Smote

Synthetic Minority Over-sampling Technique

Техники работы с несбалансированными данными

I. ***Препроцессинг данных***:

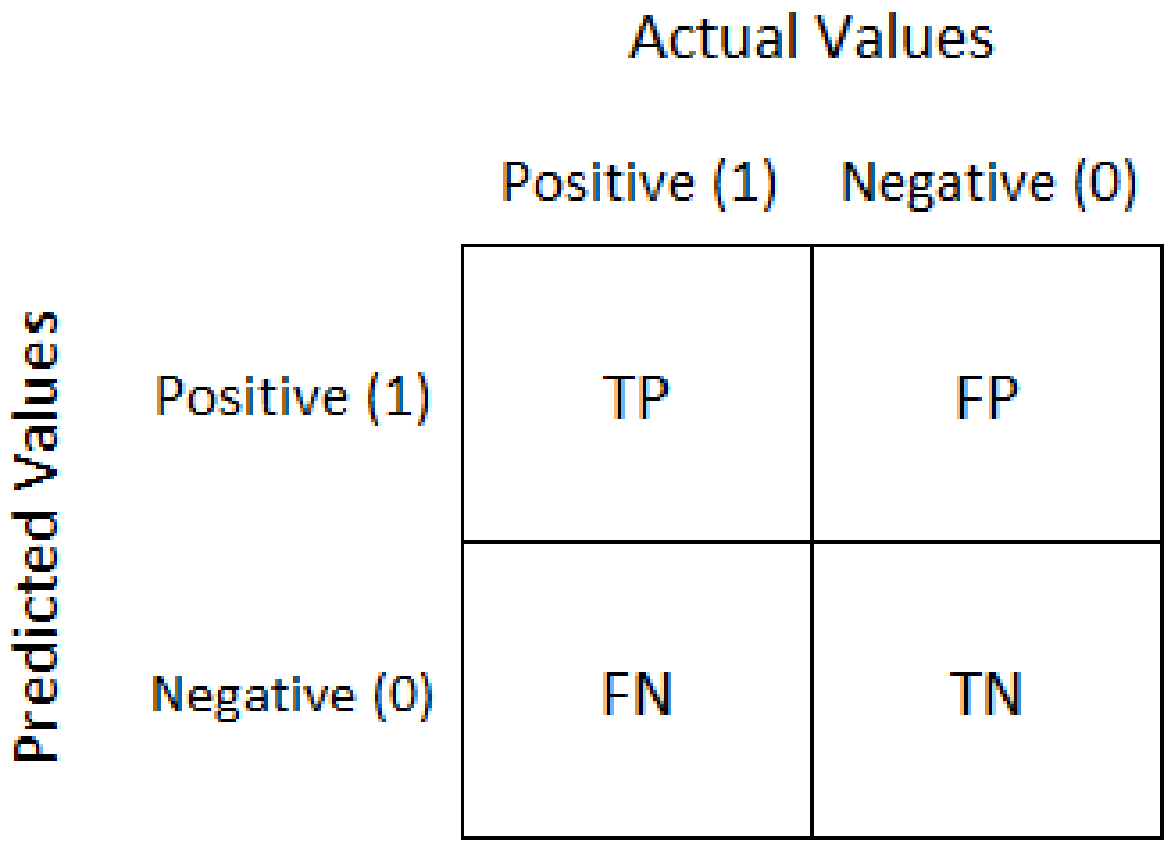
1) Smote
2) Random Under-sampling
3) Random Over-sampling
4) (на примере классификаторов Random Forest и Logistic Regression)

II. ***Модификация работы классификаторов:***
Задание априорного распределения классов (на примере классификатора Naive Bayes)

Алгоритм Smote(T,N,k)

1) ***Параметры алгоритма:***
2) а) T — число образцов меньшего класса
3) б) N% от T — число синтетических образцов, которые хотим получить
4) в) k — число ближайших соседей
5)
6) ***Суть алгоритма:***
7) Для каждого образца i меньшего класса находим k ближайших соседей из обоих классов и генерируем N/100 искусственных образцов, повторяя N/100 раз для образца i:
8)
9) Среди k соседей образца i произвольно выбираем одного — nn. Прибавляем к каждому из атрибутов i разницу между соответствующими атрибутами nn и i, умноженную на произвольное число из отрезка [0,1].
10) Получен новый искусственный меньшего класса.

