**Федеральное государственное образовательное бюджетное учреждение высшего профессионального образования**

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ**

**ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**(Финансовый университет)**

**Факультет «Факультет информационных технологий и анализа больших данных»**

**Курсовая работа**

по дисциплине: «Технологии анализа данных и машинного обучения»

на тему:

**«Предварительный анализ данных и построение признаков в задачах оценки активов»**

Вид исследуемых данных:

Набор данных «Прогноз банкротства компаний»

Выполнила:

студентка группы ПМ19-4

Качуляк Маргарита Григорьевна

Научный руководитель:

к.т.н., доцент

Болтачев Эльдар Филаридович

Москва 2022г.

Оглавление

[1. ВВЕДЕНИЕ 3](#_Toc122385507)

[2. ТЕОРЕТИЧЕСКАЯ СПРАВКА 5](#_Toc122385508)

[2.1 Активы компании и пути ее оценки 5](#_Toc122385509)

[2.2 Технические методы предобработки и обогащения признаков для построения модели машинного обучения 7](#_Toc122385510)

[2.3 Оценка эффективности преобразования данных 11](#_Toc122385511)

[3. ПРАКТИЧЕСКАЯ ЧАСТЬ 13](#_Toc122385512)

[3.1 Анализ и предобработка выбранных данных 13](#_Toc122385513)

[3.2 Работа с признаками 21](#_Toc122385514)

[4. ЗАКЛЮЧЕНИЕ 24](#_Toc122385515)

[5. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ 25](#_Toc122385516)

[6. ПРИЛОЖЕНИЕ 26](#_Toc122385517)

[6.1 Характеристики компьютера 26](#_Toc122385518)

[6.2 Коды программ 26](#_Toc122385519)

[6.3 Список файлов 26](#_Toc122385520)

## 1. ВВЕДЕНИЕ

В представленной курсовой работе будет рассмотрен предварительный анализ данных и построение признаков в задачах оценки активов.

Предварительная обработка данных является неотъемлемым этапом машинного обучения, поскольку качество данных и полезная информация, которая может быть получена из них, напрямую влияют на способность нашей модели к обучению.

Предварительная обработка данных – это процесс подготовки необработанных данных и придания им пригодности для модели машинного обучения. Это первый и решающий шаг при создании модели машинного обучения. Это метод интеллектуального анализа данных, который включает в себя преобразование необработанных данных в понятный и организованный формат. Реальные данные обычно содержат шумы, пропущенные значения и, возможно, в непригодном формате, который нельзя напрямую использовать для моделей машинного обучения. Для решения таких проблем он подготавливает исходные данные для дальнейшей обработки. Предварительная обработка данных является необходимой задачей для очистки данных и придания им пригодности для модели машинного обучения, что также повышает точность и эффективность модели машинного обучения. Предварительная обработка данных позволяет:

* Делать базу данных более точной. Устраняются некорректные или отсутствующие значения, возникшие в результате человеческого фактора или сбоя.
* Повышать плотность данных. Когда в данных есть несоответствия или дубликаты, это влияет на точность результатов.
* Сглаживать данные. Таким образом, мы упрощаем его использование и интерпретацию.

Именно поэтому чрезвычайно важно не игнорировать и не пренебрегать данным процессом, чтобы в итоге обученная модель смогла делать наиболее точный прогноз.

Целю данной работы является предварительный анализ и подготовка данных, предназначенных для дальнейшего решения задачи оценки активов того или иного предприятия или компании. В качестве модельного датасета были выбраны данные о банкротстве ряда компаний из Тайваньского экономического журнала за 2009–2019 годы. Банкротство было определено на основе правил ведения бизнеса Тайваньской фондовой биржи.

Для достижения цели курсового проекта был определен следующий ряд задач:

* Найти данные для исследования, соответствующие выбранной тематике работы
* Рассмотреть теорию о специфике данных по теме исследования
* Изучить способы предварительной обработки данных для предстоящего моделирования
* Исследовать «сырые» данные
* Применить как можно больше известных методов предобработки на выбранном датасете
* Произвести проверку качества предобработки данных

Прежде чем приступать непосредственно к предобработке данных и построению признаков будет проведена подготовительная работа. Для начала будет рассмотрена теоретическая справка по теме проекта: информация о видах активов и их экономической значимости. Затем изучим техническую сторону: способы предварительного анализа данных в машинном обучении, применяемые для дальнейшей разработки модели и наиболее качественного прогнозирования. Далее перейдем непосредственно к работе с данными и применим известные методы предобработки и обогащения датасета, а также докажем, что после преобразований исходные данные не изменили свою полноту, а, наоборот, стали только богаче и удобнее в использовании.

Практическая часть работы будет выполнена при помощи языка программирования Python.

