

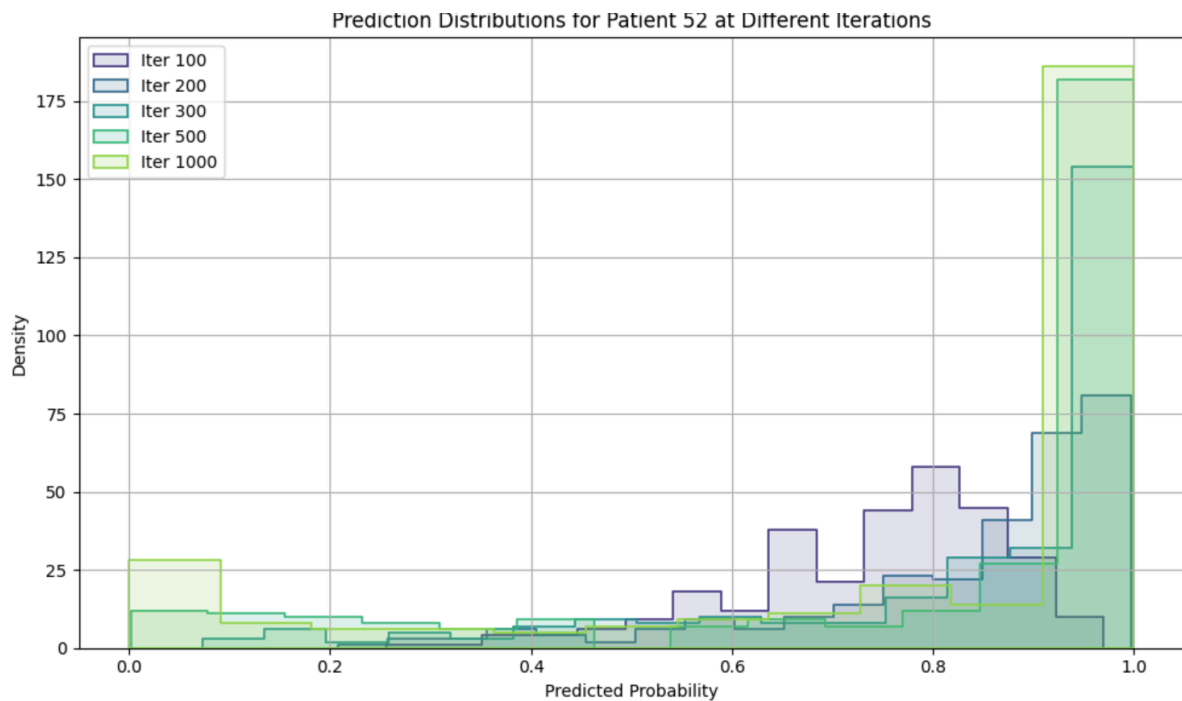
Иринчеева Татьяна Анатольевна

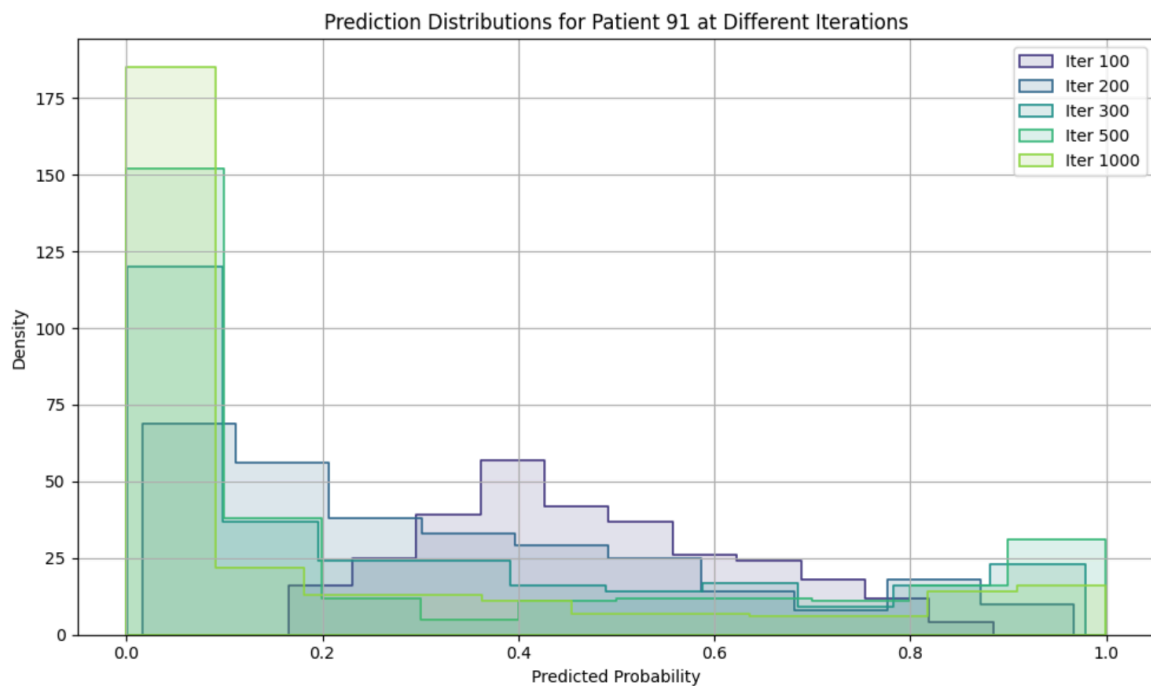
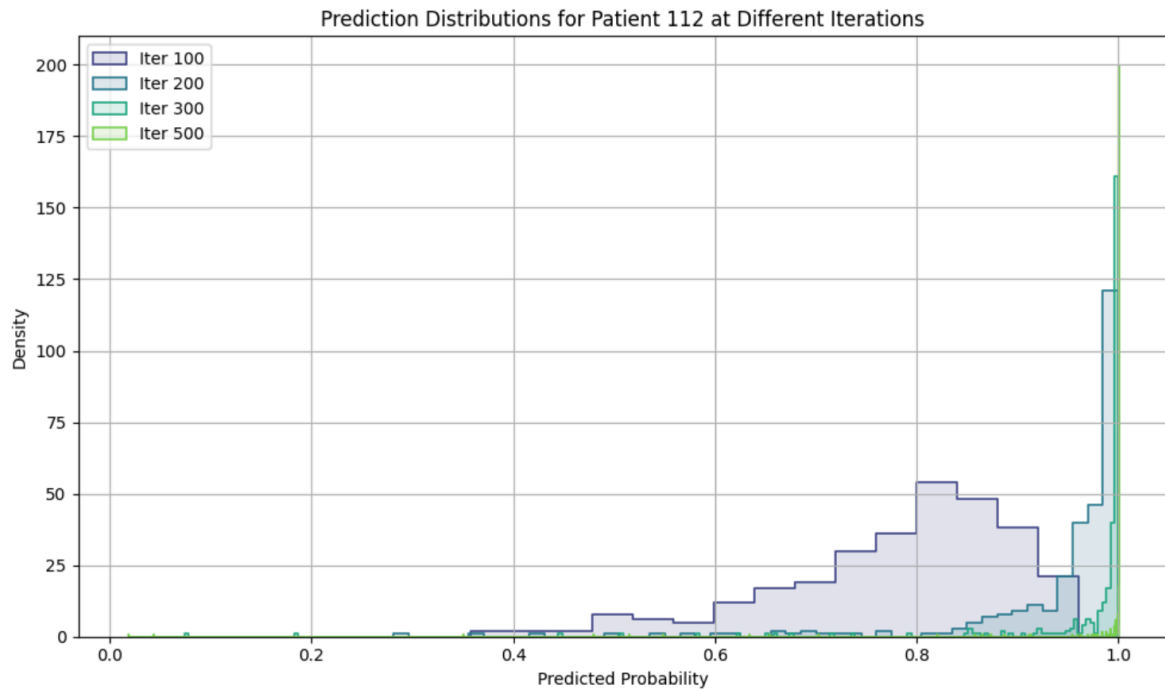
Что было кратко:

1. Смотрела на распределения предсказаний для пациентов, для которых дисперсия предсказаний максимальна → Увидела случаи с нормальным или равномерным распределением (любое распределение, отличающееся от столба в точке 0 или 1), значит модель чувствительна к сиду и можно установить факт невоспроизводимости

Что сделала:

2. Решила посмотреть на изменение формы распределений в зависимости от количества итераций. Посмотрела для тех же самых пациентов.

