

FEDERAL STATE AUTONOMOUS
EDUCATIONAL INSTITUTION OF HIGHER EDUCATION
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY
(STATE UNIVERSITY)
PHYSTECH-SCHOOL OF APPLIED MATHEMATICS AND COMPUTER
SCIENCE

Homework.

Bayes selection models

4-th year student, group B05-003
Babkin Petr, Kreinin Matvei, Nikitina Maria, Semkin Kirill

Moscow, 2023

Содержание

| | |
|---|-----------|
| 1 Выборка №1 | 3 |
| 1.1 Описание выборки | 3 |
| 1.2 Преобразования над выборкой | 3 |
| 1.3 Модель | 4 |
| 1.4 Результаты | 4 |
| 2 Выборка №2 | 5 |
| 2.1 Описание выборки | 5 |
| 2.2 Преобразования над выборкой | 5 |
| 2.3 Модель | 6 |
| 2.4 Результаты | 6 |
| 3 Выборка №3 | 7 |
| 3.1 Описание выборки | 7 |
| 3.2 Преобразования над выборкой | 7 |
| 3.3 Модель | 8 |
| 3.4 Результаты | 8 |
| 4 Выборка №4 | 9 |
| 4.1 Описание выборки | 9 |
| 4.2 Преобразования над выборкой | 9 |
| 4.3 Модель | 9 |
| 4.4 Результаты | 10 |
| 5 Выборка №5 | 11 |
| 5.1 Описание выборки | 11 |
| 5.2 Преобразования над выборкой | 11 |
| 5.3 Модель | 11 |
| 5.4 Результаты | 12 |
| 6 Выборка №6 | 13 |
| 6.1 Описание выборки | 13 |
| 6.2 Преобразования над выборкой | 13 |
| 6.3 Модель | 13 |
| 6.4 Результаты | 13 |
| 7 Выборка №7 | 15 |
| 7.1 Описание выборки | 15 |
| 7.2 Преобразования над выборкой | 15 |
| 7.3 Модель | 16 |
| 7.4 Результаты | 16 |

| | |
|--|-----------|
| 8 Выборка №8 | 18 |
| 8.1 Описание выборки | 18 |
| 8.2 Преобразования над выборкой | 18 |
| 8.3 Модель | 18 |
| 8.4 Результаты | 18 |
| 9 Выборка №9 | 20 |
| 9.1 Описание выборки | 20 |
| 9.2 Преобразования над выборкой | 20 |
| 9.3 Модель | 20 |
| 9.4 Результаты | 20 |
| 10 Выборка №10 | 22 |
| 10.1 Описание выборки | 22 |
| 10.2 Преобразования над выборкой | 22 |
| 10.3 Модель | 22 |
| 10.4 Результаты | 23 |
| 11 Выборка №11 | 24 |
| 11.1 Описание выборки | 24 |
| 11.2 Преобразования над выборкой | 24 |
| 11.3 Модель | 24 |
| 11.4 Результаты | 25 |
| 12 Выборка №12 | 26 |
| 12.1 Описание выборки | 26 |
| 12.2 Модель | 26 |
| 12.3 Результаты | 26 |
| 13 Выборка №13 | 27 |
| 13.1 Описание выборки | 27 |
| 13.2 Преобразования над выборкой | 27 |
| 13.3 Модель | 27 |
| 13.4 Результаты | 28 |
| 14 Выборка №14 | 29 |
| 14.1 Описание выборки | 29 |
| 14.2 Преобразования над выборкой | 29 |
| 14.3 Модель | 29 |
| 14.4 Результаты | 30 |
| 15 Выборка 1, 2, 4, 7, 8 и 9 | 31 |
| 15.1 Описание выборки | 31 |
| 15.2 Преобразования над выборкой | 31 |

1 Выборка №1

1.1 Описание выборки

Выборка состоит из 1000 элементов и 20 признаков. Дисбаланс классов отсутствует (доля класса 1: 0.51). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p - value \geq 0.997$ говорит о том, что выбросов нет (это было так же проверено построением гистограмм распределения). Статистика $W \geq 0.995$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

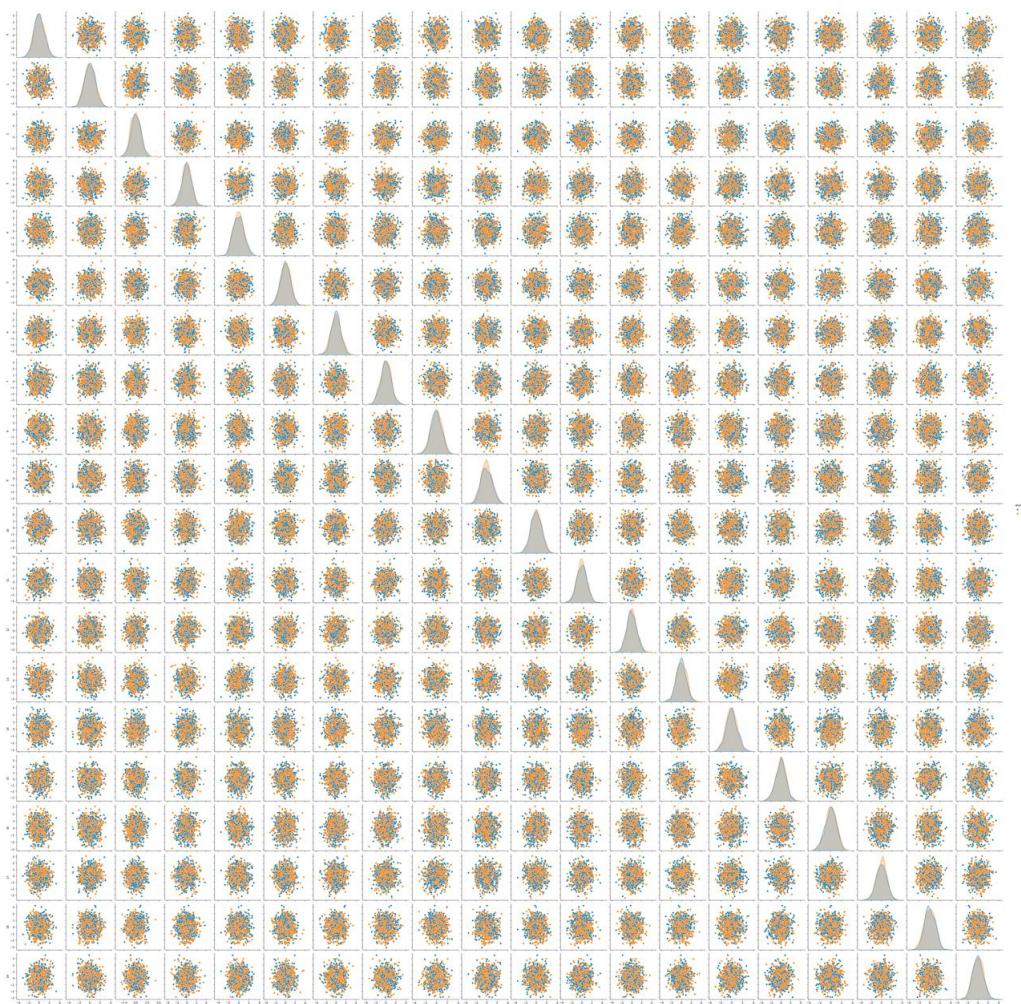


Рис. 1: Распределения признаков для выборки №1

1.2 Преобразования над выборкой

Была прверена корреляция между признаками вообще и внутри каждого класса по отдельности. Признаки скоррелированы слабо. Было посчитана KL-дивергенция для распределений одного и того же признака среди разных меток. По результатам отобраны 4 признака

