Применение наивного байесовского классификатора в задаче предсказания оттока абонентов оператора сотовой связи

Деркунский Виктор, 344 группа

Научный руководитель:
д.ф.- м.н., профессор Терехов Андрей Николаевич

Введение

- Удержание абонента дешевле привлечение нового
- Отток может достигать 50 % в год
- Машинное обучение позволяет достаточно точно предсказать поведение абонента

Постановка задачи

- Провести обзор существующих решений данной задачи
- Обработать предоставленные данные
- Рассмотреть эффективность наивного байесовского классификатора в задача предсказания оттока абонентов операторов сотовой связи
- Рассмотреть методы повышения эффективности классификатора
- Рассмотреть ансамбли, основанные на наивном байесовском классификаторе
- Рассмотреть ансамбли с другими моделями машинного обучения

Инструменты

- Pandas - для работы с данными
- Scikit-Learn - для построения алгоритмов и оценки их эффективности
- Numpy - для проведения трудоемких вычислений

Оценка точности классификации

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

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

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

$$F = (b^2 + 1)\frac{precision * recall}{b^2\, presion + recall}$$

$$AUC = \int_0^1 \frac{TP}{P}\, d\frac{FP}{N}$$

Обзор

- Лучший результат прошлого года:

  (precision = 0.75, recall = 0.72, AUC = 0.92, F1 = 0.73)

  Наиболее перспективные модели для ансамблирования:
- решающее дерево
- лес случайных деревьев
- lda
- логистическая регрессия
- k-NN

Наивный байесовский классификатор

$$c = argmax_{c \in C} P\left(C \middle| O_1 \ldots O_n\right) = argmax_{c \in C} P\left(c\right) \prod_{i \in 1..n} P\left(O_i \middle| C\right)$$