## 2. ТЕОРЕТИЧЕСКАЯ СПРАВКА

### 2.1 Активы компании и пути ее оценки

Термин «компания» (или «предприятие») можно понимать как отдельный экономический субъект или экономическую организацию, которая производит и продает за свой счет товары и услуги с целью максимизации своей прибыли. Финансовое управление современным предприятием — это максимизация его стоимости. Компанию можно назвать особой формой инвестирования, ведь ее владельцы инвестируют в свои собственные капитальные ресурсы и рассчитывают получить определенные выгоды в результате умножения вложенного таким образом капитала. Само собой, это приводит к увеличению стоимости предприятия. Экономическая сущность вопросов собственности вплотную связана с проблемами полезности и способами нахождения денежной оценки объекта собственности: специфика предприятия, его стоимость, различные методы и процедуры обучения, а также условия.

Процесс оценки бизнеса весьма сложный и требующий применения широких знаний во многих областях науки. При всем при том, что группа специалистов уже потратила больше количество усилий по стандартизации процесса оценки, все еще остается много открытых вопросов и непринятых cпорных решений. Методы оценки предприятий и их организованных частей не регламентированы законодательно как строго необходимые. Более того, не существует комплексного и единого набора правил, применимых к этому процессу. В первую очередь, отсутствие единообразных правил связано с тем, что очень трудно полностью нормализовать процесс, который может причисляться к организациям с отличными юридическими формами, спецификой, структурами или активами собственности. Но все же существуют стандарты, которые допускают его частичное структурирование. Поэтому во многих странах мира уже много лет существуют стандарты оценки бизнеса. Они были разработаны специальными организациями, в состав которых входят профессионалы по проведению оценок; касаются методологии оценки и диапазона знаний, которыми должен обладать оценщик. Эксперты связаны моделями поведения, которые гарантируют сопоставимость оценок и легкость их проверки и нормализованными процедурами. Это позволяет поддерживать безопасность деловых операций. Потребность оценки бизнеса продиктована глобализацией экономики. Одновременно с экономическим развитием, которое сопровождается интенсивным притоком капитала во все большее число стран, оценка становится необходимой для слияний и поглощений, приватизации, создания совместных предприятий или продажи и многих других процессов, связанных с предприятиями. Затруднено сформулировать определение конечной ценности объекта из-за субъективности самого понятия «ценность». Не стоит забывать, что фактическая рыночная стоимость предприятия редко когда определяется только учитываемыми в балансе активами. Фактическая оценка определяется рядом переменных факторов, таких как стратегия развития компании, привлекательность рынка и его характер, способ использования принадлежащих активов, экономическая ситуация в стране и человеческие ресурсы.

Из расчета того, что рынок подобен живому организму снова и снова появляется новая информация, влияющая на состояние и деятельность предприятий. Таким образом можно сделать вывод, что оценка бизнеса – это весьма субъективный процесс определения цены активов и выгод, приобретенных компанией в ходе их квалифицированного управления.

Именно поэтому в представленной работе для решения задачи оценки активов будут использованы данные, содержащие более 6000 наблюдений с 95 различными финансовыми параметрами для каждого и решение о значимости того-или иного показателя будет приметаться в течение работы, исследуя данные с технической точки зрения.

### 2.2 Технические методы предобработки и обогащения признаков для построения модели машинного обучения

Сейчас мир полон данных. Каждый день создаются миллиарды данных, и они хранятся в базах данных по надлежащему каналу, а также в надлежащем порядке. Они называются наборами данных. Набор данных – это не что иное, как набор выборок или наблюдений, организованных с помощью нескольких столбцов и отделенных друг от друга. Те столбцы, которые сообщают о свойствах и качествах любого объекта или конечного результата, известны как независимые столбцы или объекты. Зависимые столбцы – это конечный результат работы функций. Другими словами, объекты – это признаки набора данных, иначе называемые переменными и атрибутами.

Реальные данные часто являются неполными, непоследовательными, им не хватает определенного поведения или тенденций и, вероятно, содержат много ошибок. Чтобы лучше визуализировать и извлечь скрытый шаблон, необходимо выполнить предварительную обработку данных.

Рассмотрим ряд базовых наиболее эффективных приемов при работе с данными:

1. Кодирование нечисловые характеристик.

Алгоритмы ML работают только с числовыми значениями. Таким образом, существует необходимость полностью смоделировать проблему и ее данные в цифрах. Традиционный способ обработки категориальных столбцов — это их OneHot кодирование (бинарное). Также, для решения данной проблемы можно использовать кодировку символов. Она осуществляется с применением специальной библиотеки «category\_encoders», присваивая каждому элементу множества свой личный ID.

1. Уникальные имена. Обработка дубликатов.

Если повторяется одна и та же строка или столбец, вы также можете удалить их, сохранив первый экземпляр для того, чтобы при запуске алгоритмов машинного обучения не предоставлять этому конкретному объекту данных преимущество или предвзятость.

1. Поиск выбросов и удаление аномальных значений.

Выброс – это наблюдение, которое находится на аномальном расстоянии от других значений в случайной выборке из совокупности. В некотором смысле это определение оставляет за аналитиком право решать, что будет считаться ненормальным. Обработка выбросов является хорошей практикой перед передачей данных в модель ML, поскольку мы можем получить лучшую модель с хорошими показателями.

Чаще всего, для определения выбросов используется вычисление межквартального диапазона IQR. Выброс должен соответствовать следующим критериям:

или ,

где

1. Заполнение пропусков.