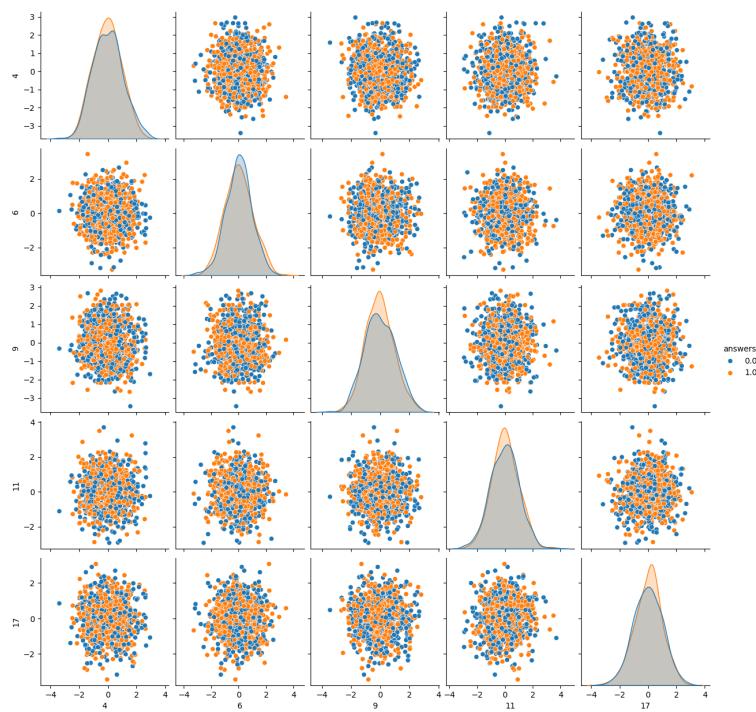


Рис. 2: Распределения отобранных признаков №1

1.3 Модель

Была выбрана модель наивного гауссовского классификатора, так как она больше всего подходит под описанную модель данных.

1.4 Результаты

Для каждой метрики (кроме AUC) была подобрана оптимальная граница первого класса.

Таблица 1: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|------------|-------------|-------|--------|--------|
| GaussianNB | 0.534 | 457 | -3902 | 526 |

2 Выборка №2

2.1 Описание выборки

Выборка состоит из 1000 элементов и 20 признаков. Присутствует небольшой дисбаланс классов (доля класса 1: 0.468). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p-value \geq 0.997$ говорит о том, что выбросов нет (это было так же проверено построением гистограмм распределения). Статистика $W \geq 0.995$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

2.2 Преобразования над выборкой

По построенным диаграммам распределения можно увидеть, что некоторые признаки заметно разделяют данные.

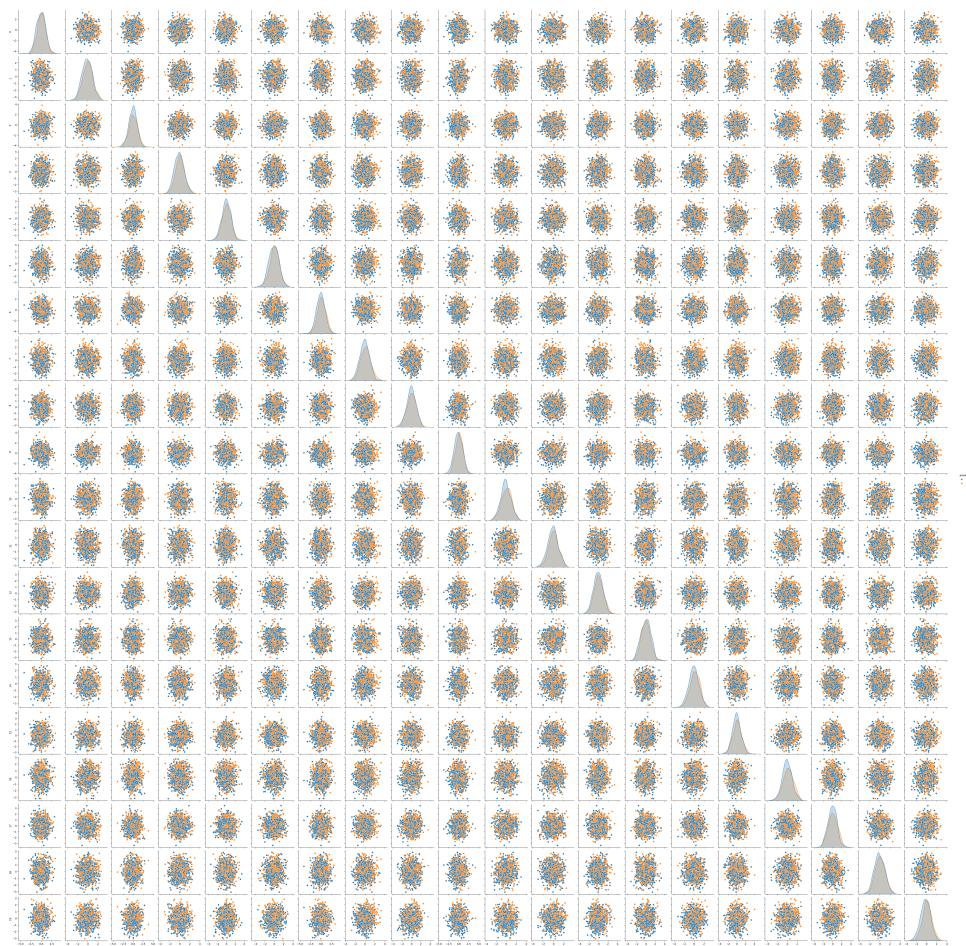


Рис. 3: Диаграмма распределений для всех признаков

Было замечено, что признаки нулевого класса в среднем больше, чем признаки первого класса. Поэтому был введен новый признак, равный сумме остальных, который хорошо разделяет данные.

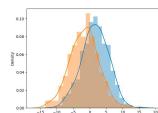


Рис. 4: Диаграмма распределений для нового признака

Предсказания делались на его основании.

2_chosen.png

Рис. 5: Диаграмма распределений для выбранных признаков

2.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB. Подсчитываются метрики. Было решено использовать GaussianNB.

2.4 Результаты

Для каждой метрики (кроме AUC) была подобрана оптимальная граница первого класса.

Таблица 2: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|------------|-------------|-------|--------|--------|
| GaussianNB | 0.709 | 344 | -4310 | 407 |

3 Выборка №3

3.1 Описание выборки

Выборка состоит из 1000 элементов и 20 признаков. Дисбаланс классов отсутствует (доля класса 1: 0.488). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p - value \geq 0.997$ говорит о том, что выбросов нет (это было так же проверено построением гистограмм распределения). Статистика $W \geq 0.995$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

3.2 Преобразования над выборкой

По построенной матрице корреляции можно заметить, что существует один признак (5), который коррелирует с целевыми значениями.

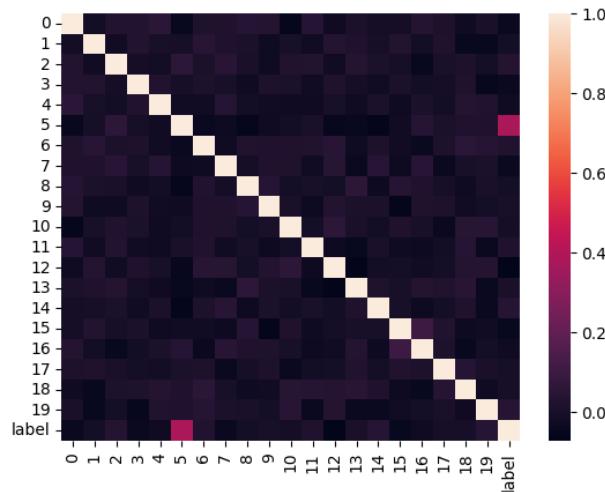


Рис. 6: Матрица корреляции для выборки №3

Был проведён ещё один анализ, подтверждающий целесообразность использования данного признака для обучения модели. Так как признаки распределены нормально для каждого класса, то следует посчитать их матожидание и дисперсию, а затем найти меру различия распределений. В качестве такой меры берём дивергенцию Кульбака-Лейблера.

$$KL = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 + \mu_1)^2}{\sigma_1^2} - 1$$

Отбирается n признаков с наибольшей дивергенцией. n находится подбором в процессе модели при разных значениях n . Итоговый результат для LogisticRegression и GaussianNB: $n = 1$.