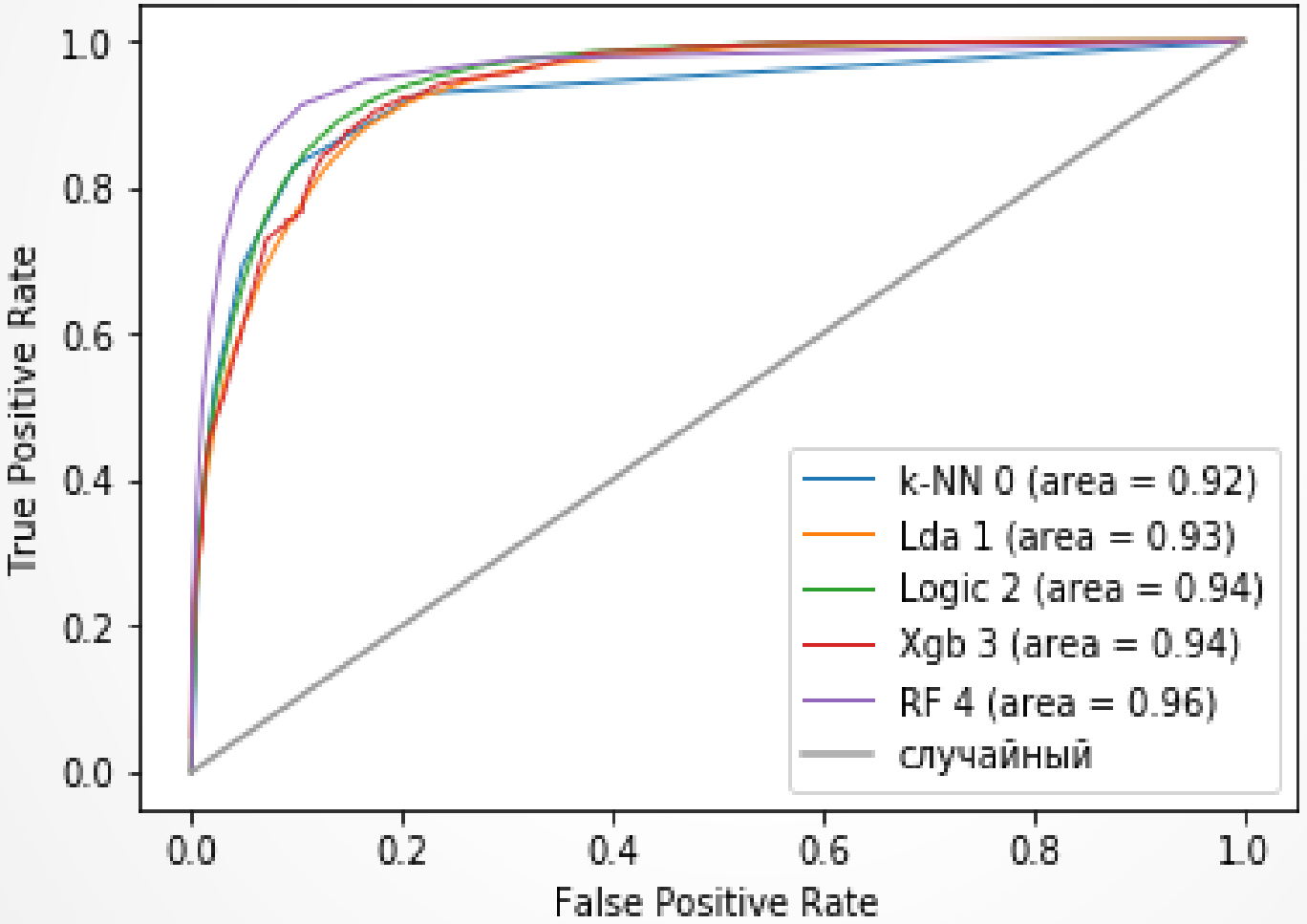
Данные

Были предоставлены данные 300000 абонентов и их активность за последние 3 месяца. Каждый абонент описывался через следующие характеристики:

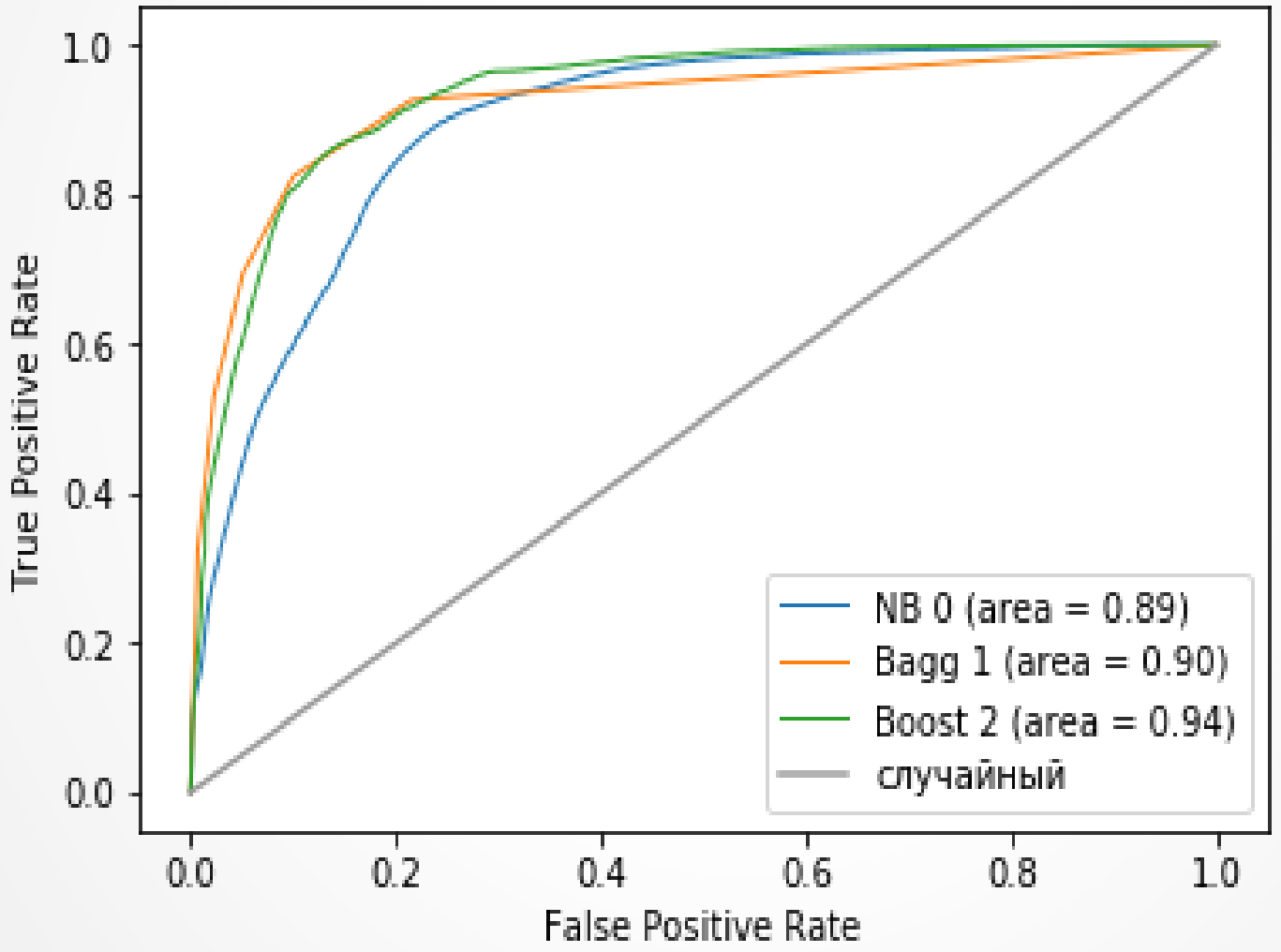
- Количество минут и стоимость исходящих
- Количество минут и стоимость входящих
- Количество и стоимость исходящих СМС
- Количество и стоимость входящих СМС
- Количество трафика мобильного интернета и его стоимость • Информация о количестве обращений клиента в справочные службы
- Личные данные