Обработка недостающих значений - одна из самых больших проблем, с которыми сталкиваются аналитики, поскольку принятие правильного решения о том, как с этим справиться, делает модели данных намного надежнее.

Основным способом замены недостающих значений является замена на среднее значение (в случае с числовыми данными) или медиану (для категориальных переменных). Вменение среднего/медианного основывается на предположении, что данные отсутствуют совершенно случайным образом. Мы решаем эту проблему, заменяя значение NaN на наиболее часто встречающиеся переменные. Также, иногда принимается решение заменять пропуски на произвольное значение. А в случае, если отсутствующие значения превышают 60% от общего объема данных в признаке/строке, то принимается решение полностью его удалить.

1. Сокращение высоко коррелирующих признаков.

Высококорреляционные функции являются распространенной проблемой в машинном обучении. Когда две черты сильно коррелируют, информация, которую они предоставляют, очень похожа, и, вероятно, избыточно включать обе.

Удаление одного из коррелированных объектов из набора функций решает проблему. В рамках нашего подхода мы сначала создаем корреляционную матрицу, которая включает все функции. Мы используем метод dataframe.corr() и смотрим на верхний треугольник корреляционной матрицы, чтобы найти пары сильно коррелированных объектов и исключить по одному объекту из каждой из этих пар.

1. Предотвращение дисбаланса данных.

Работа с несбалансированными наборами данных предполагает такие стратегии, как улучшение алгоритмов классификации или балансировка классов в обучающих данных перед предоставлением данных в качестве входных для алгоритма машинного обучения. Последний метод является предпочтительным, поскольку он имеет более широкое применение.

Основная цель балансирования классов состоит в том, чтобы либо увеличить частоту класса меньшинства, либо уменьшить частоту класса большинства. Это делается для того, чтобы получить примерно одинаковое количество экземпляров для обоих классов.

а) Удаление некоторых наблюдений из класса большинства

Удаление наблюдений делается до тех пор, пока не будет сбалансирован класс большинства и меньшинства. Такой способ может быть хорошим выбором, когда у вас есть тонна данных – например, миллионы строк. Но минусом недостаточной выборки является то, что мы можем удалить информацию, которая окажется ценной.

б) Добавление большего количества копий к классу меньшинства

Такой способ передискретизации заключается в дублировании событий меньшего класса. При недостаточной выборке следует учитывать, что это может привести к переобучению и плохому обобщению вашего тестового набора.

в) Генерация синтетической выборки (SMOTE)

Этот метод генерирует синтетические данные для класса меньшинств. SMOTE (метод синтетической передискретизации меньшинства) работает путем случайного выбора точки из класса меньшинства и вычисления k-ближайших соседей для этой точки. Синтетические точки добавляются между выбранной точкой и ее соседями.

1. Уменьшение размерности.

В случае, когда в исходном датасете имеется большое количество различных признаков мы можем столкнуться Одним из самых простых и наиболее широко используемых алгоритмов контролируемого обучения является анализ главных компонент (Рrincipal Сomponent Аnаlysis, РСА). PCA — это принципиально алгоритм уменьшения размерности, но он также может быть полезен в качестве инструмента для фильтрации шума, визуализации, проектирования и извлечения признаков.

Перед тем, как применить РСА, необходимо отмасштабировать данные таким образом, чтобы каждый признак имел единичную дисперсию, воспользовавшись StandardScaler, так как этот метод очень чувствителен к дисперсиям исходных Признаков. Если есть большие различия между диапазонами исходных переменных, те переменные с большими диапазонами будут преобладать над остальными (например, переменная в диапазоне от 0 до 100, будет доминировать над той, что лежит в диапазоне от 0 до 1), что приведет к необъективным итогам.

Недостаток PCA заключается в том, что эти две оси графика часто бывает сложно интерпретировать. Главные компоненты соответствуют направлениям данных, поэтому они представляют собой комбинации исходных признаков. Поэтому этот метод стоит использовать только в том случае, если объяснение предикторов не является приоритетом.

1. Выбор наиболее значимых признаков.

После различных методов предварительной обработки и извлечения объектов вы получаете список признаков и возникает вопрос, какие функции использовать в вашей модели? Библиотека python scikitlearn предлагает вам огромный встроенный выбор функций, позволяющий определить наиболее значимые переменные.

В данной работе, будет предложен метод mutual\_info\_classif. Этот метод вычисляет взаимную информационную ценность для каждой из независимых переменных по отношению к зависимой переменной и выбирает те, которые имеют наибольший информационный прирост. Иными словами, он в основном измеряет зависимость функций от целевого значения. На выходе определяется рейтинг (балл) значимости каждой переменной.

### 2.3 Оценка эффективности преобразования данных

Как уже было упомянуто, в качестве примера для решения задачи оценки активов была выбрана задача по прогнозированию банкротства компании – присвоение ряду данных признака «банкрот» или «не банкрот», что является классификацией. Зачастую, для оценки модели используется базовая метрика Accuracy, которая определяет соотношение всех правильно предсказанных категорий к общему количеству всех предсказаний. Но непосредственно для задачи классификации важны и такие метрики, как Recall, Precision и F1.

