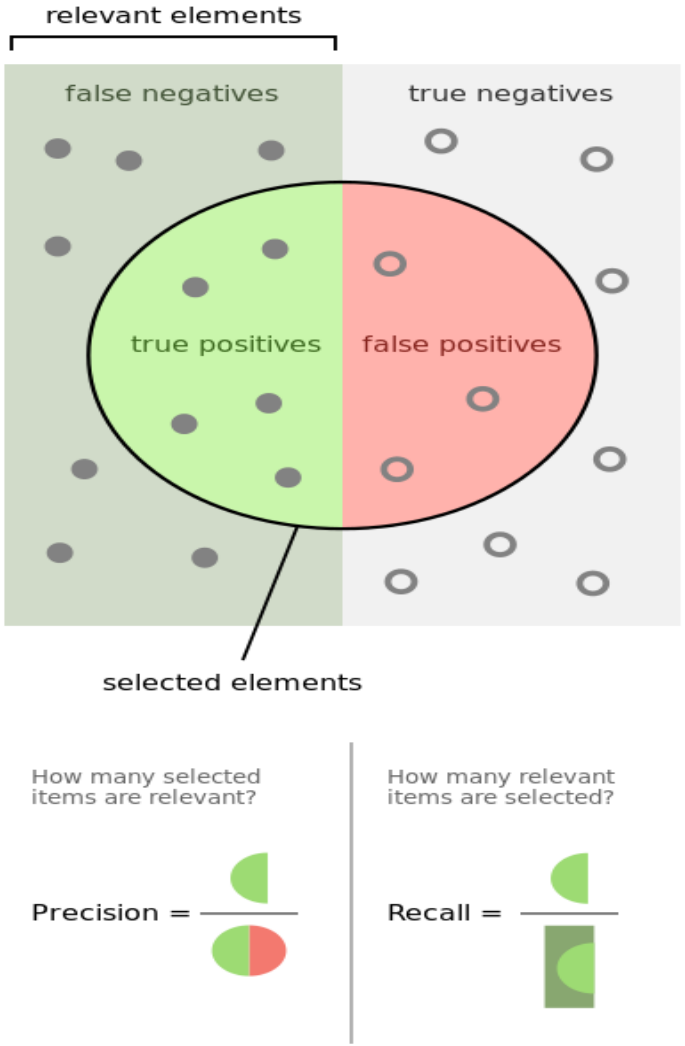
- Проведена обработка исходных данных
- Выполнено сравнение результатов обучения нескольких моделей
- Построен классификатор на основе нескольких моделей с применением группирования достигающий 0.905 ± 0.004 AUC на тестовой выборке

Оценка результатов обучения

Метрики

- precision, recall
- F1, F0.5
- AUC

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$



Результаты по группам

		Мужской			Женский			Юр. лицо		
		prec	recall	AUC	prec	recall	AUC	prec	recall	AUC
$<$ 3 месяцев	телефон	0.76	0.79	0.82	0.75	0.85	0.80	0.89	0.88	0.95
	модем	0.68	0.76	0.64	0.72	0.90	0.70	0.86	0.41	0.80
$>$ 3 месяцев	телефон	0.73	0.52	0.84	0.76	0.58	0.87	0.59	0.61	0.86
	модем	0.66	0.52	0.75	0.66	0.66	0.83	0.70	0.63	0.80
$>$ 1 года	телефон	0.63	0.60	0.88	0.78	0.65	0.86	0.75	0.69	0.85
	модем	0.80	0.53	0.77	0.74	0.74	0.84	0.63	0.70	0.79
$>$ 2 лет	телефон	0.83	0.52	0.92	0.86	0.57	0.94	0.86	0.60	0.92
	модем	0.83	0.53	0.85	0.72	0.71	0.85	0.84	0.64	0.92