

Нусратуллина Марьям Айратовна

Март 2026

1 Проделанная работа

Проверила предположение о том, что хаотичность пациентов из тестовой выборки может быть связана с их положением относительно пациентов из тренировочной выборки.

2 Введенные метрики

Метрики измеряют стабильность предсказаний пациента.

Для каждого пациента из тестовой выборки создается массив из вероятностей наличия болезни для всех random state и для всех итераций (или количества деревьев), начиная с 51.

2.1 Дисперсия

Вычисляется дисперсия этих предсказаний. Чем меньше дисперсия, тем более стабильны траектории.

2.2 Близость к 0.5

Пусть a_0 – количество предсказаний ≥ 0.5 , а $a_1 = N - a_0$ – количество предсказаний < 0.5 , где N – это общее количество предсказаний.

Смотрим на их нормированную следующим образом разницу:

$$1 - \frac{|a_0 - a_1|}{a_0 + a_1} = 1 - \frac{|N - 2a_0|}{N}$$

Чем меньше значение, тем меньше количество 'пересечений' 0.5.

3 Расстояние до train

Вначале точки нормируются (standart scaller может быть не лучшим инструментом, далее продумать, какая нормировка подойдет лучше). Далее точки из train делятся на 2 множества: X_0 – точки с таргетом 0 и X_1 – точки с таргетом 1.

Для каждой точки x из test вычисляются расстояния

$$d_{x,0} = \min_{y \in X_0} d(x, y)$$

$$d_{x,1} = \min_{y \in X_1} d(x, y)$$

Расстояния d :

1. Евклидово (l_2)
2. Манхэттенское (l_1)
3. Чебышева (l_∞)

Далее для каждой точки вычисляется модуль разности между расстоянием до X_0 и до X_1 : $|d_{x,0} - d_{x,1}|$.

Чем больше модуль разности между расстояниями, тем существеннее точка ближе к какому-то из множеств X_0 или X_1 .

4 Результаты для MLPclassifier(relu, adam)

