

Анализ многомерных данных

Первоначальный многомерный анализ

Для выявления степени линейной зависимости между переменными была построена корреляционная матрица (*Рисунок 1*), исходя из ее значений (не более 0.2 по модулю) можно сделать вывод об отсутствии указанной связи

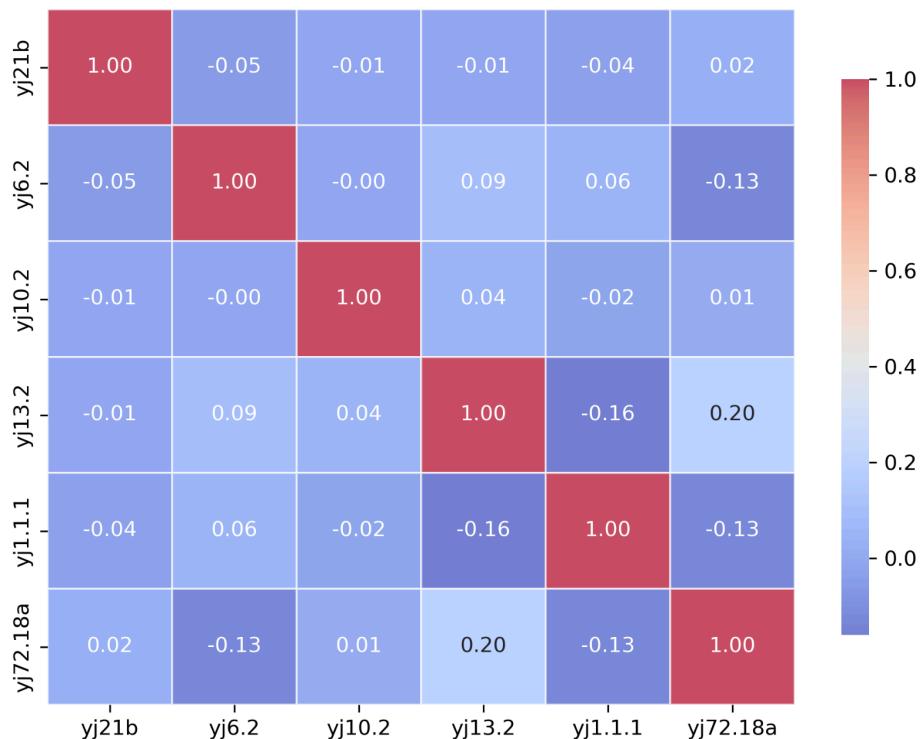


Рис. 1: Корреляционная матрица

Несмотря на отсутствие линейной связи по результатам корреляционного анализа, была дополнительно проведена проверка на наличие мультиколлинеарности, наличие которой могло бы сделать модель более чувствительной к изменениям выборки. Результаты проверки представлены ниже в Таблице 1.

Исходя из значений показателя VIF (коэффициент вариации инфляции), можно сделать вывод о вероятном отсутствии мультиколлинеарности между переменными, поскольку все они имеют значения близкие к 1, что свидетельствует о том, что они отражают уникальную часть дисперсии зависимой переменной. Значение константы можно не принимать во внимание, так как она не связана с другими переменными и была включена лишь для учета свободного члена в модели

Таблица 1: Значения показателя VIF

Переменная	Значение VIF
Константа	83.36
yj21b	1
yj6.2	1,03
yj10.2	1
yj13.2	1,06
yj72.18a	1,07

Для выявления нелинейных связей были построены диаграммы рассеяния (*Рисунок 2*) для каждой пары переменных. Визуальный анализ диаграмм показывает отсутствие каких-либо выраженных нелинейных закономерностей в виду распределения значений случайным образом