Метрики качества



В матрице неточностей меньший (более важный) класс назовем положительным, больший — отрицательным.

1) Accuracy — для сбалансированных данных

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

---

2) Для несбалансированных данных:

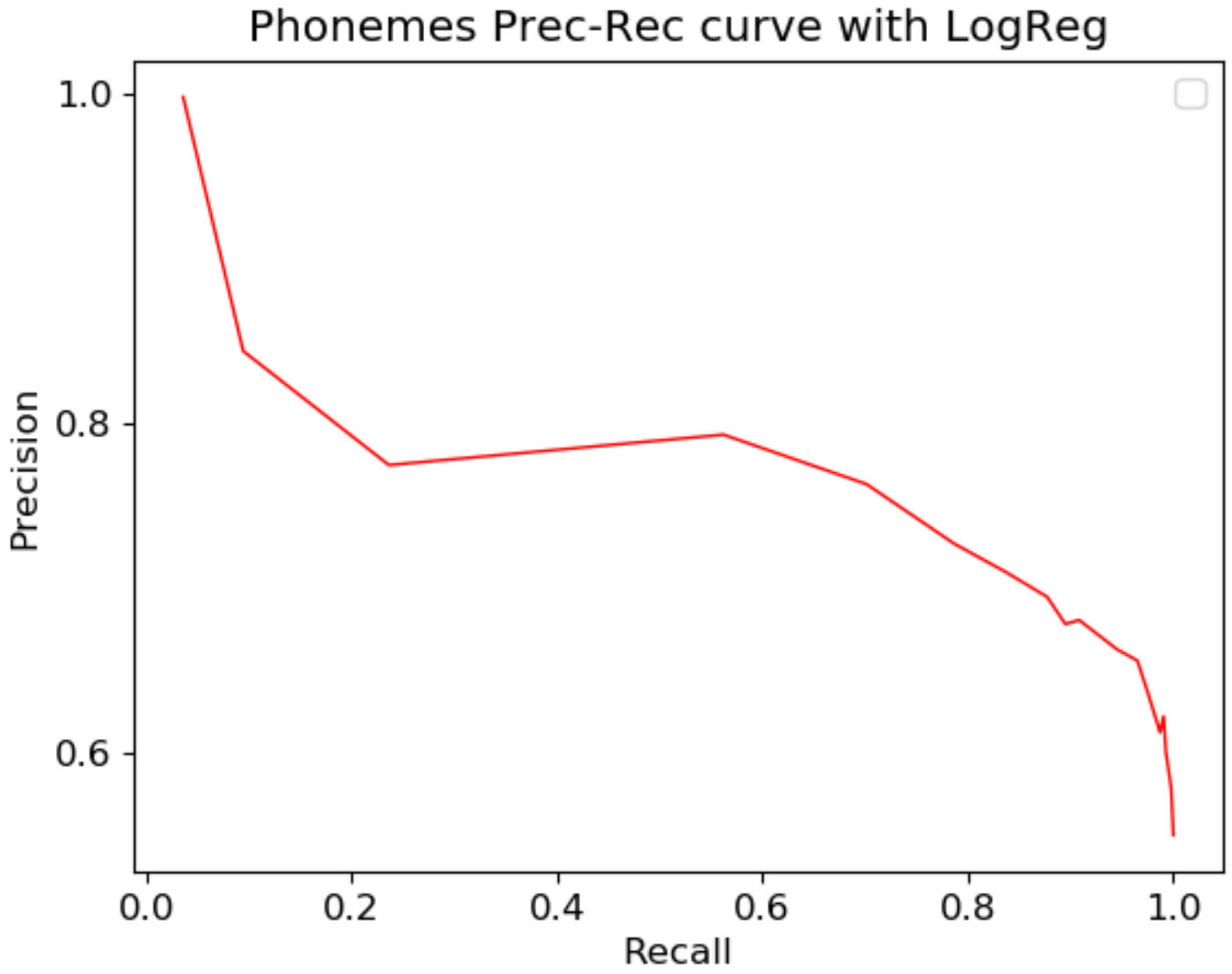
$$recall = \frac{TP}{TP + FN}$$

Полнота

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

Точность

Recall — Precision кривая



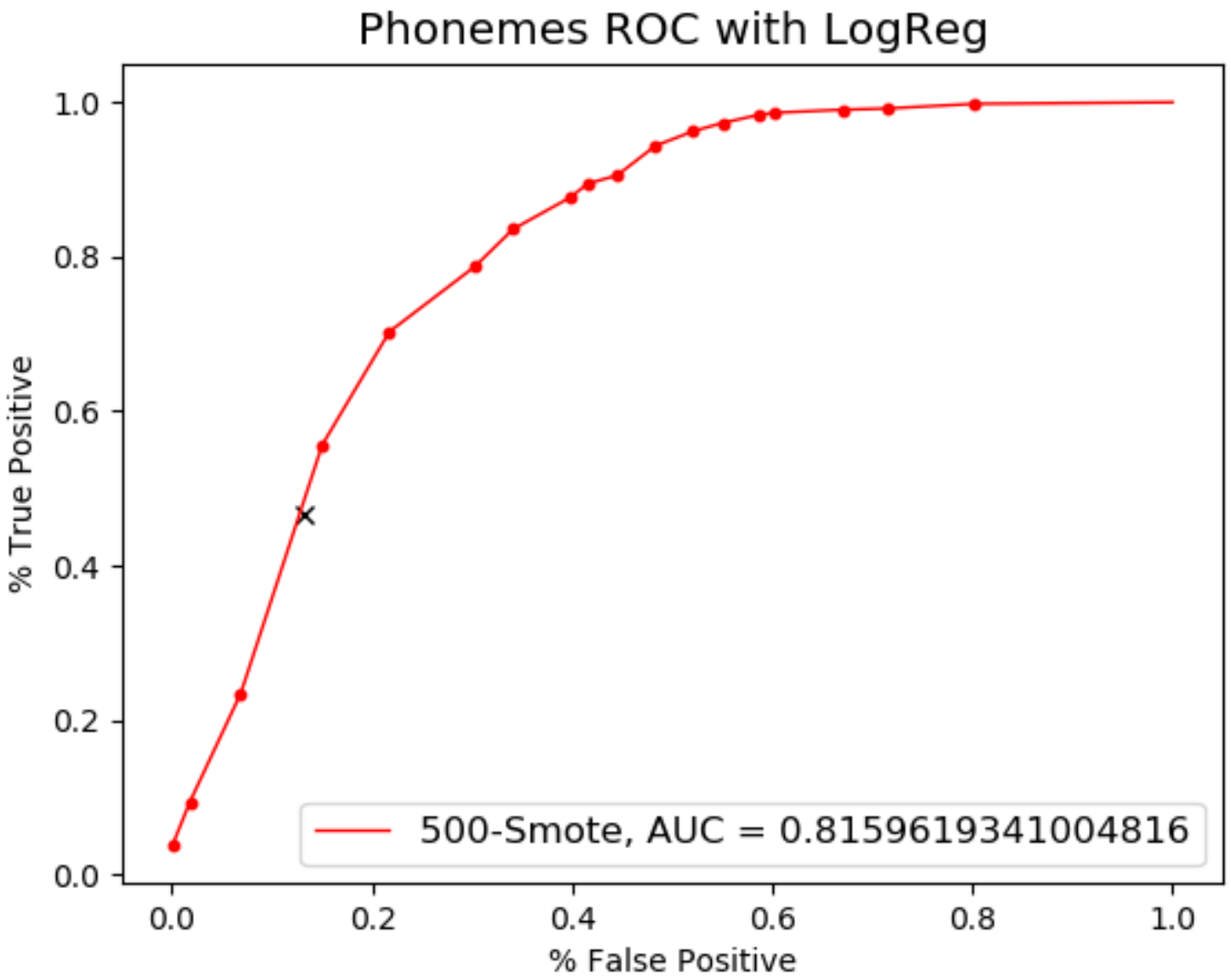
3) ROC кривая – для несбалансированных данных