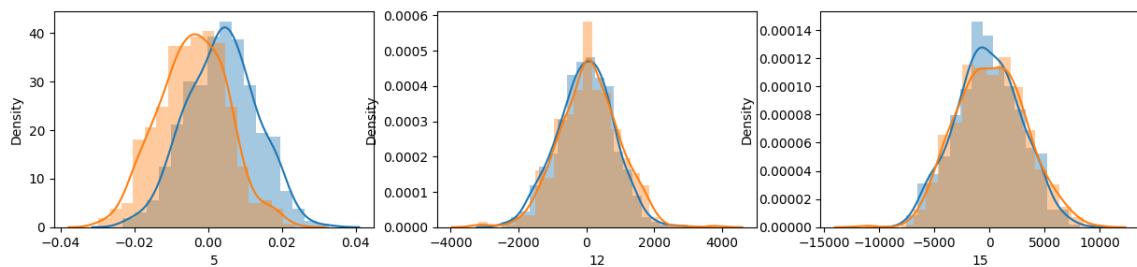


Рис. 7: Распределение 3 признаков с наибольшим значением KL для выборки №3

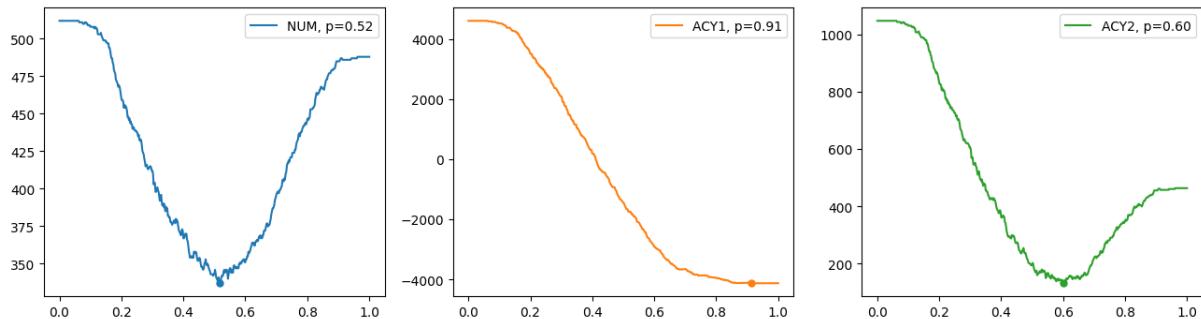


Рис. 8: Пороги отнесения к классу 1 для модели GaussianNB для выборки №3

3.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB. Подсчитываются метрики. Модели имеют практически одинаковые результаты. Возьмём GaussianNB, так как он немного лучше.

3.4 Результаты

Таблица 3: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.719 | 337 | -4122 | 132 |
| LogisticRegression | 0.719 | 343 | -4121 | 135 |

4 Выборка №4

4.1 Описание выборки

Выборка 1000 элементов и 50 признаков, дисбаланс классов присутствует доля элементов единичного класса всего 0.164. Потом проверили гипотезу о нормальном распределении признаков, для этого воспользовался критерием Шапиро-Уилка, по нему получили, что $p_{value} \geq 0.997$, т.е. выбросы в выборке отсутствуют, также это было визуально проверено, статистика же Шапиро-Уилка ≥ 0.995 , т.е. гипотеза о нормальном распределении признаков не отклоняется.

4.2 Преобразования над выборкой

Построили матрицу корреляции, видно, что есть группы признаков одинаково распределенных признаков, которые отличаются лишь порядком значений, было принято значение выбрать по одному признаку из таких групп, таким образом осталось 14 признаков. Затем было подсчитана дивергенция Кульбака-Лейблера между распределениями классов по каждой признаку, затем признаки были отсортированы по этому расстоянию между распределениями.

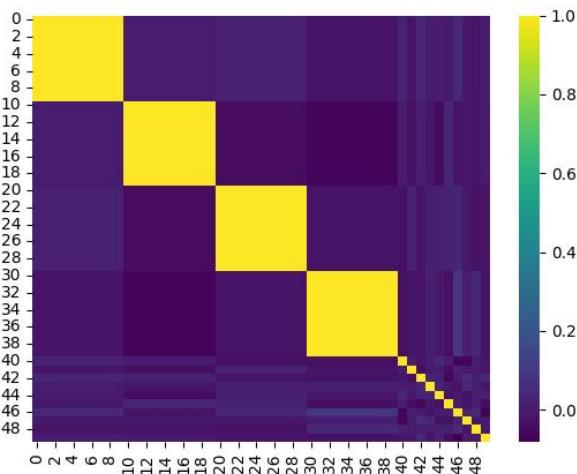


Рис. 9: Матрица корреляции для выборки №4

4.3 Модель

В качестве модели были выбран Logregression и GaussianNB из sklearn, в 10 фолдов в cross validation обучалась модель, метрика AUC усреднялась и вот по этому критерию мы измеряли качество модели, также перебиралось количество признаков, которое подается в саму модель, оказалось оптимальным число 4, логистическая регрессия оказалась хуже, чем GaussianNB. У логрегрессии получилось 0.556, у Gaussian 0.566. Также в качестве модели были попробованы черные коробка от

yandex, catboost classifier. Сначала было принято решение загрузить это все на две rtx 3080 и в тупую прогнать, чтобы итераций было побольше, но результаты были на уровне логрессии, потом в качестве оптимизируемой метрики был выбраны aucs и перебиралось оптимальное количество итераций, на удивление черная коробка показала хорошие средние результаты и aucs стал 0.584, но посмотрев на результаты от разных фолдов стало понятно, что модель крайне нестабильна и иногда она выдает aucs порядка 0.7, а иногда порядка 0.4. Далее был проанализирован трэшхолд по трем оставшимся метрикам, который нужно выбирать для минимизации значений, и из-за дизбаланса классов его нужно выбирать 1.0, это в целом логично т.к. у нас присутствует явный дизбаланс классов, поэтому последние метрики подсчитывались на все обучающей выборке.

4.4 Результаты

Таблица 4: Метрики ($ROC - AUC$ на кроссвалидации, NUM , $ASY1$, $ASY2$ на всей обучающей выборке

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.566 | 164 | 164 | -508 |
| LogisticRegression | 0.556 | 164 | 164 | -508 |
| CatboostClassifier | 0.584 | 164 | 164 | -508 |

5 Выборка №5

5.1 Описание выборки

Выборка состоит из 1000 элементов и 5000 признаков. Дисбаланс классов присутствует (доля класса 1: 0.373). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p - value \geq 0.997$ говорит о том, что выбросов нет (это было так же проверено построением гистограмм распределения). Статистика $W \geq 0.99$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

5.2 Преобразования над выборкой

Так как признаки распределены нормально для каждого класса, то следует посчитать их матожидание и дисперсию, а затем найти меру различия распределений. В качестве такой меры берём дивергенцию Кульбака-Лейблера.

$$KL = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 + \mu_1)^2}{\sigma_1^2} - 1$$

Отбирается n признаков с наибольшей дивергенцией. n находится подбором в процессе модели при разных значениях n . Итоговый результат для LogisticRegression и GaussianNB: $n = 3$.