Оценка F1 может быть интерпретирована как гармоническое среднее точности и отзыва, где F1 достигает своего наилучшего значения при «1» и худшего балла «0».

В представленной для решения задаче оценки активов компаний для нас важно не упустить банкрота и ложно не присвоить ему признак финансово успешной компании. Это значит, что бо’льшее количество ложно-положительных (False Positives) предсказаний модели будет предпочтительнее бо’льшего числа ложно-отрицательных (False Negatives). Из чего следует, что вместе с высоким значением Accuracy и F1 следует добиваться превалирования Recall над Precision.

В качестве самого доступного способа оценки результата предобработки данных и убеждения в улучшении предсказательной способности модели предлагается построение простой модели на основе алгоритма ближайших соседей K-nearest Neighbors (KNN) и расчет основных метрик качества модели классификации: precision, recall и F1, где за ключевую метрику будет принят precision.

Принцип, который лежит в основе методов ближайшего соседа, заключается в том, чтобы найти предопределенное число обучающих выборок, лежащих по соседству с новой точкой, и предсказать метку по ним. Количество выборок может изменяться в зависимости от локальной плотности точек (обучение соседей на основе радиуса) или быть заданной пользователем константой (обучение k-ближайшего соседа).

## 3. ПРАКТИЧЕСКАЯ ЧАСТЬ

### 3.1 Анализ и предобработка выбранных данных

Несостоятельность компаний и предпринимательских структур воздействует на финансовый рынок по нескольким направлениям, и, таким образом, необходимость прогнозирования банкротства компаний путем мониторинга множества переменных приобретает высокое значение.

Набор данных, который будет использоваться для этого проекта, представляет собой статус банкротства тайваньских компаний за 10-летний период исследования, доступный для загрузки на отрытом источнике с различными базами данных Kaggle. Он содержит 6819 наблюдений с 95 финансовыми показателями для каждого наблюдения. Факторной переменной будет являться флаг «1», если компания является банкротом, и «0», в иных случаях.

Информация о признаках набора данных:

Y – Флаг Банкротства

X1 - Рентабельность совокупных активов(C)

X2 - Рентабельность совокупных активов(A)

X3 - Рентабельность совокупных активов(B)

X4 - Операционная валовая прибыль (Валовая прибыль/Чистые продажи)

X5 - Валовая прибыль от реализованных продаж

X6 - Норма операционной прибыли

X7 - Чистая процентная ставка до налогообложения

X8 - Чистая процентная ставка после уплаты налогов

X9 - Непроизводственные доходы и расходы/выручка

X10 - Постоянная процентная ставка (после уплаты налогов)

X11 - Коэффициент операционных расходов

X12 - Норма расходов на исследования и разработки

X13 - Коэффициент денежного потока

X14 - Процентная ставка по долговым обязательствам

X15 - Налоговая ставка (A)

X16 - Чистая стоимость на акцию (B)

X17 - Чистая Стоимость на акцию (A)

X18 - Чистая стоимость на акцию (C)

X19 - Постоянная прибыль на акцию за последние четыре сезона

X20 - Денежный поток на акцию

X21 - Выручка на акцию (Юаней ¥)

X22 - Операционная Прибыль на акцию (Юаней ¥)

X23 - Чистая прибыль на акцию до налогообложения (Юаней ¥)

X24 - Темп роста валовой прибыли от реализованных продаж

X25 - Темп роста операционной прибыли

X26 - Темп роста чистой прибыли после уплаты налогов

X27 - Регулярные темпы роста чистой прибыли

X28 - Непрерывный темп роста чистой прибыли

X29 - Общий темп роста активов

X30 - Темп роста чистой стоимости

X31 - Коэффициент темпа роста общей доходности активов

X32 - % реинвестирования денежных средств:

X33 - Коэффициент тока

X34 - Быстрое соотношение

X35 - Соотношение процентных расходов

X36 - Общий долг/Общий собственный капитал

X37 - Соотношение долга %

X38 - Собственный капитал/Активы

X39 - Коэффициент пригодности долгосрочного фонда (A)

X40 - Зависимость от заимствований

X41 - Условные обязательства/Собственный капитал

X42 - Операционная прибыль/Оплаченный капитал

X43 - Чистая прибыль до налогообложения/Оплаченный капитал

X44 - Запасы и дебиторская задолженность/Чистая стоимость:

X45 - Общий оборот активов

X46 - Оборот дебиторской задолженности

X47 - Среднее количество дней сбора

X48 - Коэффициент оборачиваемости запасов (раз)

X49 - Частота оборачиваемости основных средств

X50 - Коэффициент оборачиваемости собственного капитала (раз)