- Кривая отображает соотношение правильно классифицированных объектов положительного класса и неверно классифицированных объектов отрицательного класса.
-
- Каждая точка с координатами FPR по оси X и TPR по оси Y — результат классификации.
-

False Positive Rate $$\mathrm{FPR\ (1\text{-}specificity)} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$

True Positive Rate $$\mathrm{TPR\ (sensitivity)} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Метрика AUC - площадь под ROC кривой



Smote для дисретных атрибутов обьектов

1) SMOTE-NC

Пример двух объектов: $\mathrm{F1} = 1\ 2\ 3\ \mathrm{A}\ \mathrm{B}\ \mathrm{C} \quad \mathrm{F2} = 4\ 6\ 5\ \mathrm{A}\ \mathrm{D}\ \mathrm{E}$

Расстояние между ними:

$$\mathrm{Eucl} = \mathrm{sqrt}[(4\text{-}1)^2 + (6\text{-}2)^2 + (5\text{-}3)^2 + \mathrm{Med}^2 + \mathrm{Med}^2]$$

---

2) SMOTE-N

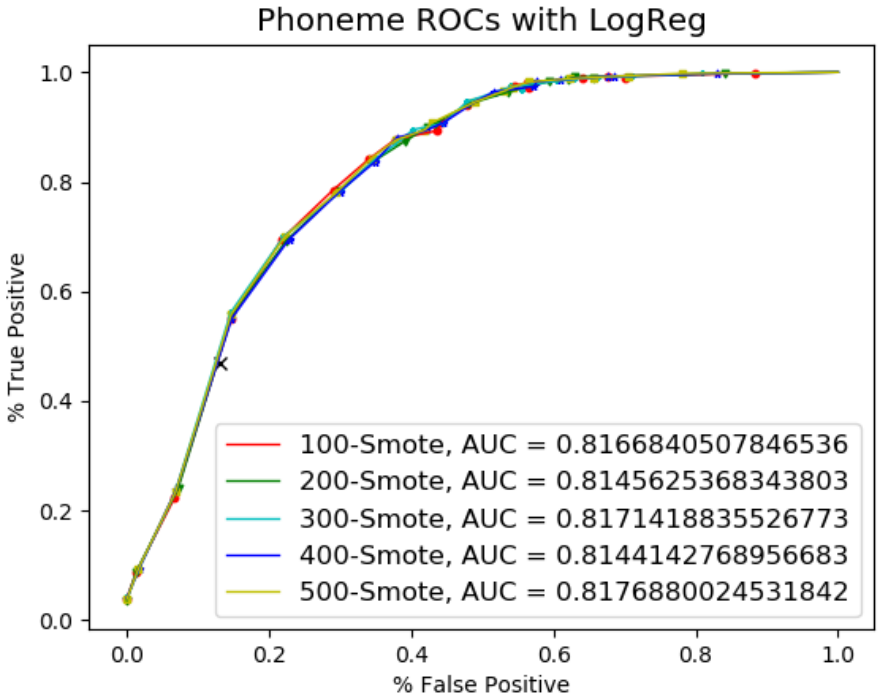
Расстояние между двумя значениями атрибутов V(1) и V(2)