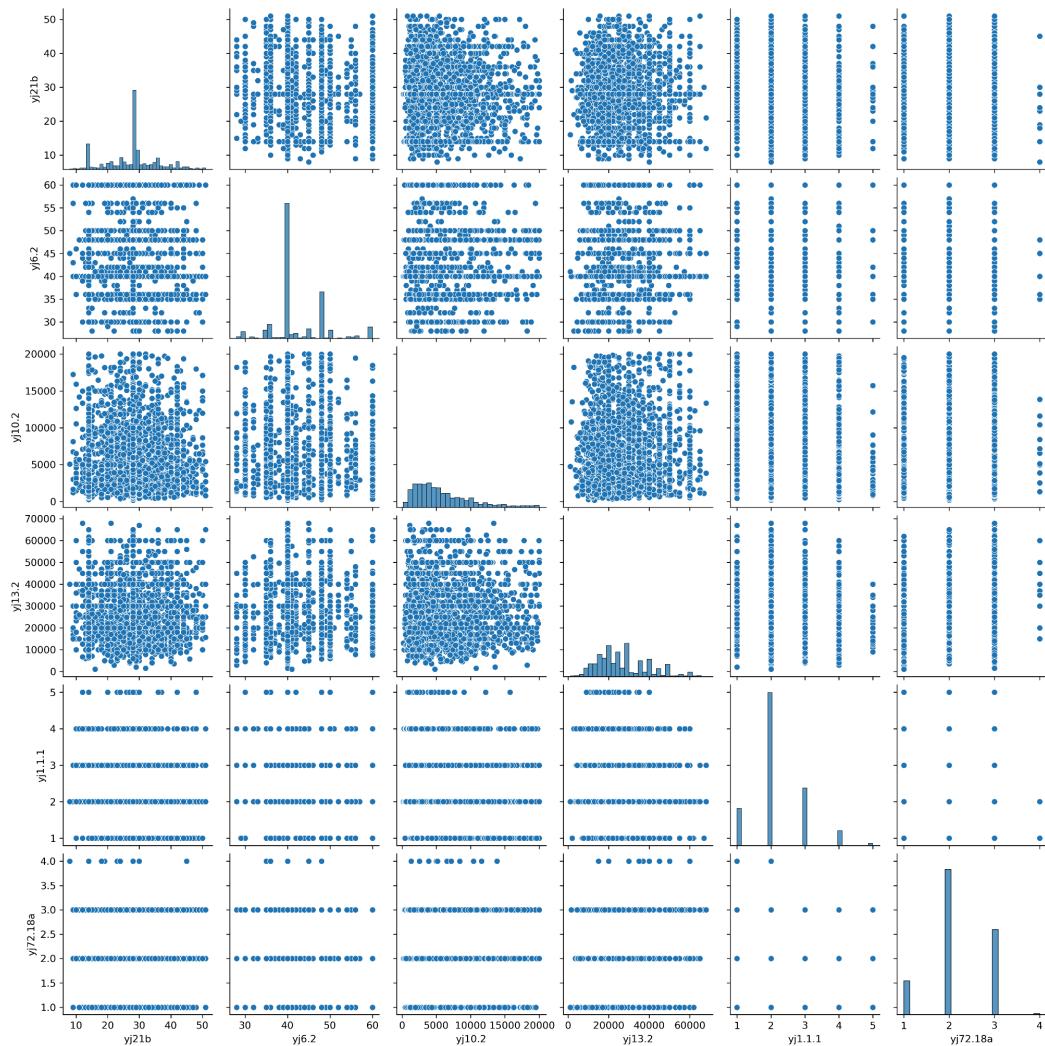


Рис. 2: Диаграммы рассеяния

Продвинутые методы многомерного анализа

Многомерные датасеты сложно интерпретировать наглядно.