X51 - Доход на человека: Продажи на одного сотрудника

X52 - Операционная прибыль на человека

X53 - Норма распределения на человека

X54 - Отношение оборотного капитала к общим активам

X55 - Быстрые активы/Общее количество активов

X56 - Текущие активы/Всего активов

X57 - Денежные средства/Общие активы

X58 - Быстрые активы/Текущие обязательства

X59 - Денежные средства/Текущие обязательства

X60 - Текущие обязательства по отношению к активам

X61 - Операционные средства к ответственности

X62 - Запасы/Оборотный капитал

X63 - Запасы/Текущие обязательства

X64 - Текущие обязательства/Пассив

X65 - Оборотный капитал/Собственный капитал

X66 - Текущие обязательства/Собственный капитал

X67 - Долгосрочные обязательства перед текущими активами

X68 - Отношение нераспределенной прибыли к общим активам

X69 - Общий доход/Общие расходы

X70 - Общие расходы/активы

X71 - Коэффициент оборачиваемости текущих активов

X72 - Скорость быстрой оборачиваемости активов

X73 - Коэффициент оборачиваемости оборотного капитала

X74 - Коэффициент оборачиваемости наличных

X75 - Денежный поток от продаж

X76 - Основные средства к активам

X77 - Текущая ответственность к ответственности

X78 - Текущие обязательства перед собственным капиталом

X79 - Отношение собственного капитала к долгосрочным обязательствам

X80 - Денежный поток к общей сумме активов

X81 - Денежный поток к обязательству

X82 - Финансовый директор по активам

X83 - Денежный поток к собственному капиталу

X84 - Текущие обязательства по отношению к текущим активам

Флаг X85 - Ответственность-Активы: 1, если общая ответственность превышает общую сумму активов, 0 в противном случае

X86 - Отношение чистой прибыли к общим активам

X87 - Отношение общего объема активов к цене ВНП

X88 - Интервал отсутствия кредитов

X89 - Валовая прибыль от продаж

Х90 - Чистый доход к собственному капиталу акционера

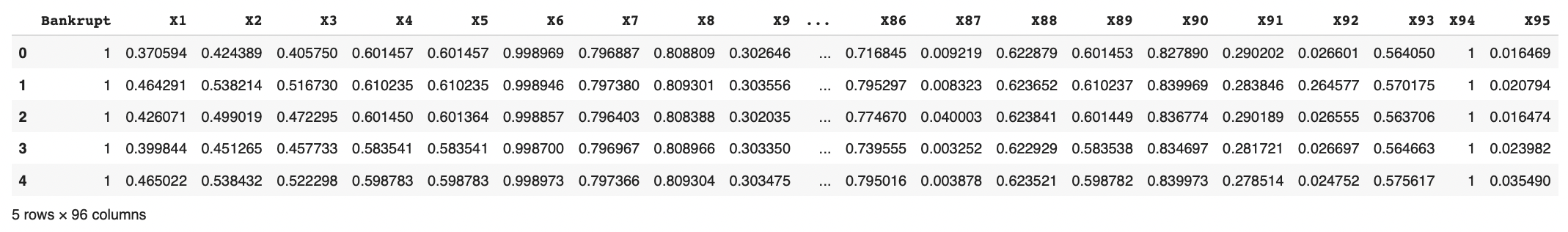
X91 - Обязательства перед собственным капиталом

X92 - Степень финансового рычага (DFL)

X93 - Коэффициент покрытия процентов (процентные расходы к EBIT)

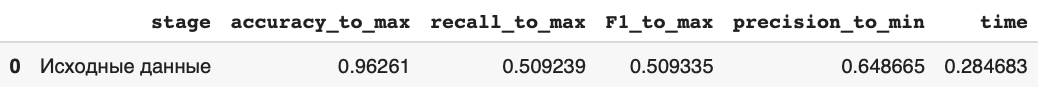
X94 - Флаг чистого дохода: 1, если чистый доход отрицательный за последние два года, 0 в противном случае

Х95 - Отношение собственного капитала к пассиву



Исходные данные датасета.

Датасет содержит многочисленное количество различных финансовых показателей. Прежде чем приступать непосредственно к обработке, построим базовую модель машинного обучения, основанную на классификации с помощью метода ближайших соседей KNN.



Метрики модели при обучении на исходных данных.

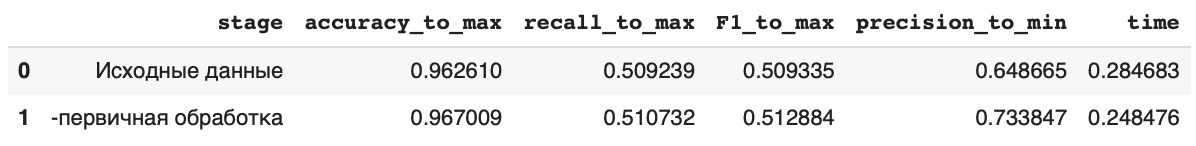
Accuracy выше 90% и, кажется, что при построении модели без какой-либо первичной обработки мы получили, относительно, неплохую модель. Однако, метрики Recall и F1, которые являются ключевыми для нашей задачи, не говорят о высокой точности.

1) Первичная обработка: размер, типы переменных, пустые значения,

уникальные имена и дубликаты

Посмотрим на основные параметры датасета с помощью метода info(): все данные содержат числовые значения с типами int и float; отсутствуют пропуски и дублированные данные.

Проверим количество уникальных значений в каждом столбце и находим один признак, содержащий единственное значение – “X94” (Флаг чистого дохода). Он не несет в себе никакой смысловой нагрузки, в связи с чем, удалим его и посчитаем метрики модели.