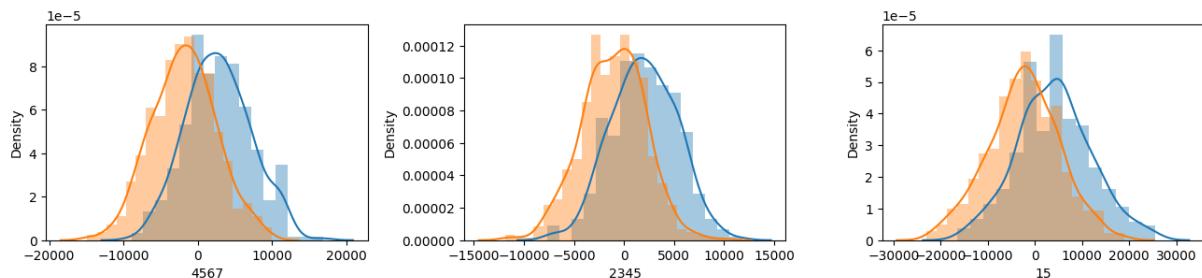


Рис. 10: Распределение 3 признаков с наибольшим значением KL для выборки №5

5.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB. Подсчитываются метрики. Обе модели показывают себя хорошо на выборках. Однако, LogisticRegression лучше показывает себя на всех метриках.

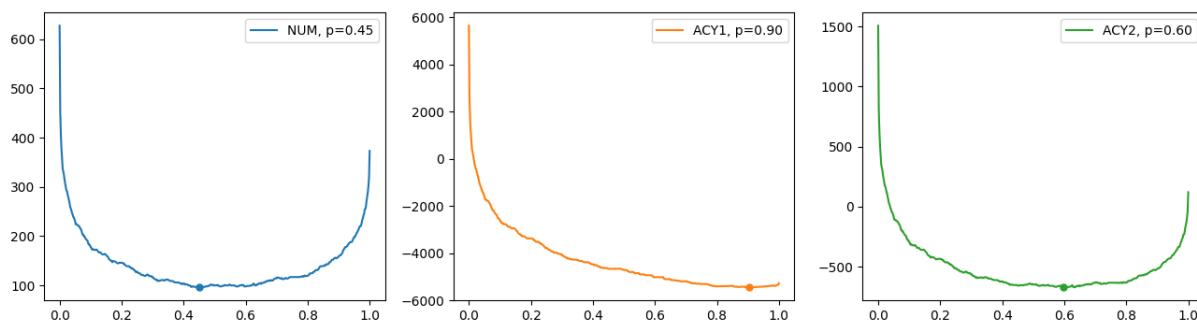


Рис. 11: Пороги отнесения к классу 1 для модели LogisticRegression для выборки №5

5.4 Результаты

Таблица 5: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.963 | 104 | -5434 | -652 |
| LogisticRegression | 0.965 | 96 | -5448 | -675 |

6 Выборка №6

6.1 Описание выборки

Выборка состоит из 1000 элементов и 200 признаков. Дисбаланс классов отсутствует (доля класса 1: 0.486). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p - value \geq 0.997$ говорит о том, что выбросов нет (это было так же проверено построением гистограмм распределения). Статистика $W \geq 0.991$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

6.2 Преобразования над выборкой

Так как признаки распределены нормально для каждого класса, то следует посчитать их матожидание и дисперсию, а затем найти меру различия распределений. В качестве такой меры берём дивергенцию Кульбака-Лейблера.

$$KL = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 + \mu_1)^2}{\sigma_1^2} - 1$$

Отбирается n признаков с наибольшей дивергенцией. n находится подбором в процессе модели при разных значениях n . Итоговый результат для LogisticRegression и GaussianNB: $n = 15$.

Несмотря на то, что много признаков, имеющих отличное распределение для разных классов, качество метрик на них не очень хорошее ввиду того, что признаки коррелируют между собой. Не очень сильно, но существенно.

6.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB. Подсчитываются метрики. Модели имеют практически одинаковые результаты. Возьмём GaussianNB, так как он немного лучше.

6.4 Результаты

Таблица 6: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.774 | 293 | -4140 | 0 |
| LogisticRegression | 0.768 | 300 | -4141 | 18 |