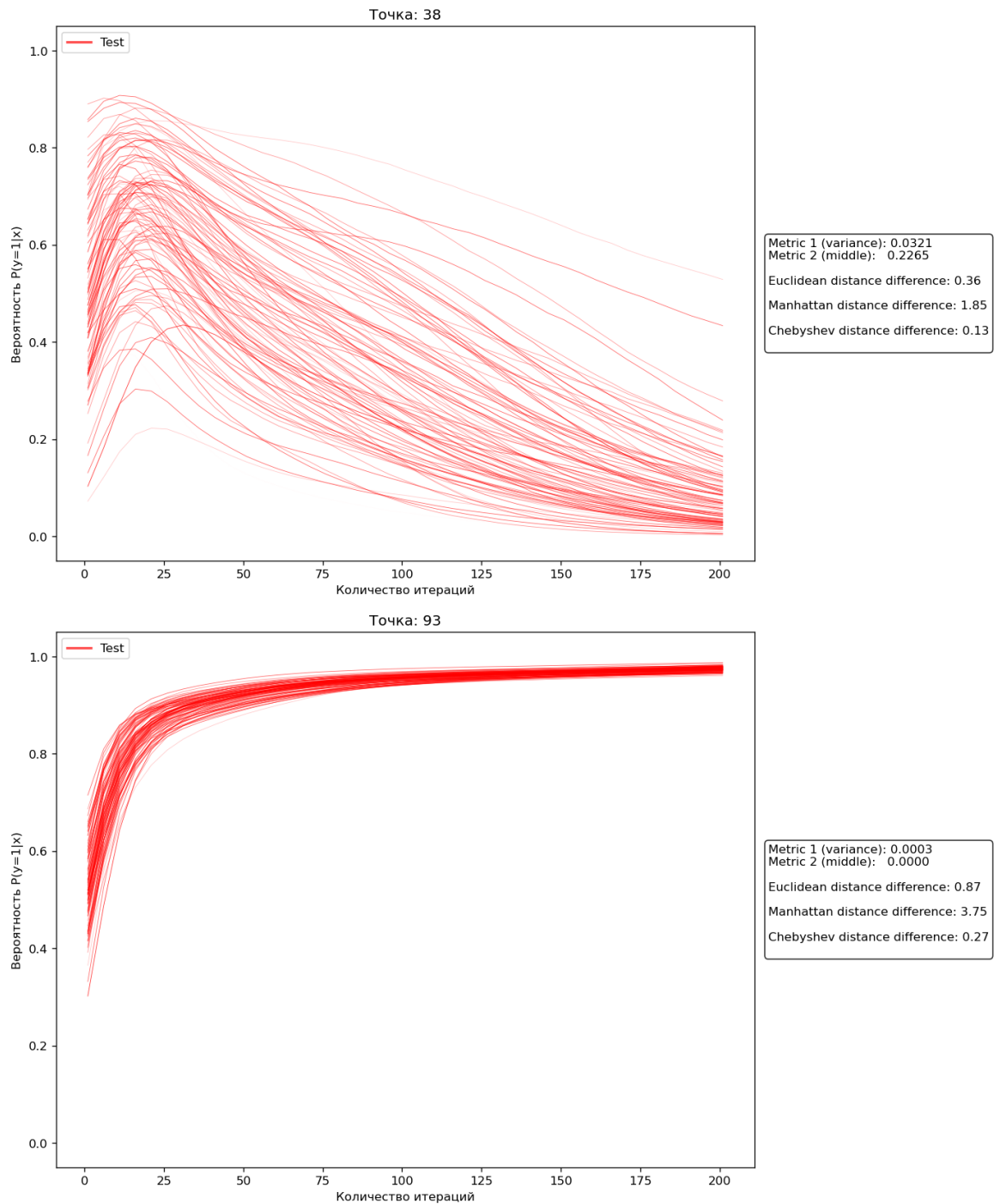
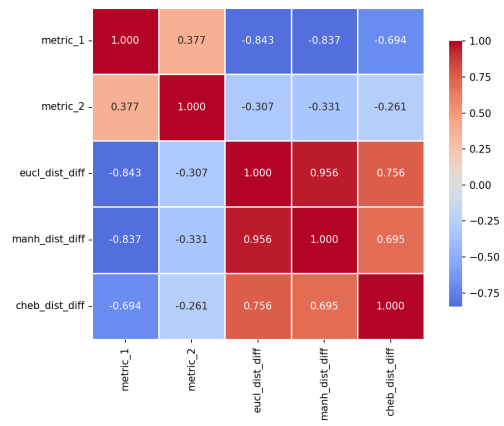


Рис. 1: Примеры траекторий и значений метрик MLPclassifier(relu, adam)

Оказалось, что существенность близости к одному из классов и значения метрики дисперсии имеют сильную отрицательную корреляцию, а значения метрики близости к 0.5 – слабую.



	metric_1	metric_2	eucl_dist_diff	manh_dist_diff	cheb_dist_diff
metric_1	0.0	0.0	0.0	0.0	0.0
metric_2	0.0	0.0	0.000044	0.00001	0.000552
eucl_dist_diff	0.0	0.000044	0.0	0.0	0.0
manh_dist_diff	0.0	0.00001	0.0	0.0	0.0
cheb_dist_diff	0.0	0.000552	0.0	0.0	0.0