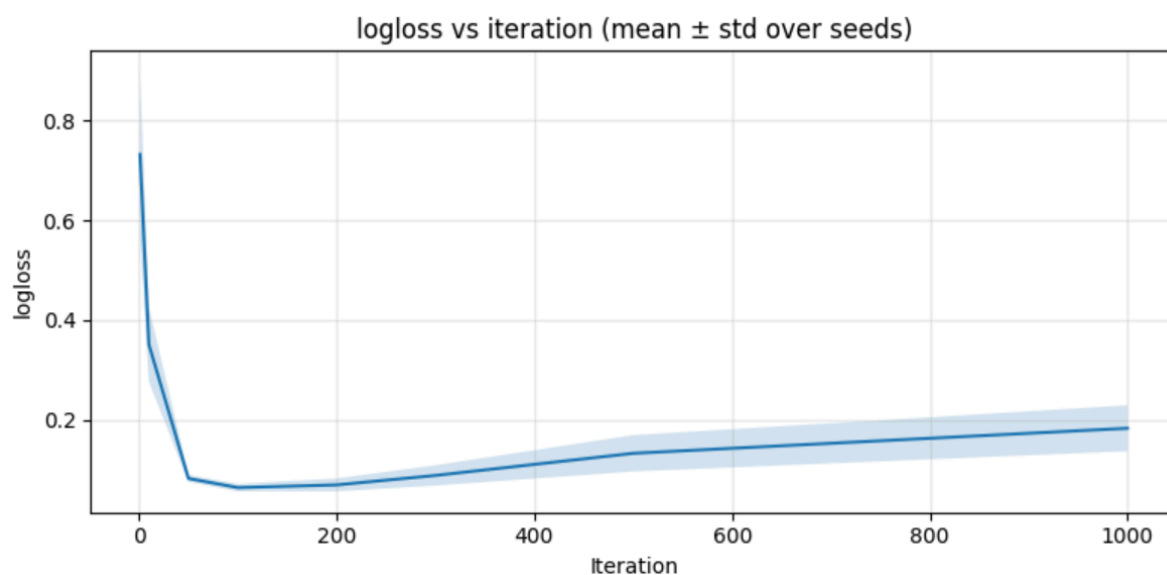
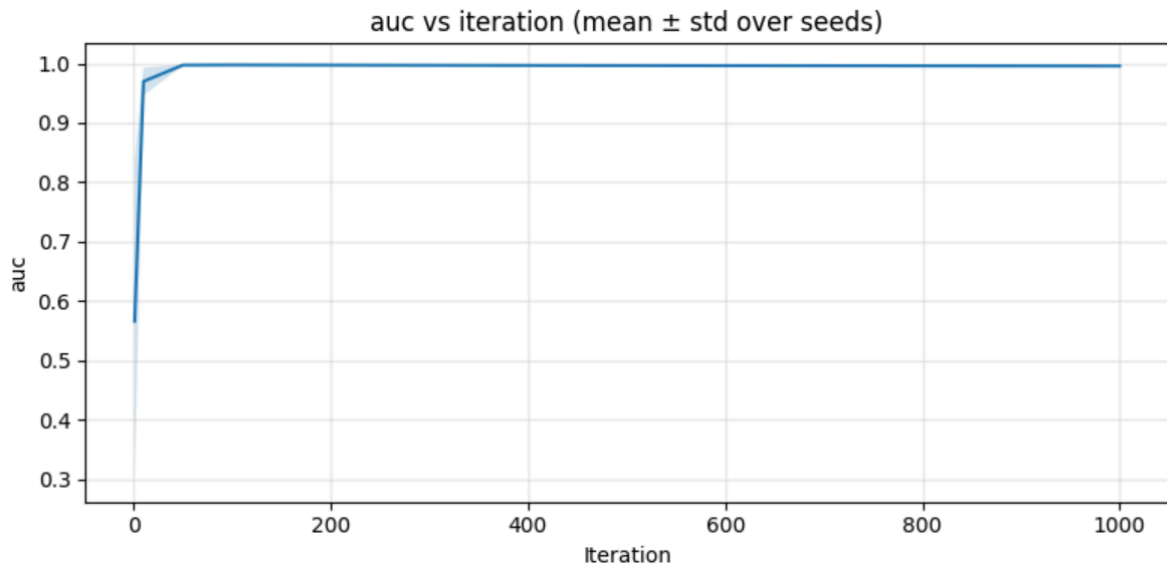




Вывод: по пациенту 112 на **100-ой** итерации распределение широкое и разброс большой, но с ростом числа итераций распределение сжимается к 1 и разброс снижается, поэтому можно сказать, что невоспроизводимость исчезает при обучении. Однако все же есть случаи (пациент 91 и 52), для которых невоспроизводимость не исчезает даже при большом кол-ве итераций.

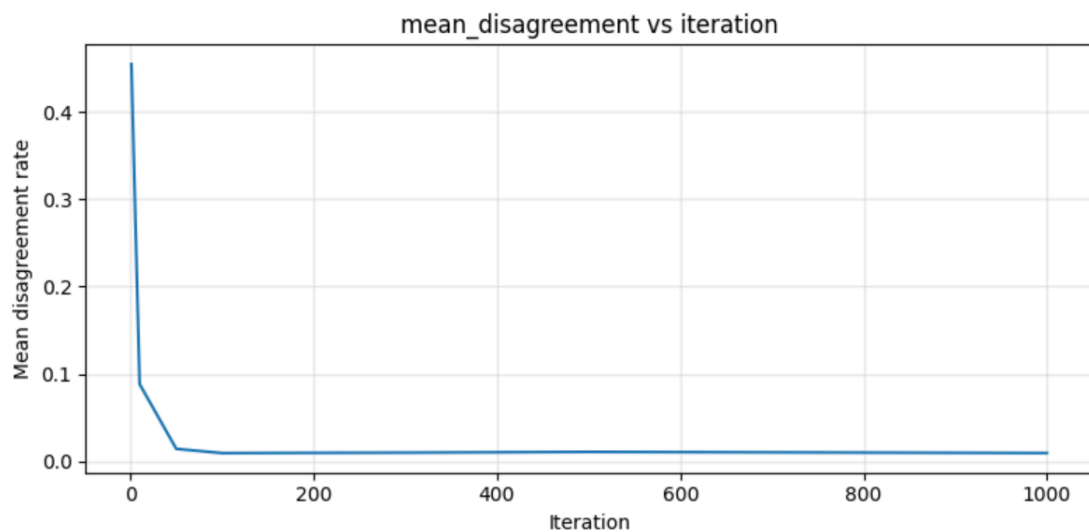
- Из этого я поняла, что до этого смотрела на распределения, обучив MLP classifier с дефолтными параметрами, то есть `max_iter = 100`. До этого внимания

на это не обращала, так как ассигасу у модели все равно был высокий. Стоит посмотреть на метрики качества в зависимости от числа итераций. Синим - среднее значение, голубая заливка вокруг кривой - отклонения по сидам.