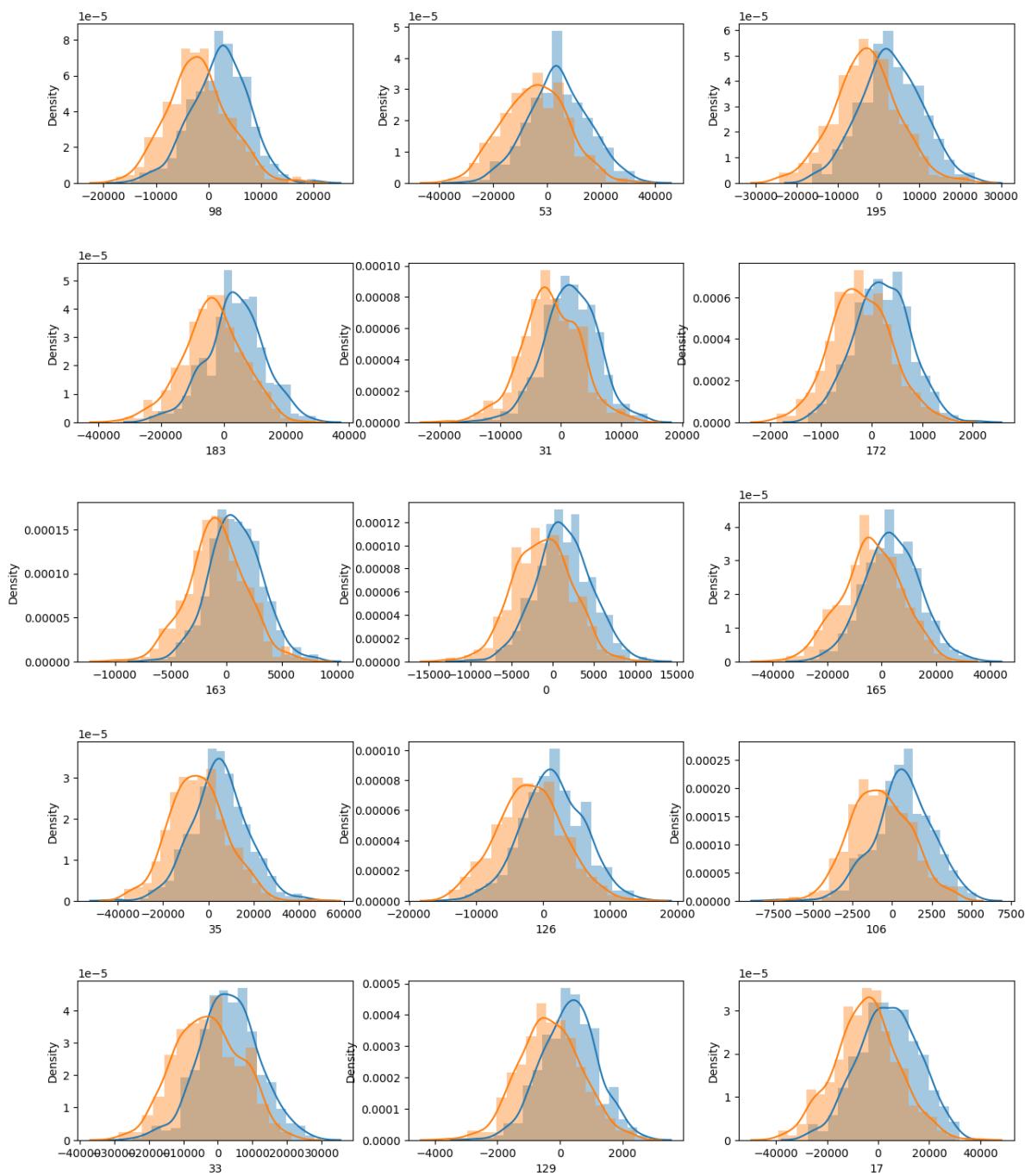


Рис. 12: Распределение 15 признаков с наибольшим значением KL для выборки №6

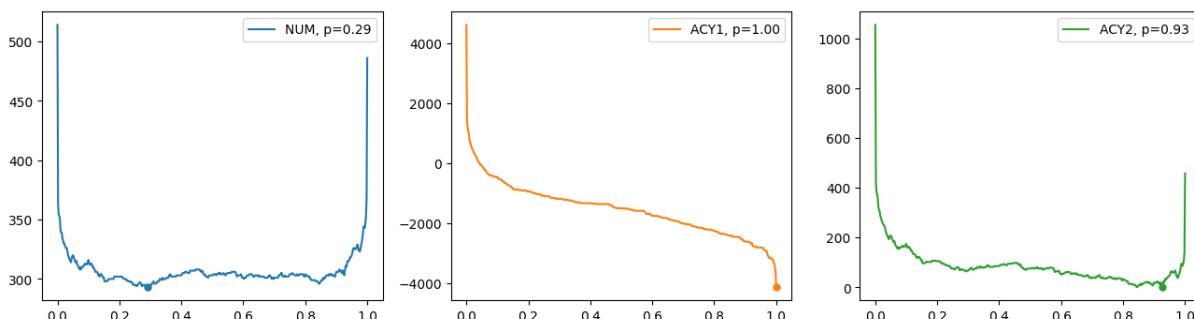


Рис. 13: Пороги отнесения к классу 1 для модели GaussianNB для выборки №6

7 Выборка №7

7.1 Описание выборки

Выборка состоит из 1000 элементов и 50 признаков. Дисбаланс классов присутствует (доля класса 1: 0.132). В выборке присутствуют выбросы.

Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака после замены выбросов на среднее значение признака. Для её проверки использовался критерий Шапиро-Уилка. $p-value \geq 0.997$, статистика $W \geq 0.991$. Такие значения говорят о том, что скорее всего признаки и правда распределены нормально для каждого класса.

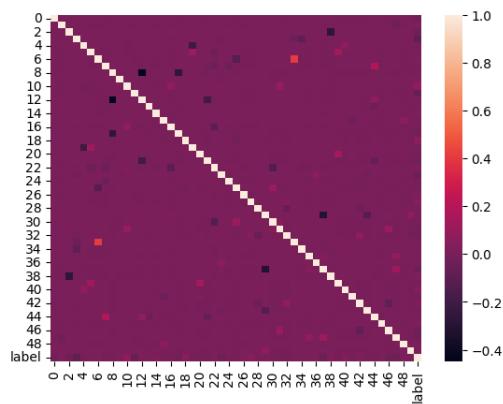


Рис. 14: Матрица корреляции для неочищенных данных для выборки №7

7.2 Преобразования над выборкой

Сначала следует избавиться от выбросов. С помощью методов `.quantile(0.05)` и `.quantile(0.95)` убираем значения, которые не попадают в 90% интервал значений признака для каждого класса. Затем к оставшимся элементам применяем критерий Шапиро-Уилка, чтобы убедиться, что они распределены нормально, и из начального массива выкидываем все элементы, что не попадают в 5σ интервал. Вместо выбросов вписываем средне по вписывающимся в интервал элементам.

Заметим, что после удаления выбросов, признаки перестают коррелировать между собой.

Так как после удаления выбросов признаки распределены нормально для каждого класса, то следует посчитать их матожидание и дисперсию, а затем найти меру различия распределений. В качестве такой меры берём дивергенцию Кульбака-Лейблера.

$$KL = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 + \mu_1)^2}{\sigma_1^2} - 1$$

Отбирается n признаков с наибольшей дивергенцией. n находится подбором в процессе модели при разных значениях n . Итоговый результат для LogisticRegression

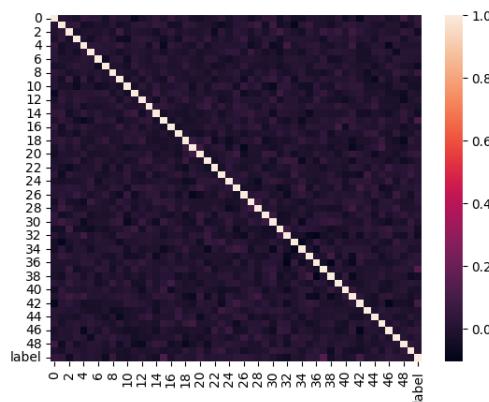


Рис. 15: Матрица корреляции для очищенных данных для выборки №7

$n = 9$ и GaussianNB $n = 10$.

7.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB. Подсчитываются метрики. Обе модели решили, что стоит предсказывать класс 0 почти для каждого элемента. Возьмём GaussianNB, так как он немного лучше.

7.4 Результаты

Таблица 7: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.651 | 132 | -7681 | -607 |
| LogisticRegression | 0.640 | 131 | -7681 | -607 |