5 Результаты для RandomForest

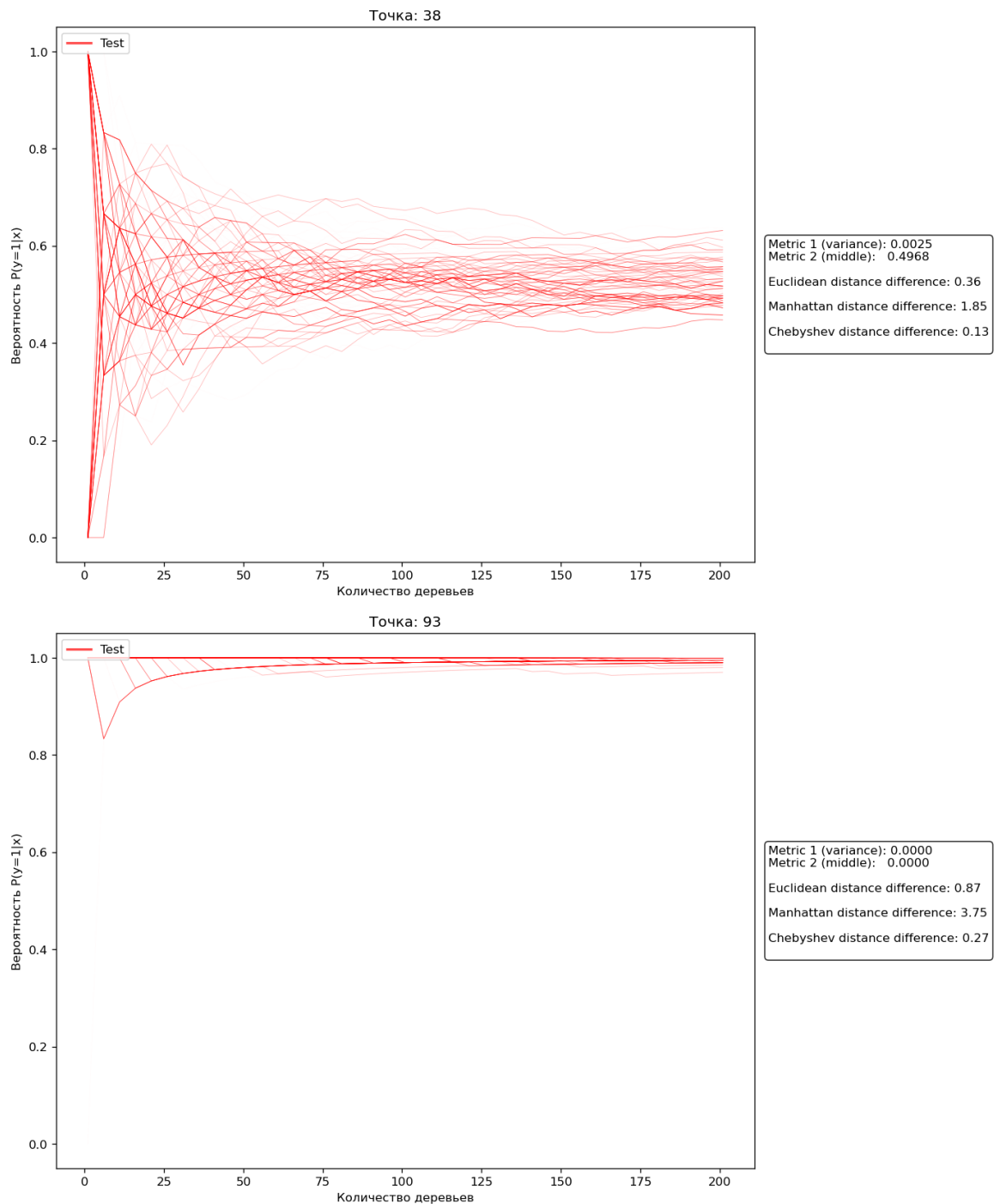
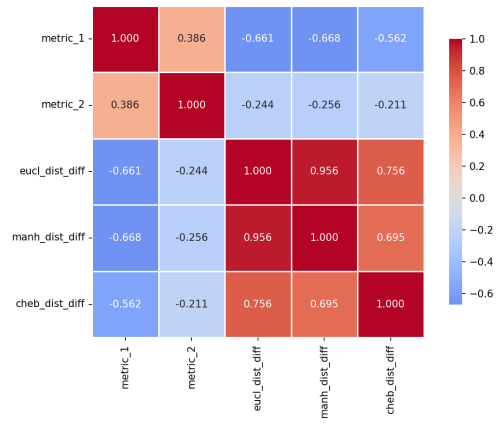


Рис. 2: Примеры траекторий и значений метрик RandomForest

Здесь значения коэффициентов корреляции меньше.



	metric_1	metric_2	eucl_dist_diff	manh_dist_diff	cheb_dist_diff
metric_1	0.0	0.0	0.0	0.0	0.0
metric_2	0.0	0.0	0.001271	0.00071	0.005587
eucl_dist_diff	0.0	0.001271	0.0	0.0	0.0
manh_dist_diff	0.0	0.00071	0.0	0.0	0.0
cheb_dist_diff	0.0	0.005587	0.0	0.0	0.0