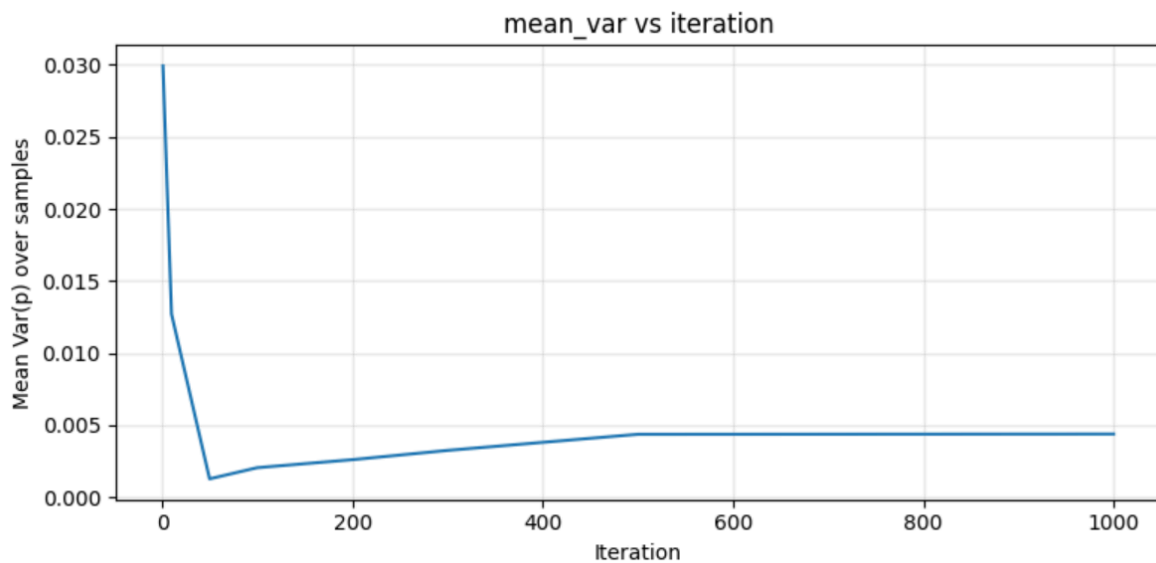


Вывод: видно, что для этого датасета (breast cancer) MLP classifier достигает высокого качества уже на 50 итерации и где-то после 200-ой итерации ошибка начинает увеличиваться, то есть модель начинает переобучаться. Логично предположить, что будет выбрано то количество итераций, которое обеспечивает хорошее качество и предотвращает модель от переобучения, хотя обученная с таким количеством итераций модель будет менее воспроизводима по предсказаниям.

4. При этом если смотреть на общую картину, а не предсказания по отдельным пациентам, то с ростом числа итераций классовые предсказания в среднем становятся практически полностью воспроизводимыми



и дисперсия тоже сходится к 0, но потом выходит на плато



2 глобальных вывода:

- Воспроизводимость предсказаний действительно повышается с ростом числа итераций, но все же остается подмножество объектов с высокой чувствительностью к сдвигу.

И вообще у меня возникло небольшое противоречие в голове: как будто при переобучении модель наоборот должна стать менее воспроизводимой. И я

пришла к тому, что высокая уверенность вообще не гарантирует воспроизводимость и нужно посмотреть на decision boundary.

- Такое число итераций, которое позволит обеспечить почти полную воспроизводимость предсказаний, не всегда возможно выбрать, так как оно часто приводит к переобучению. Получается такой трейд-офф и мы хотим повысить воспроизводимость модели, не доводя ее до переобучения. Логично предположить, что это стоит связать с регуляризацией.

5. Нашла [статью](#), которая подтверждает мысль из вывода 1. Ключевые выводы статьи такие:

1. Нестабильность решений напрямую связаны с тем, насколько близко decision boundary подходит к данным.
2. По мере обучения decision boundary приближается к обучающим и тестовым объектам, причём это продолжается даже на поздних эпохах, когда ассигасу уже почти не меняется.
3. сближение decision boundary с данными - не неизбежное свойство архитектуры, а следствие стандартного обучения с cross-entropy.

То есть можно сформулировать такую гипотезу: наблюдаемое сжатие вероятностей к 0 и 1 по мере обучения отражает рост уверенности классификатора, но не обязательно соответствует геометрической устойчивости и воспроизводимости. Это происходит потому что cross-entropy стимулирует рост логитов, но при этом геометрическая граница решений (decision boundary) может одновременно приближаться к данным.

Провела эксперимент, чтобы проверить что граница действительно отдаляется от объектов по мере обучения.