Далее данные были сжаты методом главных компонент до 58 параметров

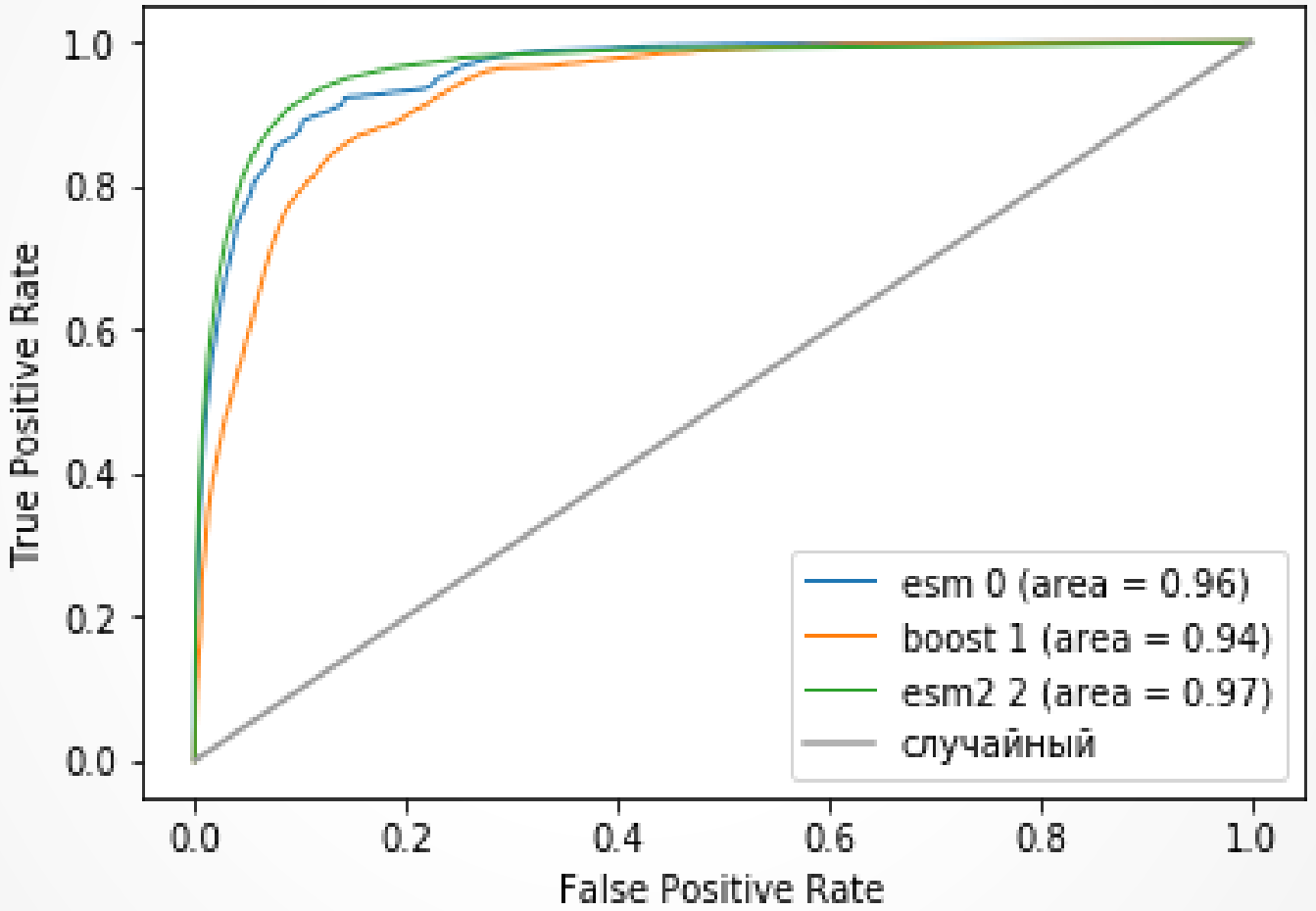
Оценка моделей из обзора



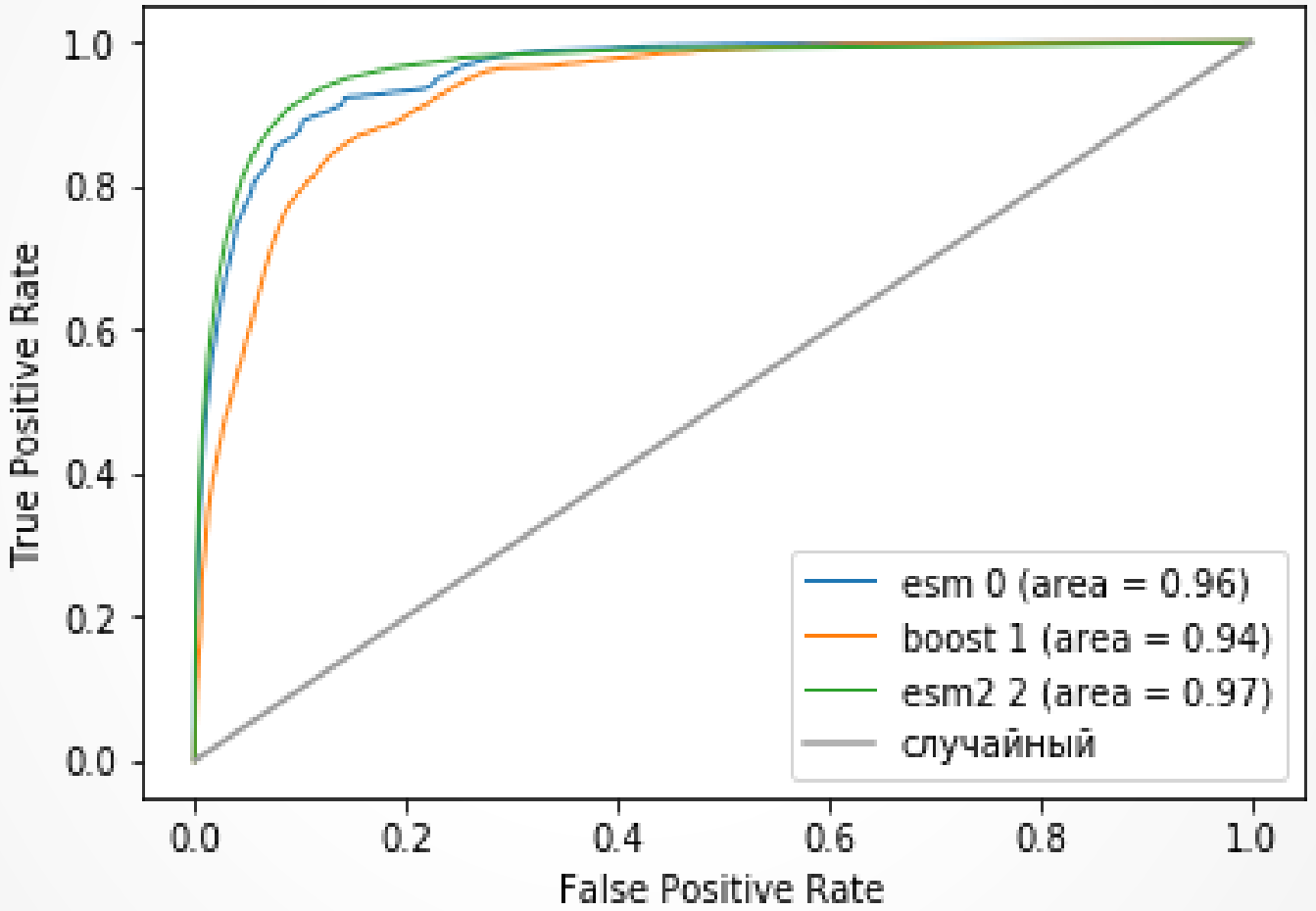
Оценка Naive Bayes



Результаты ансамблирования



Результаты ансамблирования



Общий результат :

	Accuracy	Precision	Recall	AUC	F1
Лучший результат прошлых лет[6]	0.88	0.72	0.75	0.92	0.73
NB	0.89	0.54	0.66	0.89	0.6
Бэггинг NB	0.9	0.56	0.67	0.9	0.61
Бустинг NB	0.91	0.72	0.76	0.94	0.74
Ансамбль с NB(*)	0.92	0.75	0.83	0.96	0.79
Ансамбль без NB(**)	0.93	0.76	0.84	0.97	0.8