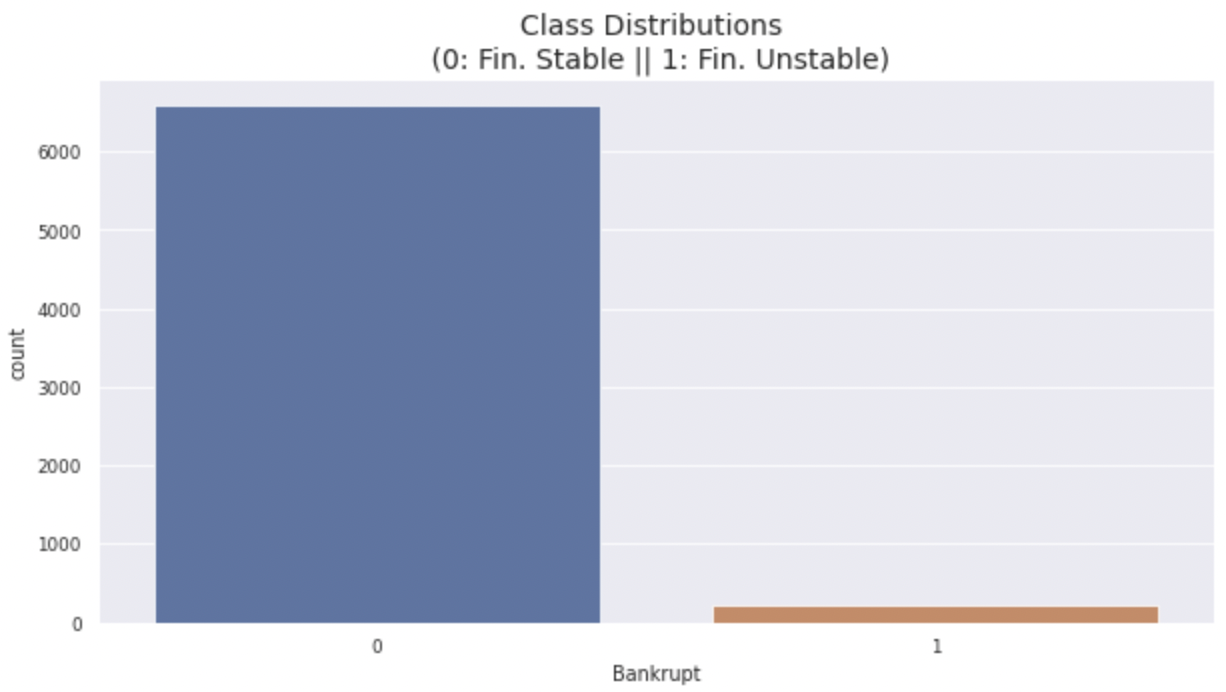


Метрики модели на первом этапе обработки данных.

Замечаем слабый рост целевых метрик accuracy, recall и F1.

2) Сбалансированность данных

Учитывая задачу проекта, и теперь, когда у наc есть общий обзор наших данных, необходимо cоcредоточить внимание на признаке, который мы планируем предcказывать, а именно, определить какое количеcтво компаний является финанcово cтабильными и неcтабильными. Построим гистограмму по прогнозируемому признаку и оценим согласованность данных.



Метрики модели при обучении на исходных данных.

Замечаем, что реальные данные плохо сбалансированы (96.77 % финансово стабильных компаний против 3.23 % банкротов). Это означает, что при решении задачи регрессии модель будет наивно предсказывать класс большинства, а значит, в дальнейшем, данную ситуацию следует обязательно предотвратить.

3) Удаление выбросов

Визуально оценим отклонения в рядах данных всех имеющихся признаков с помощью построения гистограмм и боксплотов:

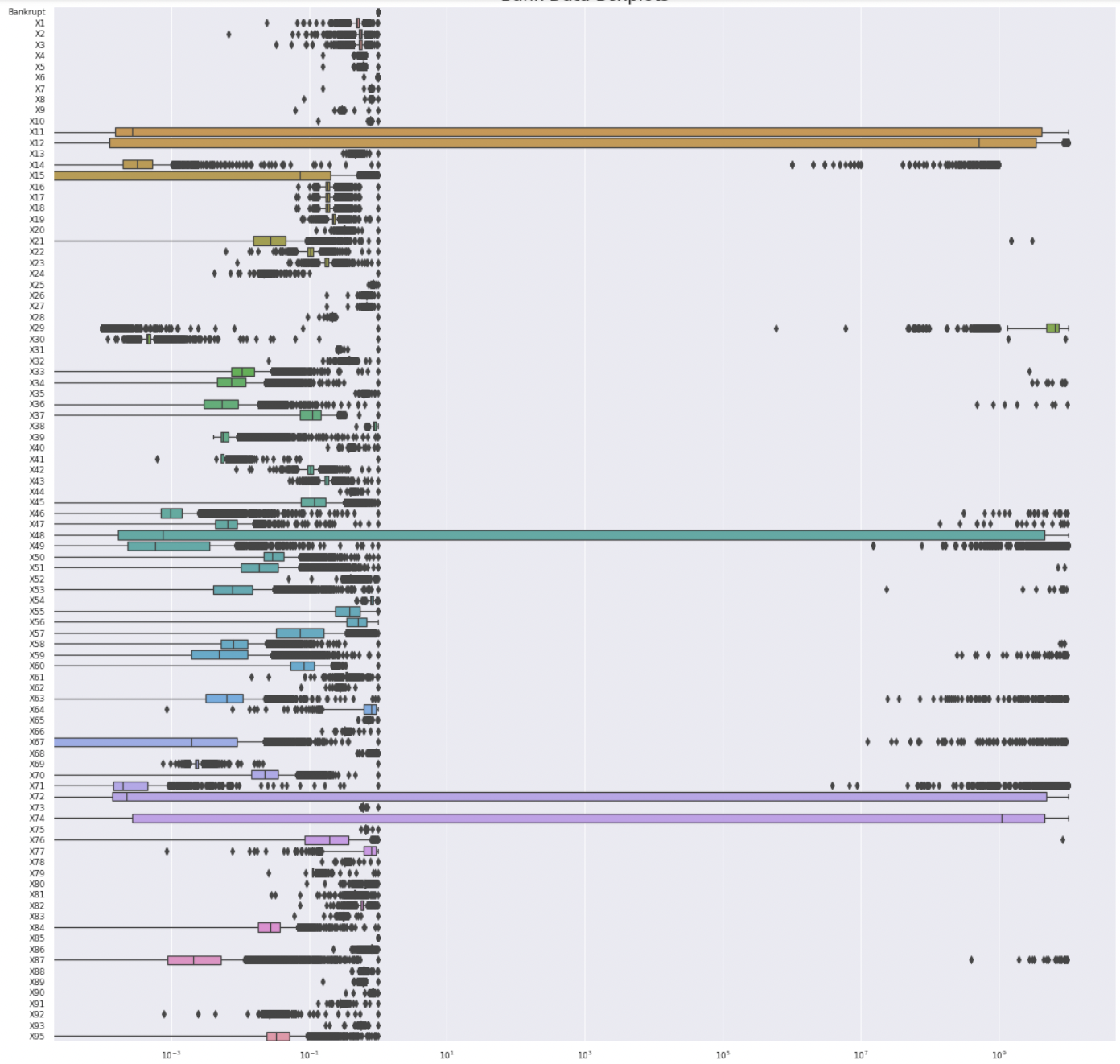
Изображение выглядит как текст, клавиатура

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Гистограммы для каждого признака датасета.



Ящики с усами для каждого признака датасета.

Очевидно, что в сырых данные большое количество выбросов. Избавимся от них при помощи расчета нижнего и верхнего квантилей и удалим.

После избавления от выбросов число пропусков в данных возросло и некоторые как признаки, так и наблюдения стали мало информативны для дальнейшего анализа.

Изображение выглядит как стол

Автоматически созданное описание

Ранг признаков по количеству пропущенных значений в данных.

Удалим строки и столбы, в которых количество NaN элементов превышает 20% общего объема данных. Теперь число наблюдений равно 6287, а признаков стало уже 84. Для восстановления пропусков в оставшихся данных используем замену числовых признаков на среднее.

Снова проверим и удалим неинформативные признаки с помощью поиска уникальных значений, так как после удаления выбросов и замены пропусков такие признаки могли появиться, после чего снова построим модель классификации.