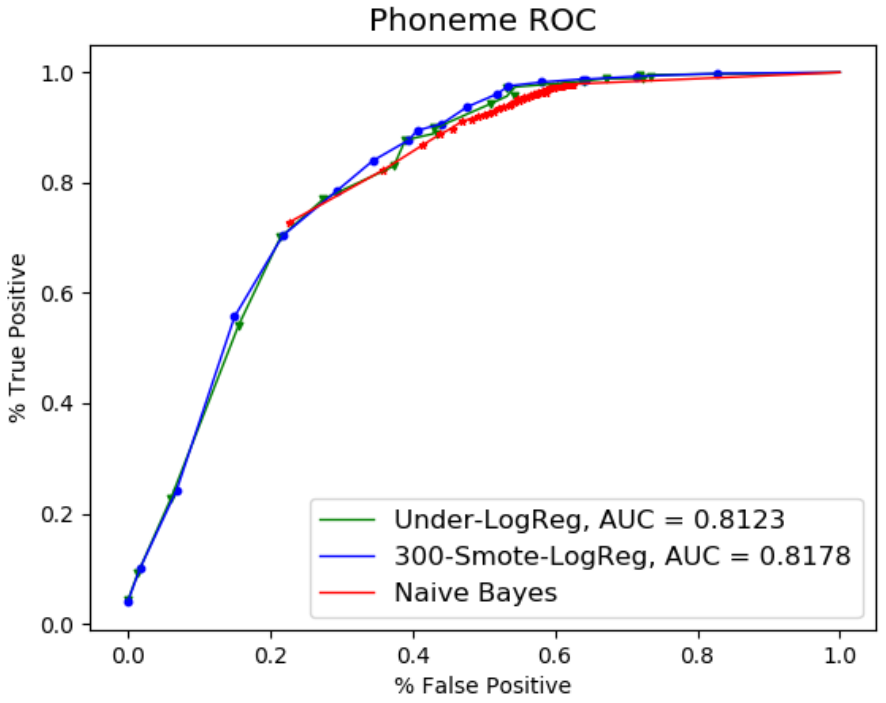
$$\delta(V_1, V_2) = \sum_{i=1}^{n} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right|^k$$