Используемые в данном разделе продвинутые методы многомерного анализа могут помочь выявить скрытые закономерности между признаками, а также представить их в более наглядной форме

Метод главных компонент

Представленное на рисунке 3 распределение классов после применения метода главных компонент (PCA) не позволяет полноценно выделить отдельные группы

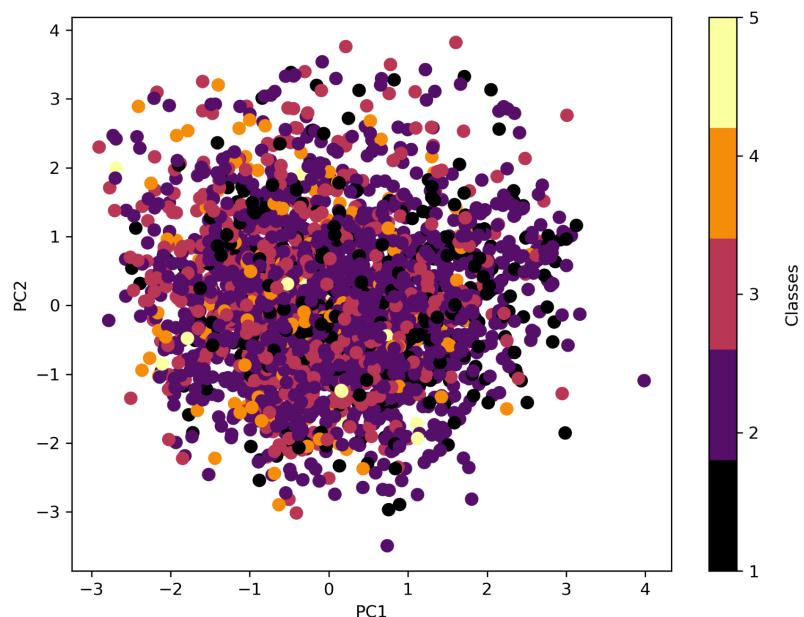


Рис. 3: PCA анализ

t-SNE

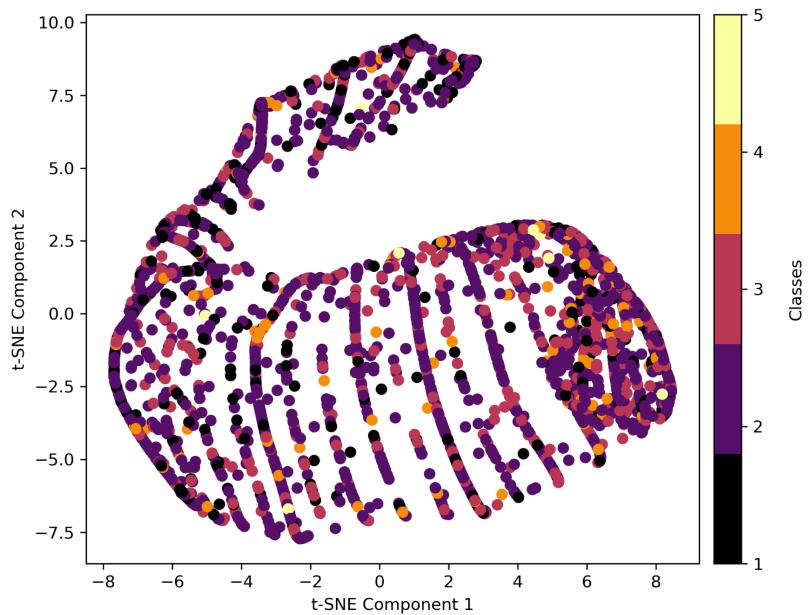


Рис. 4: t-SNE анализ

Подбор параметров в t-SNE помог выделить некоторые кластеры, отражающие распределение сэмплированных данных (Рисунок 4)