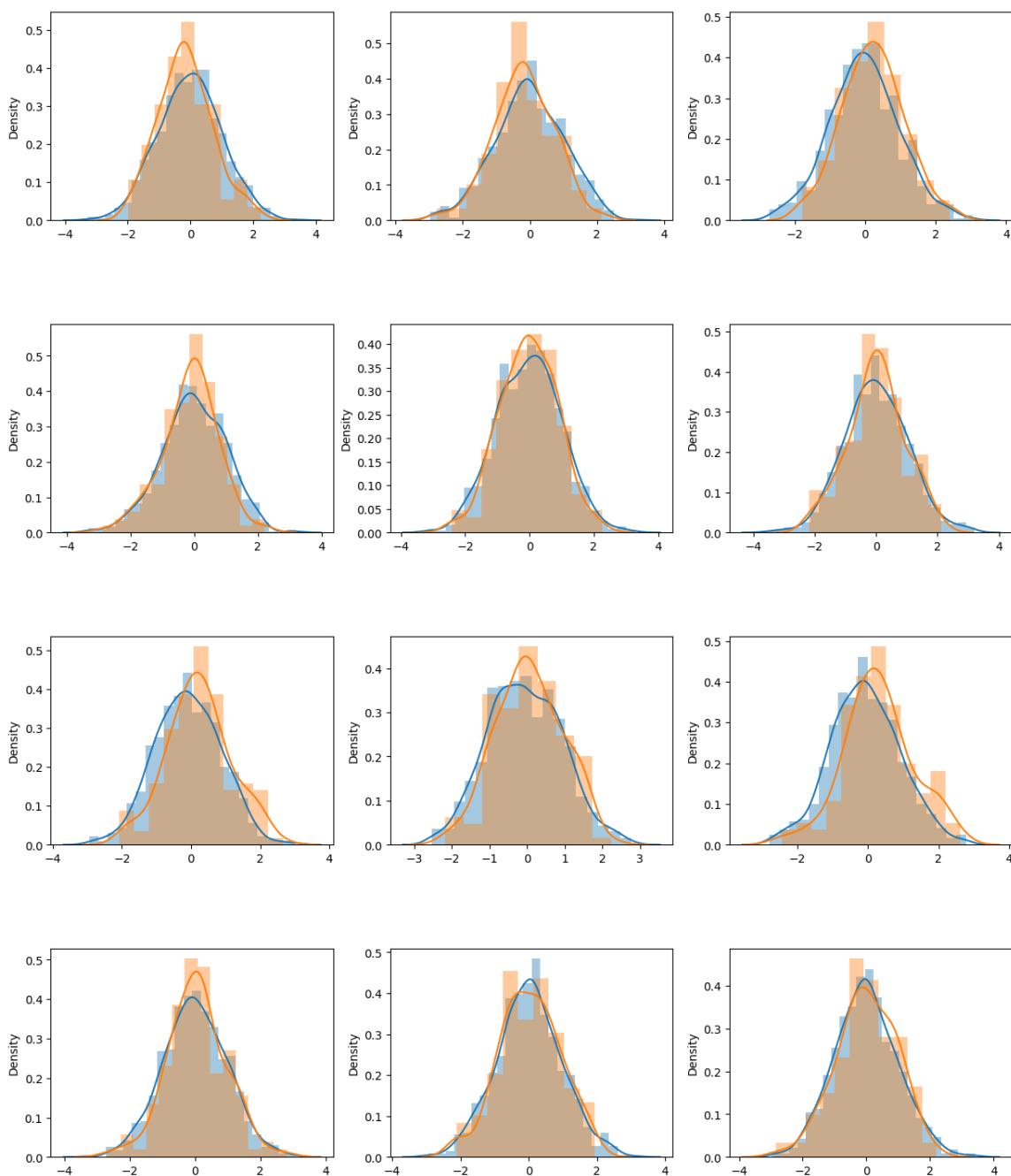


Рис. 16: Распределение 12 признаков с наибольшим значением KL для выборки №7

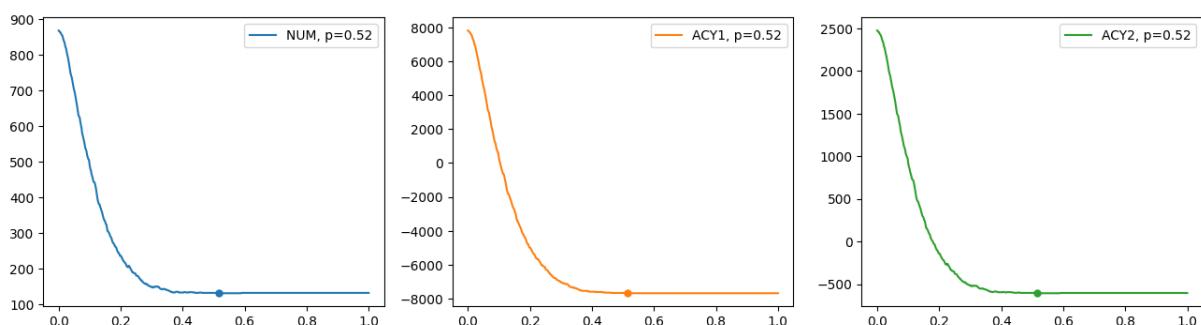


Рис. 17: Пороги отнесения к классу 1 для модели GaussianNB для выборки №7

8 Выборка №8

8.1 Описание выборки

В выборке 1000 объектов и 1500 фичей. В выборке содержатся выбросы, причём у каждого объекта есть хотя один выброс по какому-то признаку, так что полностью чистых объектов нет. Т.о. можно работать с каждым признаком по отдельности.

8.2 Преобразования над выборкой

1. Первый вариант - сделать РСА над исходной выборкой, отобрать 100 главных компонент, преобразовать выборку, попробовать обучить различные модели.

8.3 Модель

1. : Были проверены следующие модели: лог. регрессия с l1, l2 регуляризаторами (гипотеза линейной разделимости с априорным распределением на параметры модели), SVM-классификатор с rbf и poly ядрами (гипотеза линейности модели в новых признаковых пространствах, соответствующим выбранным ядрам), наивный байес над всеми фичами (они уже преобразованы), бустинг, гипотеза мусорных объектов (метка класса не зависит от объекта и генерируется случайным образом)

8.4 Результаты

1. Лучший результат по всем характеристикам имеет SVM-rbf

| NUM | ASY_1 | ASY_2 | AUC |
|-----|---------|---------|--------|
| 68 | -6980 | -394 | 0.8316 |

Таблица 8: Метрики

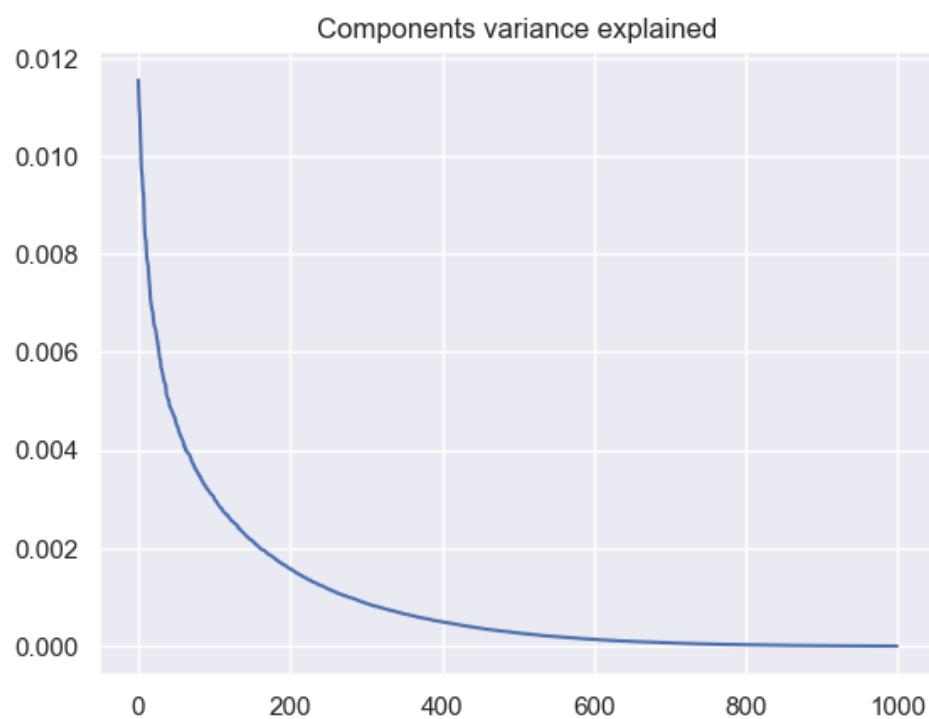


Рис. 18: PCA

9 Выборка №9

9.1 Описание выборки

В выборке 1000 объектов и 1500 фичей. В выборке содержатся выбросы, причём у каждого объекта есть хотя один выброс по какому-то признаку, так что полностью чистых объектов нет. Т.о. можно работать с каждым признаком по отдельности.

9.2 Преобразования над выборкой

1. Первый вариант - сделать РСА над исходной выборкой, отобрать 100 главных компонент, преобразовать выборку, попробовать обучить различные модели.