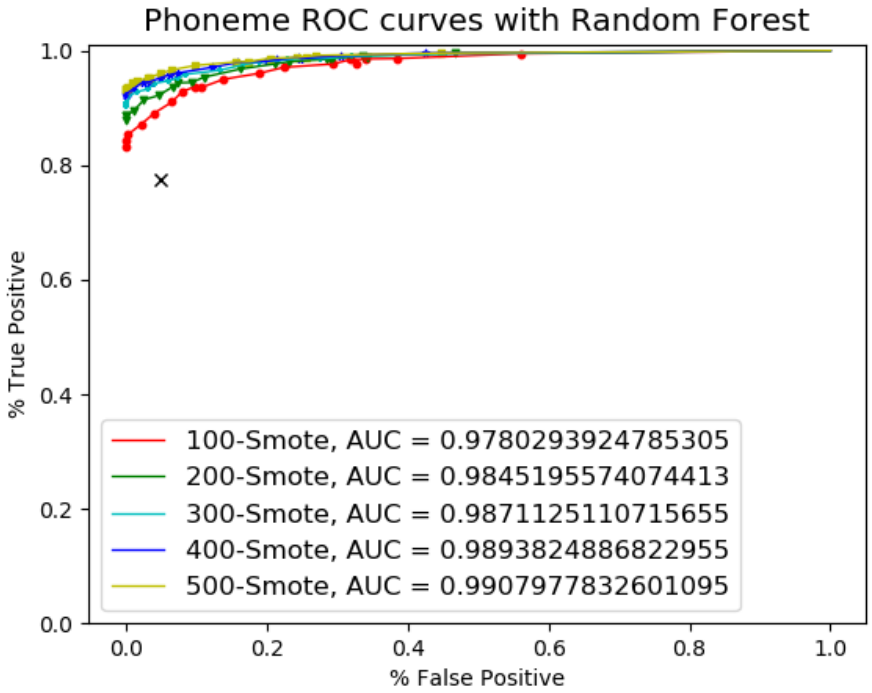
Расстояние между двумя объектами X и У

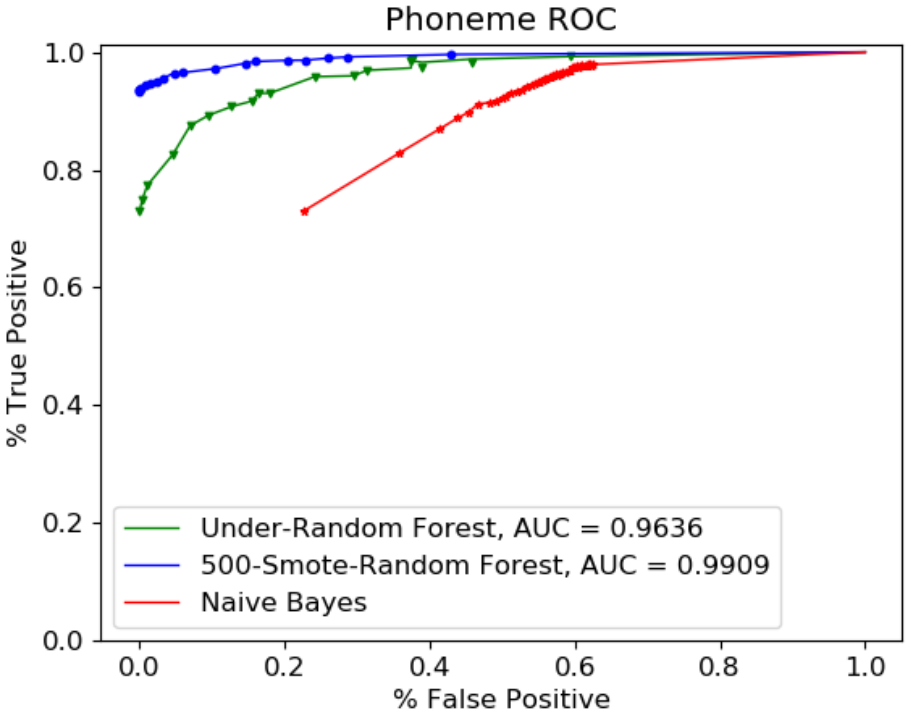
$$\Delta(X, Y) = \sum_{i=1}^{N} \delta(x_i, y_i)^r$$