UMAP

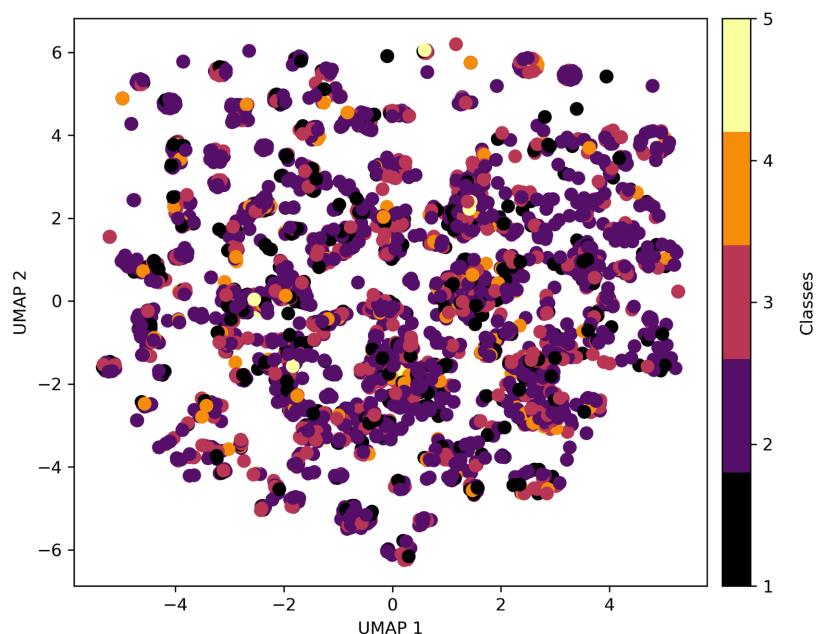


Рис. 5: UMAP анализ

UMAP сохранил больше локальных зависимостей и отразил области содержащие все классы одновременно, что может говорить о схожих характеристиках в признаках (рисунок 5)

Факторный анализ

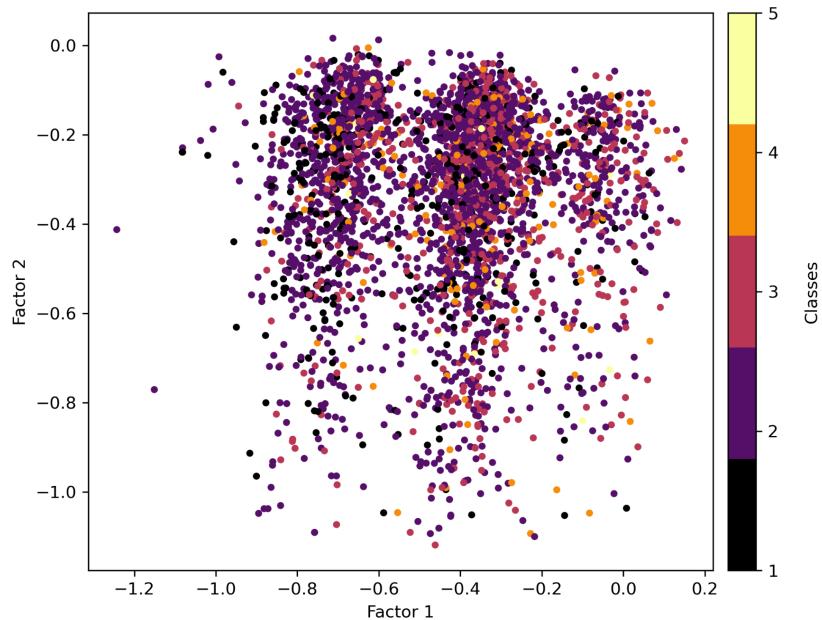


Рис. 6: Факторный анализ

Факторный анализ выделил основные зависимости между переменными. Однако найденные классы в результирующем пространстве перемешаны и отражают значительных зависимостей (Рисунок 6)