9.3 Модель

1. : Были проверены следующие модели: лог. регрессия с l1, l2 регуляризаторами (гипотеза линейной разделимости с априорным распределением на параметры модели), SVM-классификатор с rbf и poly ядрами (гипотеза линейности модели в новых признаковых пространствах, соответствующим выбранным ядрам), наивный байес над всеми фичами (они уже преобразованы), бустинг, гипотеза мусорных объектов (метка класса не зависит от объекта и генерируется случайным образом)

9.4 Результаты

1. Лучший результат по NUM , ASY_1 , ASY_2 имела l1-логрегрессия, лучший AUC - l2-логрегрессия

| NUM | ASY_1 | ASY_2 | AUC |
|-------|---------|---------|--------|
| 154 | -7201 | -539 | 0.6527 |

Таблица 9: Метрики

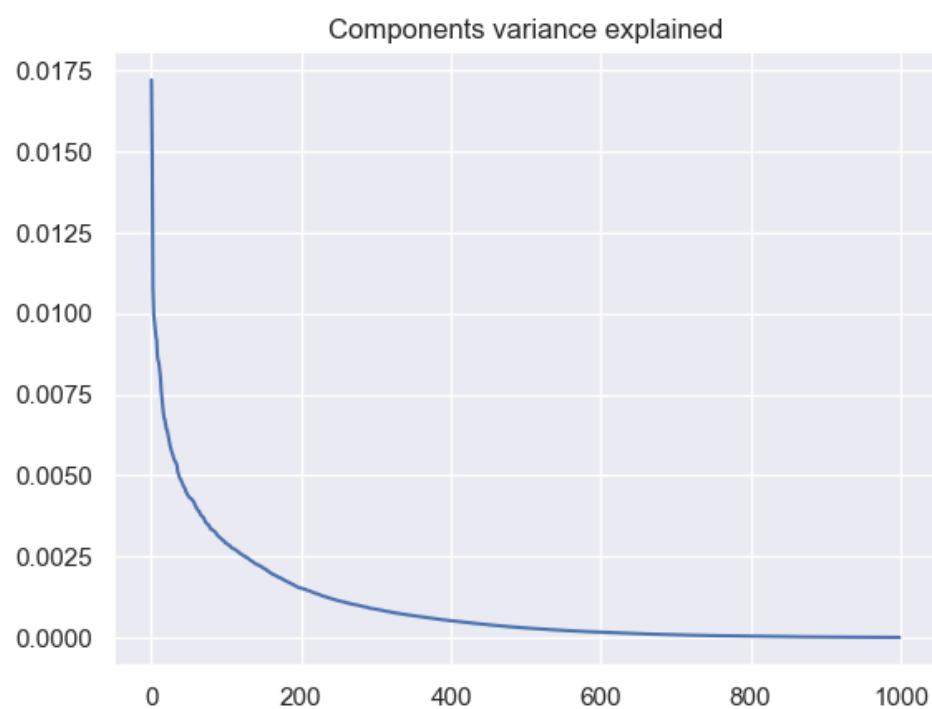


Рис. 19: PCA

10 Выборка №10

10.1 Описание выборки

Выборка состоит из 1000 элементов и 1500 признаков (много признаков, такая ситуация называется перепараметризация). Дисбаланс классов присутствует (доля класса 1: 0.39). Проверили распределение некоторого признака, по картинке очевидно присутствие выбросов.

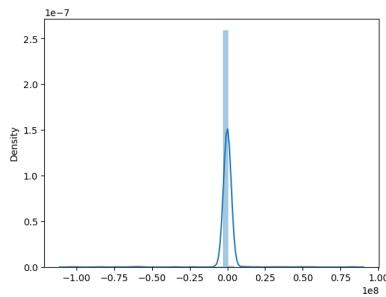


Рис. 20: распределение 0 признака

10.2 Преобразования над выборкой

Выбросы были очищены (те, что больше/меньше $0.05/0.95$ квантили были приравнены к $-3\sigma / 3\sigma$). После этого признаки были проверены на нормальность: они распределены нормально.

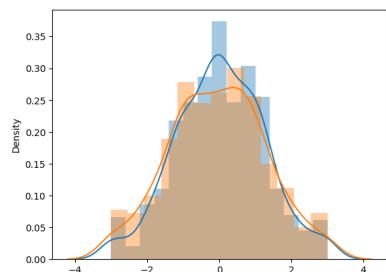


Рис. 21: распределение 1392 признака

После этого требовалось уменьшить размерность данных. Данные были проверены на корреляцию между собой, но сильной корреляции выявлено не было. Было проверено, есть ли признаки с большой KL-дивергенцией, но у всех дивергенция одного порядка. Затем был применен РСА, который дал результат:

Было решено, что это знак. Мы взяли 4 главные компоненты, спроектировали данные на них и предсказывали по получившимся 4 признакам.

10.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB.

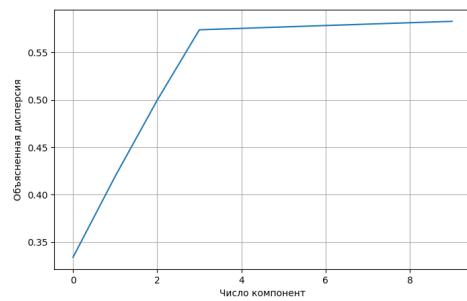


Рис. 22: Результат PCA (дисперсия, объясняемая признаками)

10.4 Результаты

Для каждой метрики (кроме AUC) была подобрана оптимальная граница первого класса.

11 Выборка №11

11.1 Описание выборки

Выборка состоит из 1000 элементов и 20 признаков. Дисбаланс классов присутствует (доля класса 1: 0.219). Была выдвинута гипотеза о нормальности распределений для каждого класса внутри признака. Для её проверки использовался критерий Шапиро-Уилка. $p-value \geq 0.995$ говорит о том, что признаки распределены нормально (это было так же проверено построением гистограмм распределения).

11.2 Преобразования над выборкой

Анализ матрицы корреляции показал, что признаки не сильно коррелируют между собой, то есть не являются линейно зависимыми

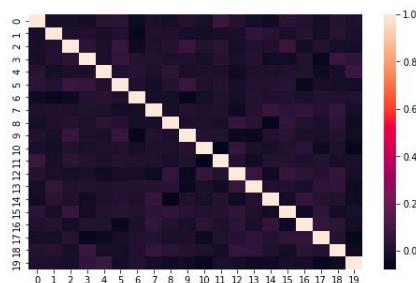


Рис. 23: Корреляция 20 признаков

Так как признаки распределены нормально для каждого класса, то следует посчитать их матожидание и дисперсию, а затем найти меру различия распределений. В качестве такой меры берём дивергенцию Кульбака-Лейблера.

$$KL = \log \frac{\sigma_1}{\sigma_0} + \frac{\sigma_0^2 + (\mu_0 + \mu_1)^2}{\sigma_1^2} - 1$$

Было проведено отсечение выбросов (на всякий случай) данная выборка достаточно стабильна, но некоторые выбросы мешали проведению KL-анализа. После этого были посчитаны значения KL-дивергенции для признаков, после отсечения выбросов. Минимальное значение 0.05, а максимальное 0.176. То есть значения девергенции отличаются менее чем в 3 раза, что не очень много, поэтому решено не производить отбор признаков, все признаки в некоторой степени влияют на ответ.

11.3 Модель

Проводится кросс-валидация для моделей: LogisticRegression, GaussianNB, CatBoost. Подсчитываются метрики. Логистическая регрессия предсказывала всегда 0 класс, что говорит, что она не очень умная, так что было решено использовать дальше catBoost и GNB.

11.4 Результаты

Для каждой метрики (кроме AUC) была подобрана оптимальная граница первого класса.

Таблица 10: Метрики на кросс-валидации

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|------------|-------------|-------|--------|--------|
| GaussianNB | 0.904 | 114 | -6821 | -379 |

для CatBoost статистика не записана, скорее всего, он плохо себя покажет на инференсе, так как он слишком сильно переобучается, но попытка не пытка.

12 Выборка №12

12.1 Описание выборки

В выборке 1000 объектов и 100 фичей. Выбросы отсутствуют, признаки не коррелируют между собой. Т.о. никаких преобразований выборки не проводилось.