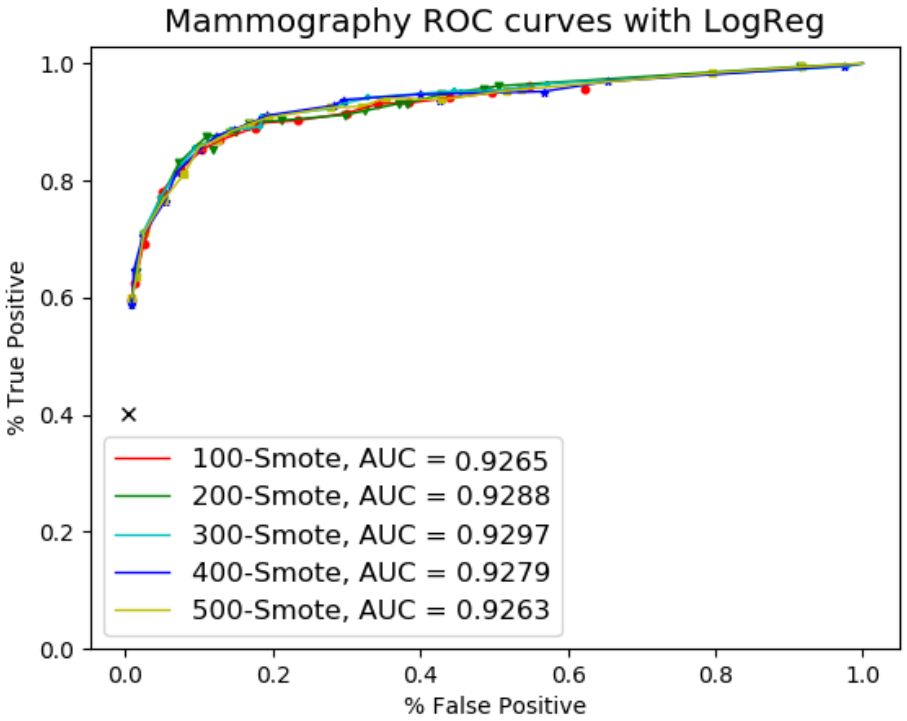
Эксперимент

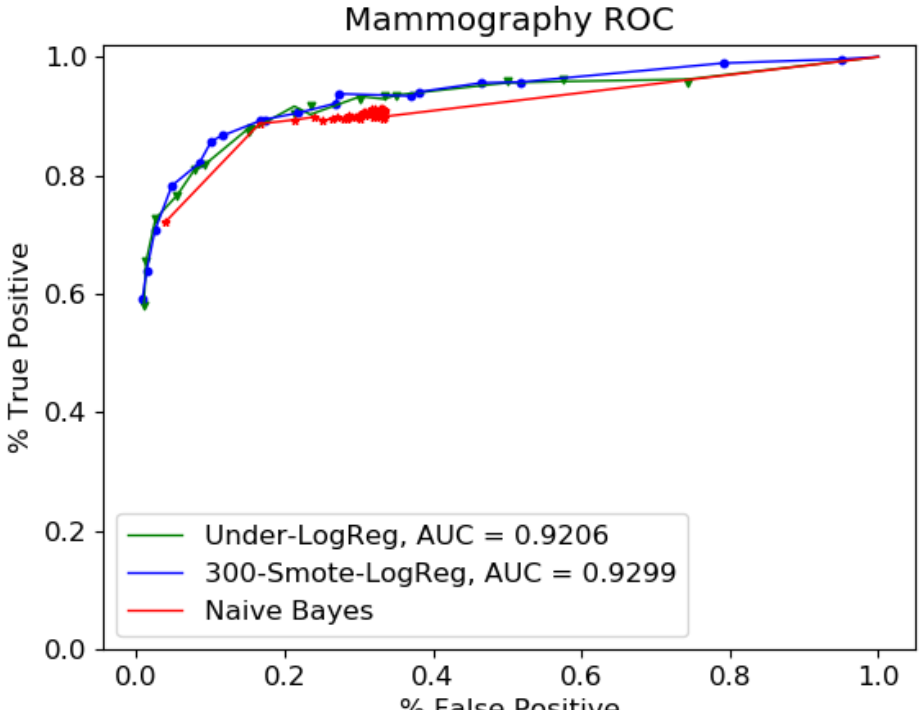


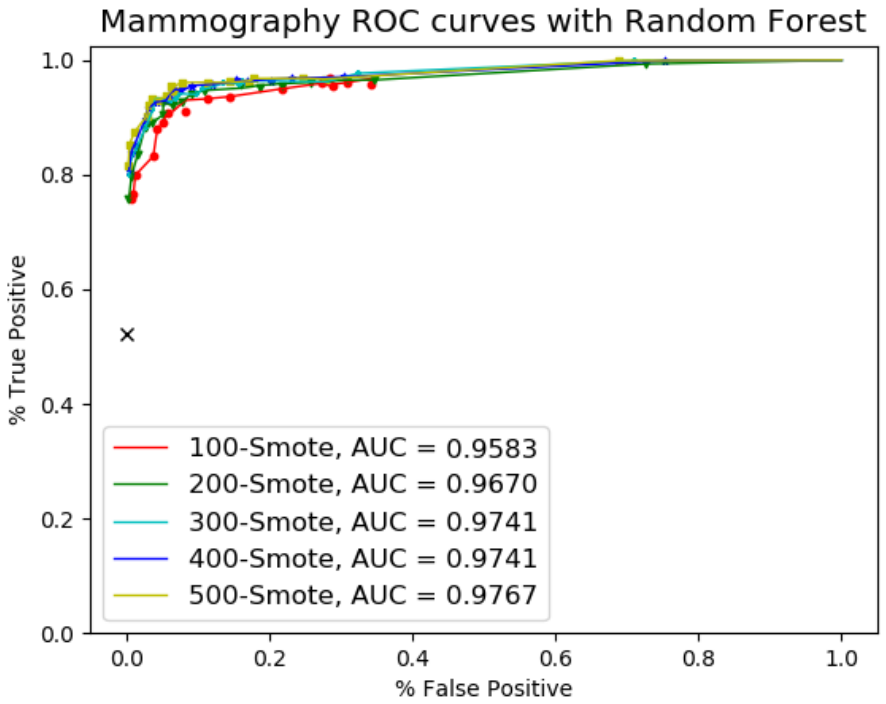


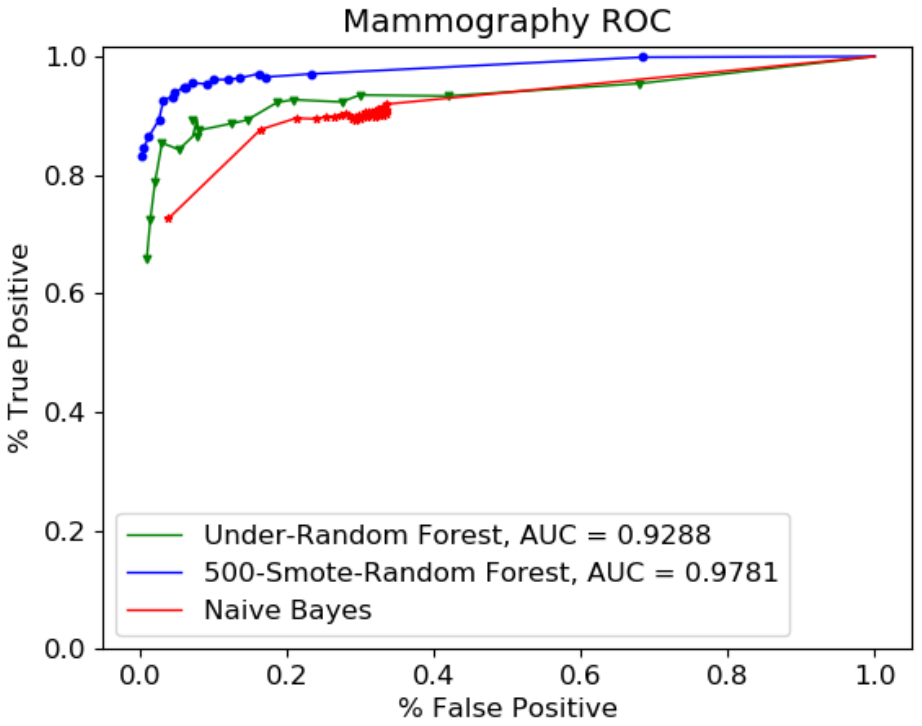

Phoneme ROC curves with Random Forest
% True Positive
% False Positive
100-Smote, AUC = 0.9780293924785305
200-Smote, AUC = 0.9845195574074413
300-Smote, AUC = 0.9871125110715655
400-Smote, AUC = 0.9893824886822955
500-Smote, AUC = 0.9907977832601095


Phoneme ROC
% True Positive
% False Positive
Under-Random Forest, AUC = 0.9636
500-Smote-Random Forest, AUC = 0.9909
Naive Bayes


Mammography ROC curves with LogReg
% True Positive
% False Positive
100-Smote, AUC = 0.9265
200-Smote, AUC = 0.9288
300-Smote, AUC = 0.9297
400-Smote, AUC = 0.9279
500-Smote, AUC = 0.9263


Mammography ROC
% True Positive
Under-LogReg, AUC = 0.9206
300-Smote-LogReg, AUC = 0.9299
Naive Bayes


Mammography ROC curves with Random Forest
% True Positive
% False Positive
100-Smote, AUC = 0.9583
200-Smote, AUC = 0.9670
300-Smote, AUC = 0.9741
400-Smote, AUC = 0.9741
500-Smote, AUC = 0.9767


Mammography ROC
% True Positive
% False Positive
Under-Random Forest, AUC = 0.9288
500-Smote-Random Forest, AUC = 0.9781
Naive Bayes

Результаты

Dataset	Majority Class	Minority Class	Ratio	Best result
Phoneme	3818	1586	5:2	Smote + UnderSampling
Mammography	10923	260	41:1	Smote + UnderSampling