Кластерный анализ

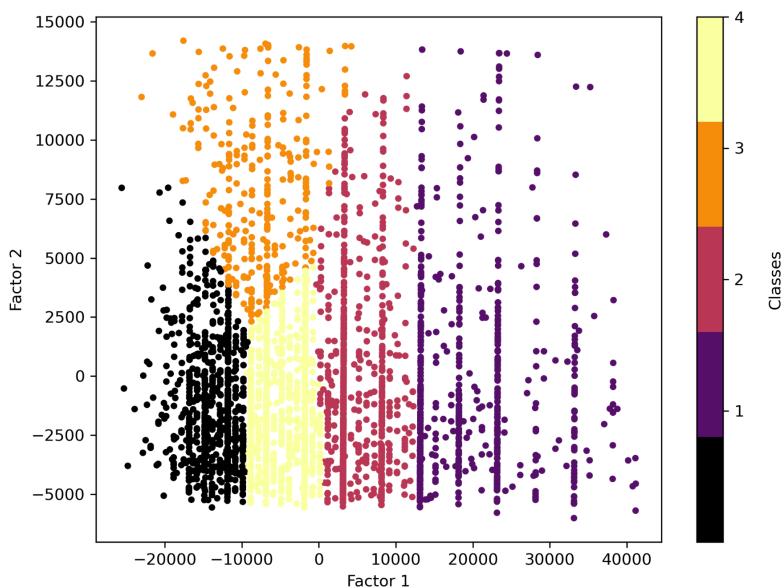


Рис. 7: Кластерный анализ

Кластерный анализ показал, что данные можно разделить на несколько явно заметных групп в новом двумерном пространстве (*Рисунок 7*)

Дополнительные методы создания признаков

Были сгенерированы новые признаки - полиномиальные (со степенью 3) комбинации исходных. Все непрерывные признаки были преобразованы в категориальные. Корреляционная матрица получившихся признаков показана на Рисунке 8