6 Результаты для MLPclassifier(logistic, ibfgs)

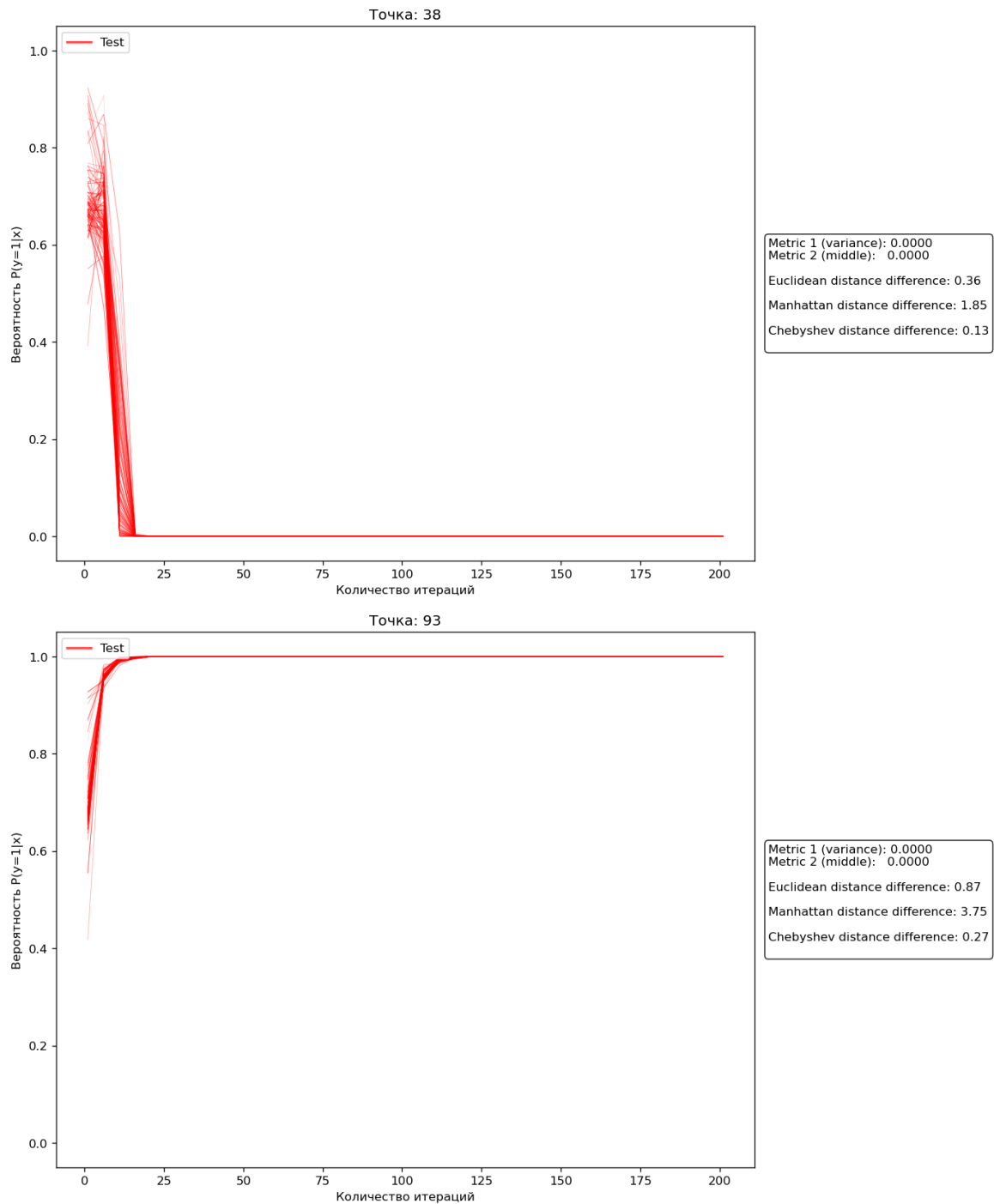
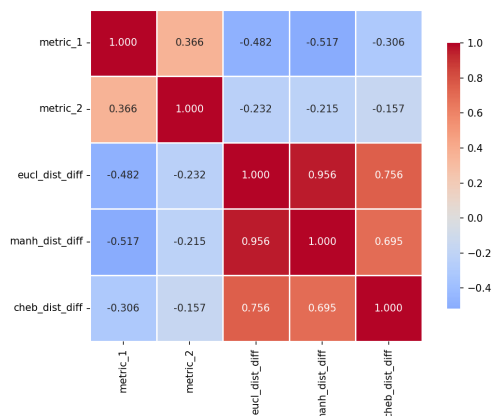