1. Определим расстояние до decision boundary таким образом:

$$d(x) \approx \frac{|f(x)|}{\|\nabla_x f(x)\|},$$

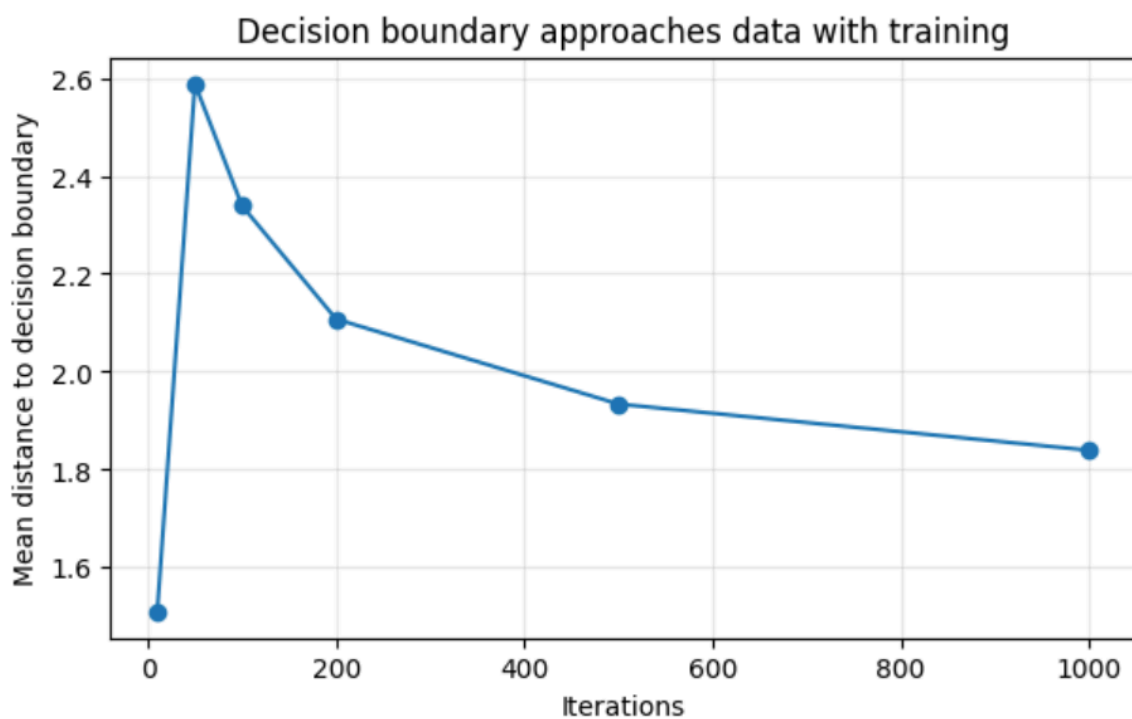
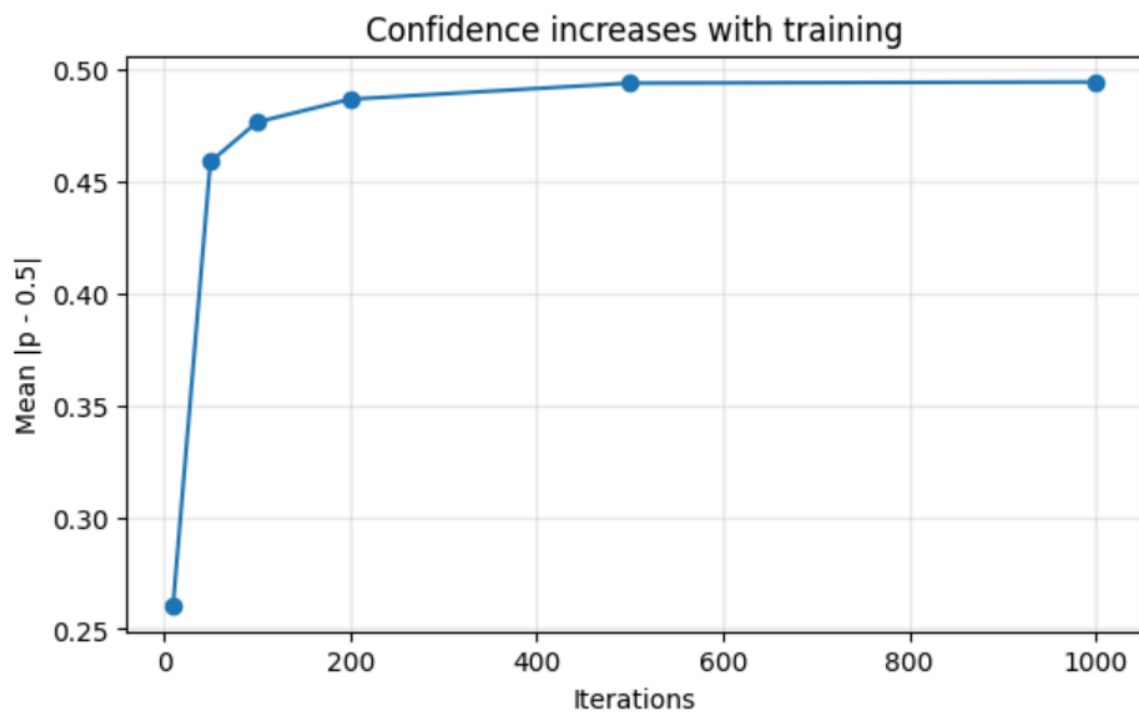
где $f(x) = 0$ - это граница решений (разность логитов) и

$\nabla_x f(x)$ - градиент этого score по входу.