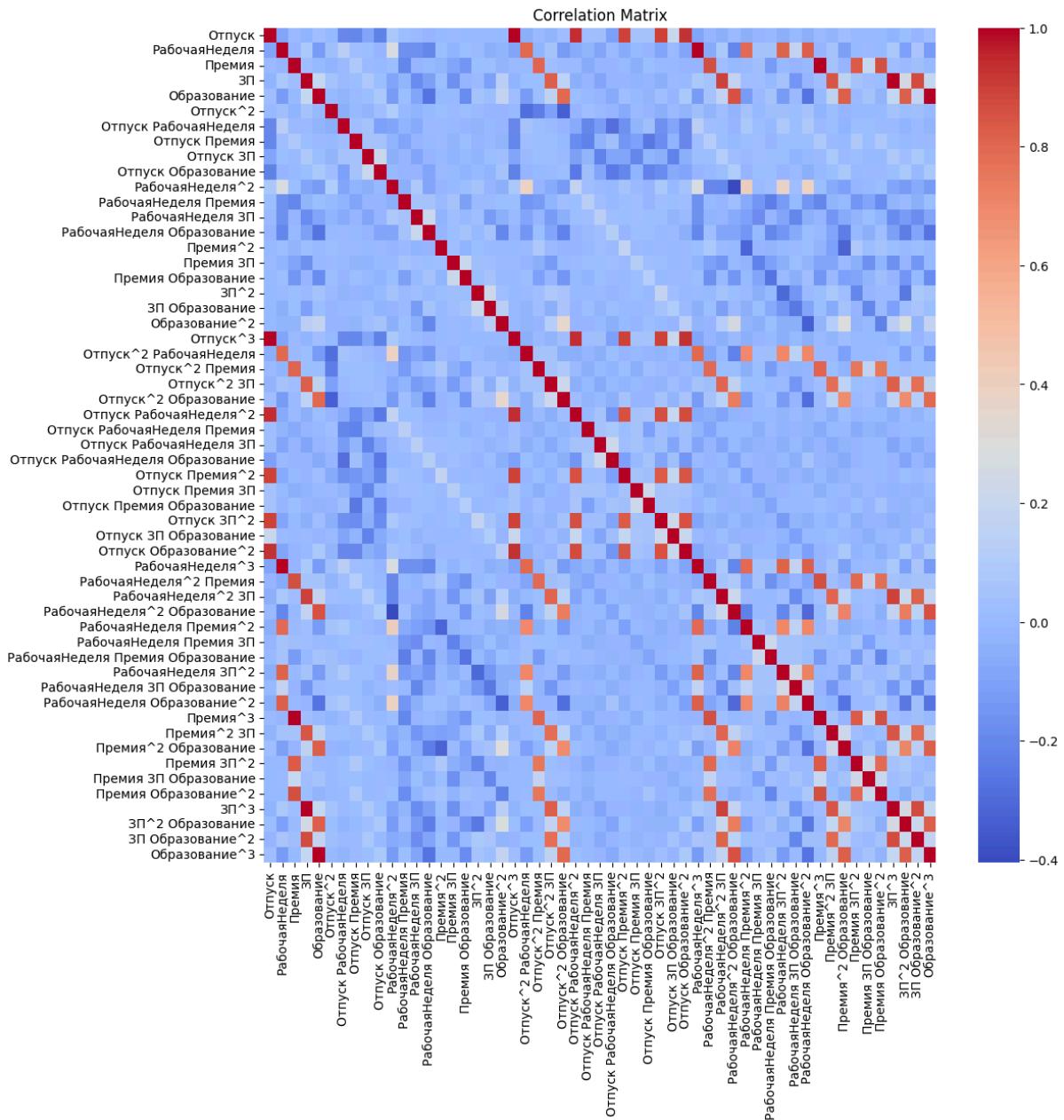


Рис. 8: Корреляционная матрица добавленных полиномиальных признаков

Далее был составлен новый датасет после отбора признаков с помощью регуляризационного метода Lasso.

Байесовская сеть

В ходе выполнения лабораторной работы была обучена структура байесовской сети на сэмплированных данных, т.к. они показали наилучший результат при обучении модели

В качестве библиотеки для обучения использовалась ВАМТ. У нас смешанные данные, поэтому для наилучшего обучения непрерывные данные были дискретизированы.

В качестве алгоритма нахождения структуры использовался Hill-Climbing. Для представления условного распределения непрерывных переменных использовались смеси Гауссовых распределений, чтобы в дальнейшем при сэмплировании синтетические данные лучше отображали изначальную выборку.

В результате обучения структуры получился DAG, представленный на Рисунке 9

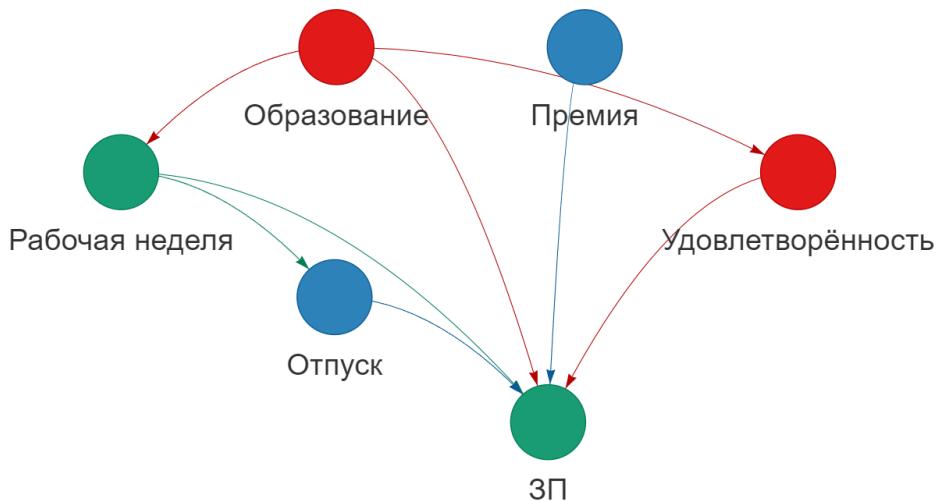


Рис.9: Структура Байесовской сети

По данной структуре можно заметить, что удовлетворённость, которую мы предсказываем, условно независима от заработной платы и длины рабочей недели. Также следует заметить, что удовлетворённость независима от премии и отпуска, но только до тех пор, пока неизвестна заработка. Прямую зависимость от образования можно объяснить тем, что обычно чем выше образование, тем более престижная работа, но также следует отметить, что данный график может не полностью отражать действительность, т.к. датасет очень несбалансирован, что видно в Таблице 2

Табл.2: Несбалансированность датасета

Удовлетворенность	Кол-во значений
1	475
2	1966
3	742
4	193
5	26

Поэтому были пересэмплированы синтетические данные.