Рис. 3: Примеры траекторий и значений метрик MLPclassifier(logistic, ibfgs)

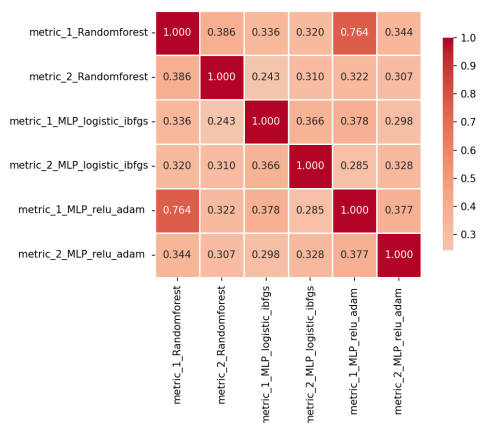
Здесь значения еще меньше.



	metric_1	metric_2	eucl_dist_diff	manh_dist_diff	cheb_dist_diff
metric_1	0.0	0.000001	0.0	0.0	0.000048
metric_2	0.000001	0.0	0.002229	0.004687	0.040258
eucl_dist_diff	0.0	0.002229	0.0	0.0	0.0
manh_dist_diff	0.0	0.004687	0.0	0.0	0.0
cheb_dist_diff	0.000048	0.040258	0.0	0.0	0.0

7 Общее сравнение

У дисперсии для MLPclassifier(relu, adam) и Randomforest высокая положительная корреляция. Для остального – умеренная и низкая. Может быть связано либо с данными, либо с устройством моделей.



	metric_1_Randomforest	metric_2_Randomforest	metric_1_MLP_logistic_ibfgrs	metric_2_MLP_logistic_ibfgrs	metric_1_MLP_relu_adam	metric_2_MLP_relu_adam
metric_1_Randomforest	0.0	1.8083390168032347e-07	7.240640973348188e-06	1.0189667411888548e-05	6.197872079670514e-34	4.169367887304463e-06
metric_2_Randomforest	8.083390168032347e-07	0.0	0.0013673688503355253	3.789304090189373e-05	7.071072440208454e-05	4.387710463478596e-05
metric_1_MLP_logistic_ibfgrs	7.240640973348188e-06	0.0013673688503355253	0.0	8.569919342854608e-07	4.520581587185897e-07	7.612365763684276e-05
metric_2_MLP_logistic_ibfgrs	1.0189667411888548e-05	3.789304090189373e-05	8.569919342854608e-07	0.0	0.0016087621426503197	2.177392588251985e-05
metric_1_MLP_relu_adam	6.197872079670514e-34	7.071072440208454e-05	4.520581587185897e-07	0.0016087621426503197	0.0	1.7108654457917383e-07
metric_2_MLP_relu_adam	4.169367887304463e-06	4.387710463478596e-05	7.612365763684276e-05	2.177392588251917e-05	1.7108654457917383e-07	0.0

8 Выводы

Нестабильность (хаотичность) предсказаний нейросетевых и ансамблевых моделей на тестовых данных может быть частично объяснена геометрическими свойствами расположения объектов относительно обучающей выборки.

9 Что планирую делать далее

1. (ГЛАВНОЕ) Заняться реальными данными
2. Понять, как лучше нормализовать данные для подсчета расстояний
3. Лучше смотреть на расстояние до k ближайших соседей или на плотность, а не на расстояние до ближайшего элемента

4. Доработать метрики хаоса