Насколько маленькое L2-изменение нужно, чтобы пересечь границу.

2. Уверенность определим как $(|p-0.5|)$
3. Для разного числа итераций обучаем MLP с разными seed
4. Считаем уверенность и расстояние до decision boundary
5. Усредняем по объектам и seed и визуализируем зависимость

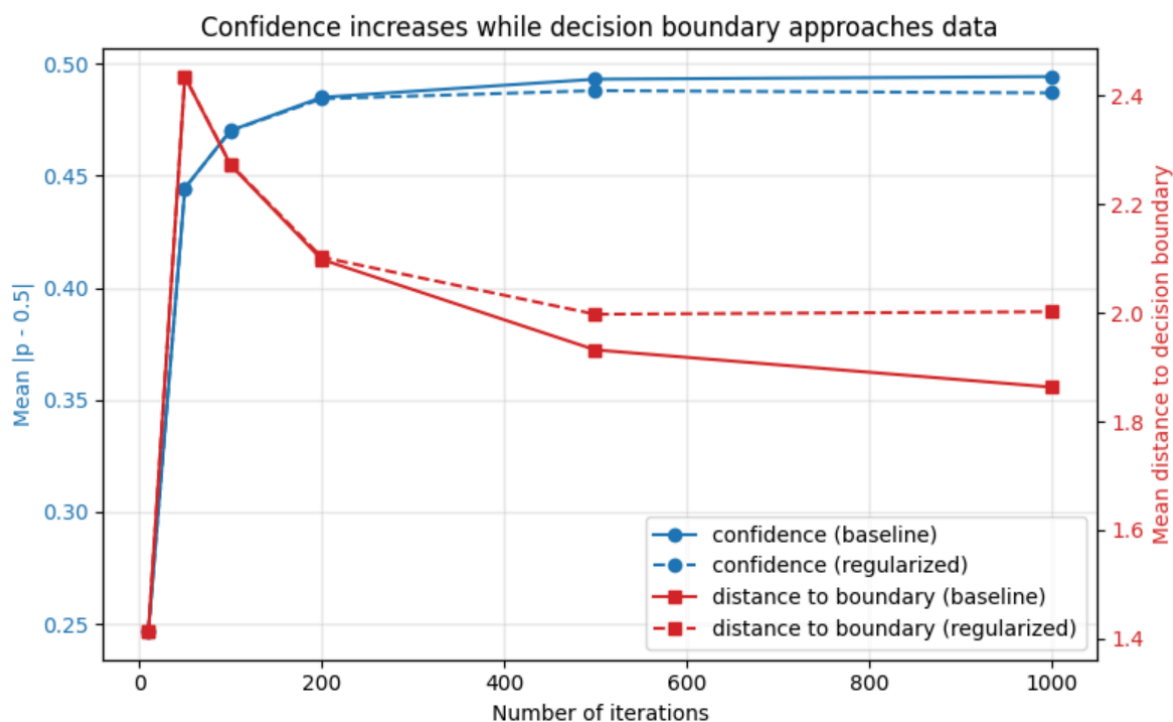
И действительно, уверенность растёт, тогда как decision boundary отдаляется от объектов



Это порождает некоторый trade-off, и как я уже сказала, возможно стоит это связать с регуляризацией.

Исследование влияния регуляризации на уверенность модели и устойчивость модели (расстояние до decision boundary)

1. baseline - MLP (alpha = 1e-4)
2. regularized - MLP (alpha = 1e-2)



Полученные результаты демонстрируют расхождение между вероятностной уверенностью и геометрической устойчивостью модели. Несмотря на то что предсказания становятся более уверенными по мере обучения, decision boundary систематически приближается к данным.

Регуляризация не мешает росту уверенности, но при этом обучение становится более устойчивым, то есть boundary decision не так быстро и не так близко подходит к данным и на более поздних итерациях остается на одном расстоянии.