Сэмплирование происходило следующим образом: на сабсете, в котором содержались только строки с определённым значением удовлетворенности, которое требует синтетических данных, обучалась структура байесовской сети, а затем по этому же сабсету были уточнены параметры распределений. Далее происходило семплирование до нужного количества данных и эти данные добавлялись в основной датасет.

У удовлетворенности “5.0” данных слишком мало, поэтому семплирование могло привести к некачественным данным.

Для валидации сэмплирования был построен qq-plot, который говорит о том, что синтетические данные неидеально описывают исходные данные, но при этом всё равно достаточно хорошие. График представлен ниже на Рисунке 10

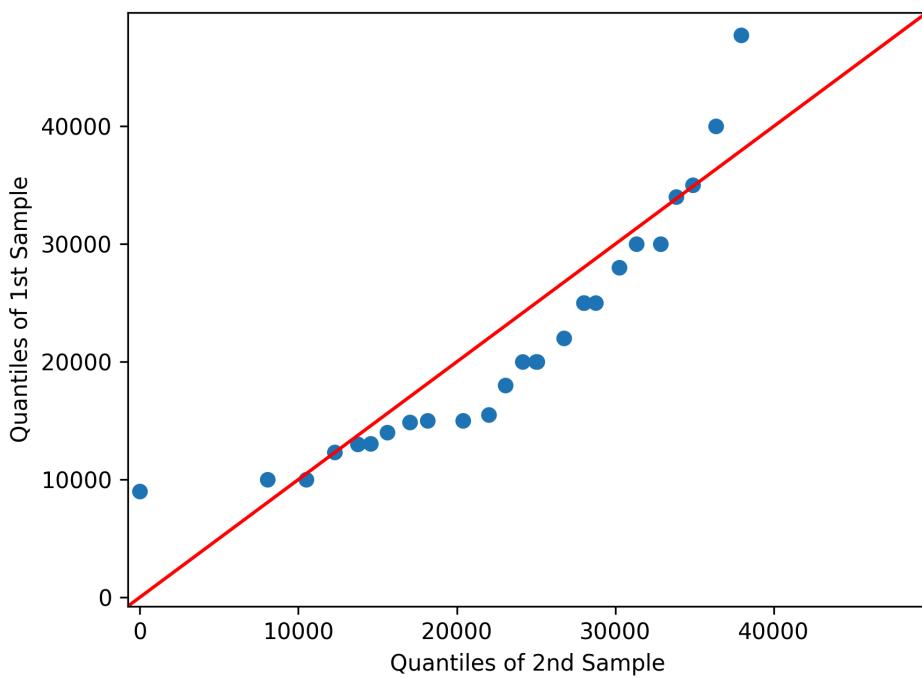


Рис.10: qq-plot для заработной платы

Повторное построение модели

Выделение дополнительных признаков

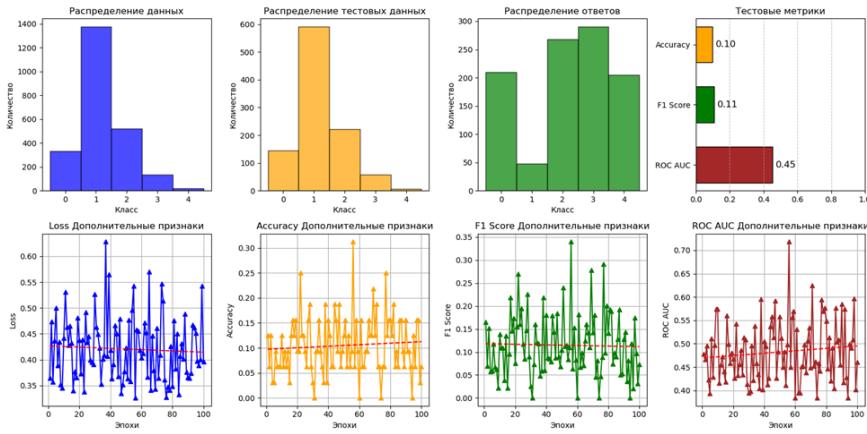


Рис. 11: Графики метрик модели на датасете с дополнительными признаками

Модели, основанные на выделении многомерных признаков, отразили новые зависимости между признаками и повысили гибкость модели при обучении (Рисунок 11). Однако из за низкой предсказательной способности самой модели это не повысило её метрики.