12.2 Модель

Лучшей моделью оказалась SVM с экспоненциальным ядром, провалидированная на CV-5 и с подобранным параметром l2-регуляризации

12.3 Результаты

Привожу средние метрики на кросс-валидации

| NUM | ASY_1 | ASY_2 | AUC |
|-----|---------|---------|-----|
| 0 | -2430 | 460 | 1.0 |

Таблица 11: Метрики

13 Выборка №13

13.1 Описание выборки

Выборка 1000 элементов и 10 признаков, дисбаланс классов присутствует доля элементов единичного класса всего 0.18. Потом проверили гипотезу о нормальном распределении признаков, для этого воспользовался критерием Шапиро-Уилка, по нему получили, что $p_{value} \geq 0.998$, т.е. выбросы в выборке отсутвует, также это было визуально проверено, статистика же Шапиро-Уилка ≥ 0.995 , т.е. гипотеза о нормальном распределении признаков не отклоняется.

13.2 Преобразования над выборкой

Построили матрицу корреляции, видно, что есть группы признаков одинаково распределенных признаков, которые отличаются лишь порядком значений, было принято значение выбрать по одному признаку из таких групп, таким образом осталось 6 признаков. Затем было подсчитана дивергенция Кульбака-Лейблера между распределениями классов по каждой признаку, затем признаки были отсортированы по этому расстоянию между распределениями.

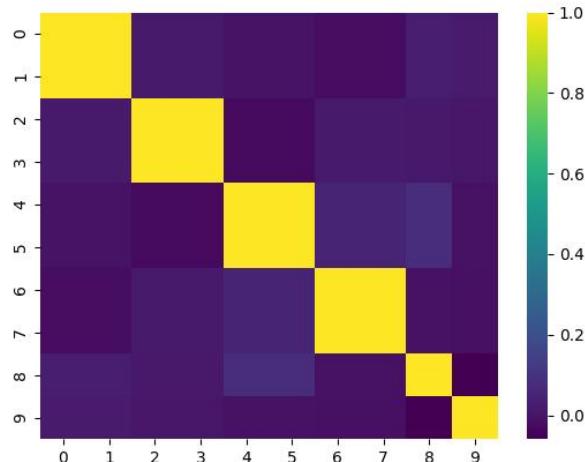


Рис. 24: Матрица корреляции для выборки №13

13.3 Модель

В качестве модели были выбран Logregression и GaussianNB из sklearn, в 10 фолдов в cross validation обучалась модель, метрика AUC усреднялась и вот по этому критерию мы измеряли качество модели, также перебиралось количество признаков, которое подается в саму модель, оказалось оптимальным число 1 для логрегресии и 2 для GaussianNb. У логрегресии получилось 0.527, у Gaussian 0.517. Также в качестве модели были попробована черная коробка от yandex, catboost classifier.

В качестве оптимизируемой метрик был выбран $ROC - AUC$, перебиралось количество итераций и какие фичи подавать. Оптимальным параметром стало 2 фичи и 5 итераций, получилось лучшее значение 0.532, но также, как и в четвертой выборке этому не стоит доверять, т.к. модель крайне нестабильна, но попробовать её один раз отправить стоит. При переборе трешхолда разделения классов опять получилось, как и в 4-ой выборке, что лучшим параметром является 1, это связано с дизбалансом классов.

13.4 Результаты

Таблица 12: Метрики ($ROC - AUC$ на кроссвалидации, NUM , $ASY1$, $ASY2$ на всей обучающей выборке

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|--------|--------|
| GaussianNB | 0.517 | 180 | 180 | -460 |
| LogisticRegression | 0.527 | 180 | 180 | -460 |
| CatboostClassifier | 0.532 | 180 | 180 | -460 |

14 Выборка №14

14.1 Описание выборки

Выборка 300000 элементов и 100 признаков, дисбаланс классов присутствует для элементов единичного класса всего 0.20839. Потом проверили гипотезу о нормальном распределении признаков, для этого воспользовался критерием Шапиро-Уилка, по нему получили, что $p_{value} \geq 1.0$ [ну вот так :)], т.е. выбросы в выборке отсутствуют, также это было визуально проверено, статистика же Шапиро-Уилка ≥ 0.995 , т.е. гипотеза о нормальном распределении признаков не отклоняется.

14.2 Преобразования над выборкой

Построили матрицу корреляции, видно, что признаки не коррелируют между собой и лишь четыре из них коррелирует с лейблом класса. Также как и в предыдущих выборках было измерено расстояние Кульбака-Лейблера и лишь у четырех признаков оно отлично от нуля, что согласуется с 4-мя признаками, которые коррелируют с ответом.

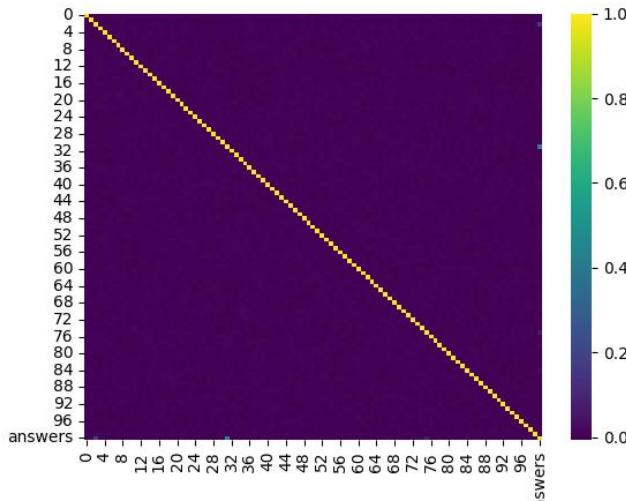


Рис. 25: Матрица корреляции для выборки №14

14.3 Модель

В качестве модели были выбран Logregression и GaussianNB из sklearn, в 10 фолдов в cross validation обучалась модель, метрика AUC усреднялась и вот по этому критерию мы измеряли качество модели, также перебиралось количество признаков. для GaussianNb. У логрегрессии получилось 0.847, у Gaussian 0.846. Перебирались количество фичей, которые имеют влияние на выборки, качество значительно растет при увеличении до 4-х фичей, потом тоже немного начинает

улучшаться, но незначительно. Это связано с тем, что мы добавляем шум и пытаемся этот шум выучить, поэтому оставим 4 признака. Также в качестве модели были попробована черная коробка от yandex, catboost classifier, но при всем переборе параметров он не дал $ROC - AUC$ выше 0.845, поэтому в результаты он не был включен.

14.4 Результаты

Таблица 13: Метрики ($ROC - AUC$ на кроссвалидации, NUM , $ASY1$, $ASY2$ на всей обучающей выборке

| Модель | $ROC - AUC$ | NUM | $ASY1$ | $ASY2$ |
|--------------------|-------------|-------|----------|---------|
| GaussianNB | 0.846 | 48927 | -2075113 | -144261 |
| LogisticRegression | 0.847 | 48769 | -2075126 | -144339 |

15 Выборка 1, 2, 4, 7, 8 и 9

15.1 Описание выборки

Описание выборок есть ранее.

15.2 Преобразования над выборкой

Были сделано следующее предположение о том, что в каждой из этих выборок есть выбросы от них нужно избавиться, заменив на среднее минус 3 сигмы, если оно слева, или среднее плюс 3 сигмы, если оно находится справа. Среднее и дисперсию мы считали, не учитывая 5-й квантиль, слева и справа. Такая итерация была одна, если делать больше, то было хуже. Далее мы отсортировали признаки по значимости, используя дивергенцию кульбака-лейблера. И сделали из всех признаков один по следующей формуле,

$$\text{feature} = \frac{1}{l} \sum_{i=1}^l \text{sign}(\text{mean}_1 - \text{mean}_0) * \sin(KL_i)$$

Где l – это такие фичи, у которых p_{value} по Шапиро был больше 0.97.

Почему именно синус я не знаю, я перебирал, он лучше, чем косинус, экспоненты, полиномы и линейные версии и ещё какие-то комбинации. В качестве модели был выбран GaussianNB из sklearn.naive_bayes, которая предсказывала собственно по этой одной новой фиче.

Трещхолды были выбраны перебором, как и в других выборках.