Связь воспроизводимости и устойчивости

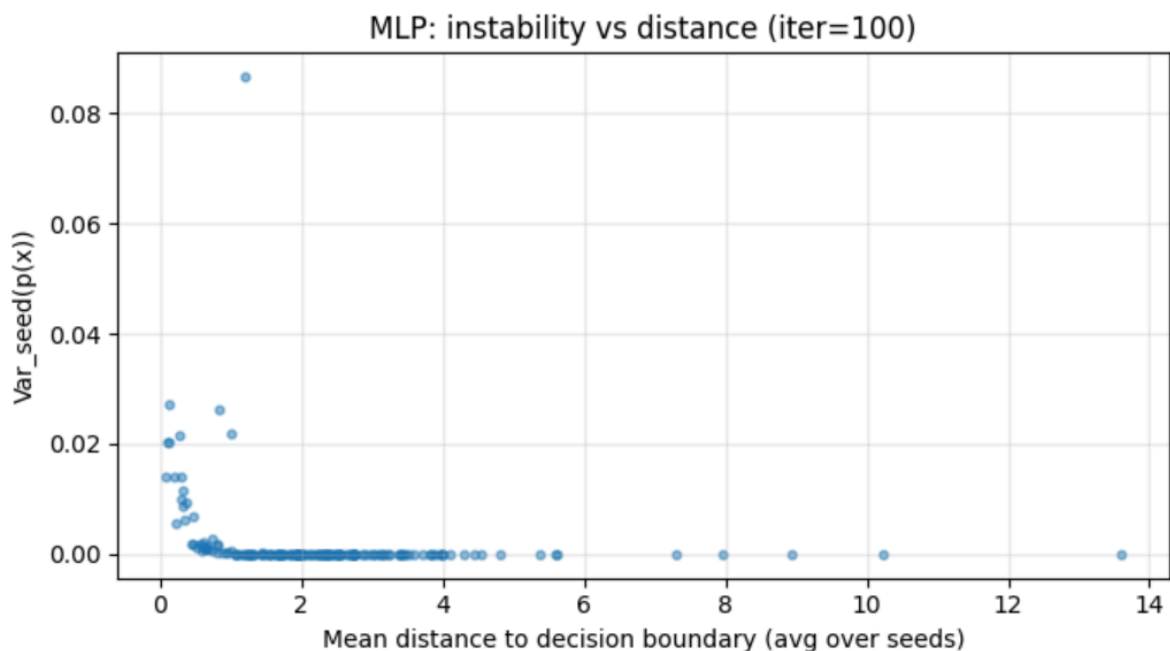
1. За устойчивость отвечает метрика расстояние от decision boundary
2. Воспроизводимость мы измеряли с помощью дисперсии предсказаний

Посчитали корреляцию между этими метриками:

Spearman $\rho = -0.975$, $p \approx 7 \cdot 10^{-113}$

То есть чем ближе объект к decision boundary, тем выше дисперсия вероятностных предсказаний при изменении random seed. Когда-то давно мы выдвигали такую гипотезу, и она подтвердилась.

Хотя график не такой убедительный:



Посмотрели на объекты, которые находятся рядом с decision boundary и далеко от нее:

Группа	mean var_p
Bottom 10% (ближе всего к boundary)	0.0109
Top 10% (дальше всего)	≈ 0

Разница в значениях небольшая, но она все-таки показывает некоторую закономерность:

- предсказания для объектов, которые далеко от decision boundary, почти полностью воспроизводимы
- для объектов, которые близко находятся к decision boundary, предсказаний более вариативны

Эксперимент для ансамблей

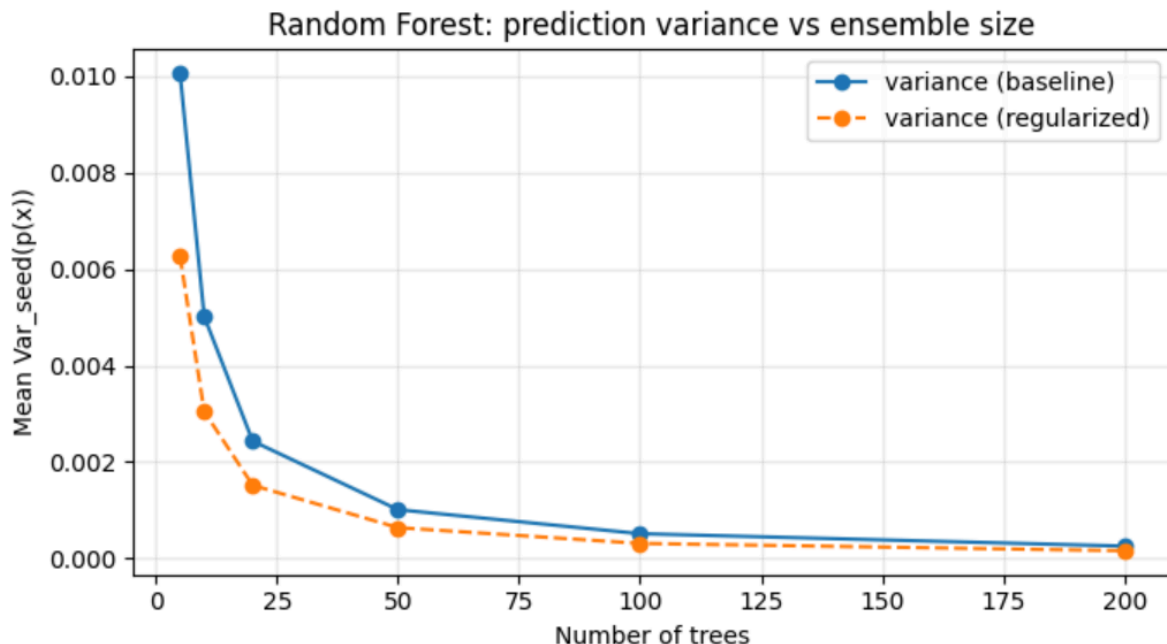
Чем этот эксперимент отличается от эксперимента с MLP?