Байесовская сеть

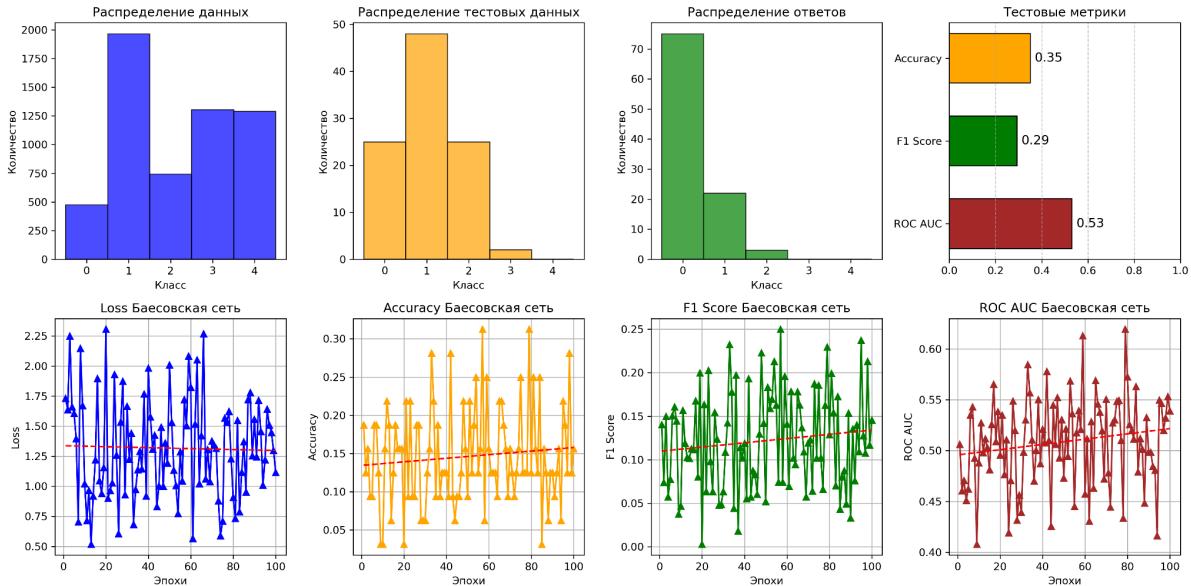


Рис. 12: Графики метрик модели на датасете с байесовской сетью

Байесовская сеть позволила скорректировать распределение редких классов в датасете (Рисунок 12). Это позволило лучше отразить каждый класс

при обучении, что положительно отразилось на итоговых метриках модели, на уровне с сэмплированными данными

Заключение

По итогу применения методов многомерного анализа можно заключить, что итоговые результаты не показали значительного улучшения метрик.

Наилучшим образом показала себя байесовская сеть, однако синтетические данные оказались не лучшего качества, ввиду дисбаланса классов в изначальном датасете, к тому же некоторые признаки были определены недостаточным количеством наблюдений. Принцип создания синтетических данных аналогичен сэмплированию, которое при одномерном анализе показало наивысшие показатели метрик.

Дополнительное создание признаков и их последующий отбор изменил структуру признаков датасета, что позволило добавить гибкости модели, однако это по итогу привело к ухудшению целевых результатов обучения модели.