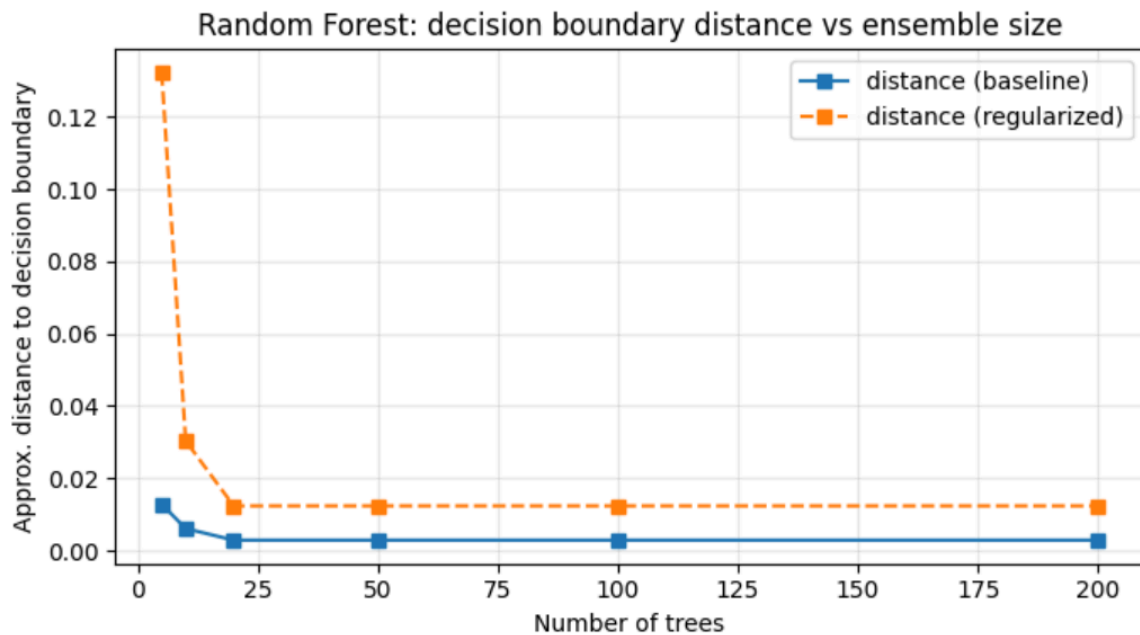
1. Расстояние до boundary decision определяется по-другому. Определение через градиент нам больше не подходит, потому что функция не дифференцируема как в MLP, а кусочно-постоянна, градиент почти везде равен 0. Но интуиция остается та же: минимальная величина возмущения входа при которой меняется предсказание. Посмотрела в интернете, как можно определить это для деревьев/ансамблей, нашла только такой [способ](#) (для отдельного дерева): рекурсивно обходятся все пути дерева, для каждого альтернативного пути

минимально изменяем признак, чтобы перепрыгнуть порог, доходим до листа, и если класс другой → кандидат. Однако для ансамблей деревьев граница решений формируется комбинацией решений множества деревьев, и ближайшая граница ансамбля не обязана совпадать с границей какого-либо отдельного дерева. В связи с этим для ансамблей используется приближенная оценка расстояния, мы просто берем минимум по кандидатам и говорим, что расстояние до decision boundary для ансамбля - минимум по кандидатам. В GB я не смогла пока перенести этот подход потому что в Gradient Boosting Classifier базовые модели DecisionTreeRegressor, а не классификаторы.

2. До этого я рассматривала трейдофф между уверенностью и расстоянием до границы решений. В случайном лесу же немного нелогично смотреть на уверенность, потому что число деревьев почти не влияет на уверенность, а влияет на дисперсию предсказаний (она уменьшается), поэтому будем смотреть на эту метрику. Соответственно, графики для двух метрик я разнесла в два, потому что тренд у этих метрик одинаковый с ростом числа деревьев, на одном графике они сливаются.

Какой вывод по этому эксперименту?





Радует, что результаты как будто бы повторяются и для этой модели:

- Дисперсия падает с ростом числа деревьев → модель становится более воспроизводимой по предсказаниям
- При этом граница подходит к данным (но не сильно, масштаб значений слишком крупный и потом она стабилизируется). Однако стоит помнить, что мы не совсем точно определили расстояние до границы решений, поэтому более точная интерпретация будет выглядеть так: при большем числе деревьев возрастает шанс, что хотя бы одно дерево имеет порог рядом с точкой → минимум по деревьям уменьшается
- С регуляризацией и дисперсия предсказаний меньше, и граница к данным не так близко подходит, но опять же масштаб крупный, разница не такая уж и большая.

Вопросы:

1. Можно ли таким образом определять расстояние до границы? Имеет ли результат с такой аппроксимацией вообще ценность?
2. Как можно и нужно ли вообще определить расстояние до границы у GB?

Следующий шаг:

Понять, как все мои эксперименты связать с поставленной задачей по построению устойчивой модели минимальной сложности для данных.