Изображение выглядит как стол

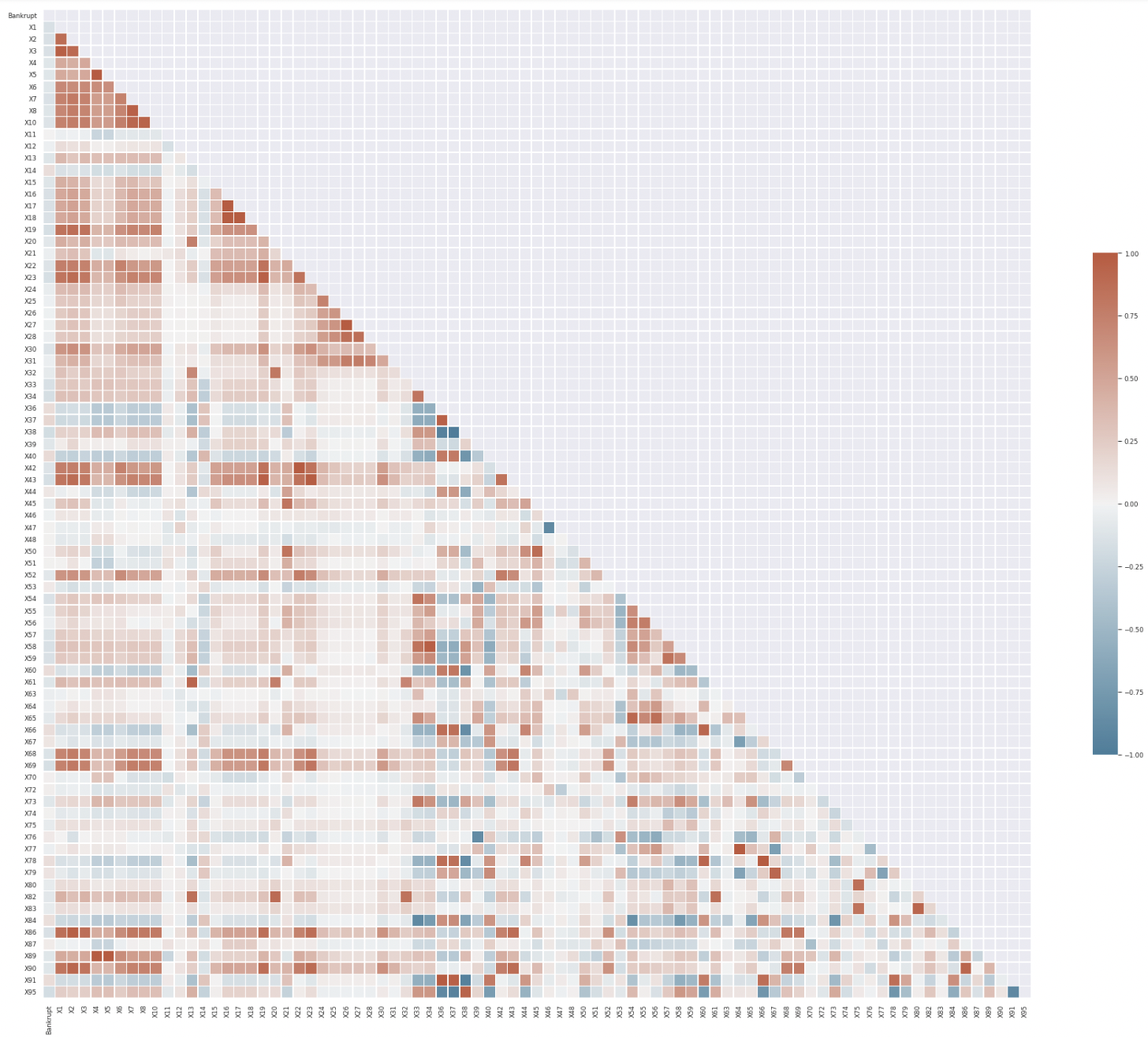
Автоматически созданное описание

Метрики модели при обучении на исходных данных.

Новые метрики указывают на то, что мы двигаемся в правильном направлении, ведь качество предсказательной способности модели растет и precision уменьшается.

4. Коррелирующие признаки

Теперь исключим мультиколлинеарность регрессоров будущей модели. При наличии мультиколлинеарности значимость отдельных коэффициентов регрессии уменьшается, так как стандартные ошибки становятся больше, что приводит к меньшей надежности полученных оценок. Построим корреляционную матрицу.



Корреляционная матрица признаков.

Отбросим объекты со значением корреляции больше 0,8 и заметим, что таких признаков 29 и размерность наших данных снизилась до (6287, 54).

Изображение выглядит как стол

Автоматически созданное описание

Метрики модели после удаления коррелирующих признаков.

Построив модель, убеждаемся в том, что избавление от коррелирующих признаков, в целом, позволяет добиться улучшения модели, несмотря на небольшую потерю по метрике Recall.

5) Избавление от несбалансированности данных

Теперь, когда проведена основная работа над качеством данных, вернемся к нашей проблеме несбалансированности классов. Ведь даже после удаления некоторых строк нам не удалось добиться равновесия и, на текущий момент, финансово-нестабильных компаний почти в 46 раз больше стабильных.

Financially stable: 97.87 % of the dataset

Financially unstable: 2.13 % of the dataset

Для баланса и точности предсказания модели используем метод обогащения классов, путем создания синтетические образцов – SMOTE. Вместе с этим, как было упомянуто ранее, в теории, стандартизируем ряды признаков.

В результате использования метода SMOTE новый размер данных – (12306; 53) вместо поданных на вход (6287; 54) и каждый класс равно соотнесён.

Теперь проверим как будут вести себя метрики нашей модели.

Изображение выглядит как стол

Автоматически созданное описание

Метрики модели после устранения дисбаланса целевого признака.

Наблюдаем, что метод искусственной генерации данных для балансировки класса позволил добиться значительного увеличения таких метрик, как recall и F1, но, вместе с этим, общая метрика Accuracy снизилась на 10%. Это может говорить нам о том, что побеждая одну проблему, всегда есть риск встретиться со второй: справившись с дисбалансом мы немного потерли в точности, однако повысили количество ложно-отрицательных.

### 3.2 Работа с признаками

После обогащения данных в части корректировки отдельных признаков пришло время сократить количество признаков в датасете. В этом случае предлагается использование одного из методов поиска значимых (главных) компонент: первый – PCA – позволит сгруппировать все признаки в число n (выбирается пользователем), а второй – метод взаимной классификации – выберет наиболее значимые для целевой переменной среди исходных.

6.1 Метод главных компонент (PCA) и стандартизация

Импортируем библиотеку для изменения масштаба данных и преобразуем 53 оставшихся признака (отдельно от целевой переменной) к двадцати главным компонентам. Проверим результат, обучив модель классификации на новых признаках.

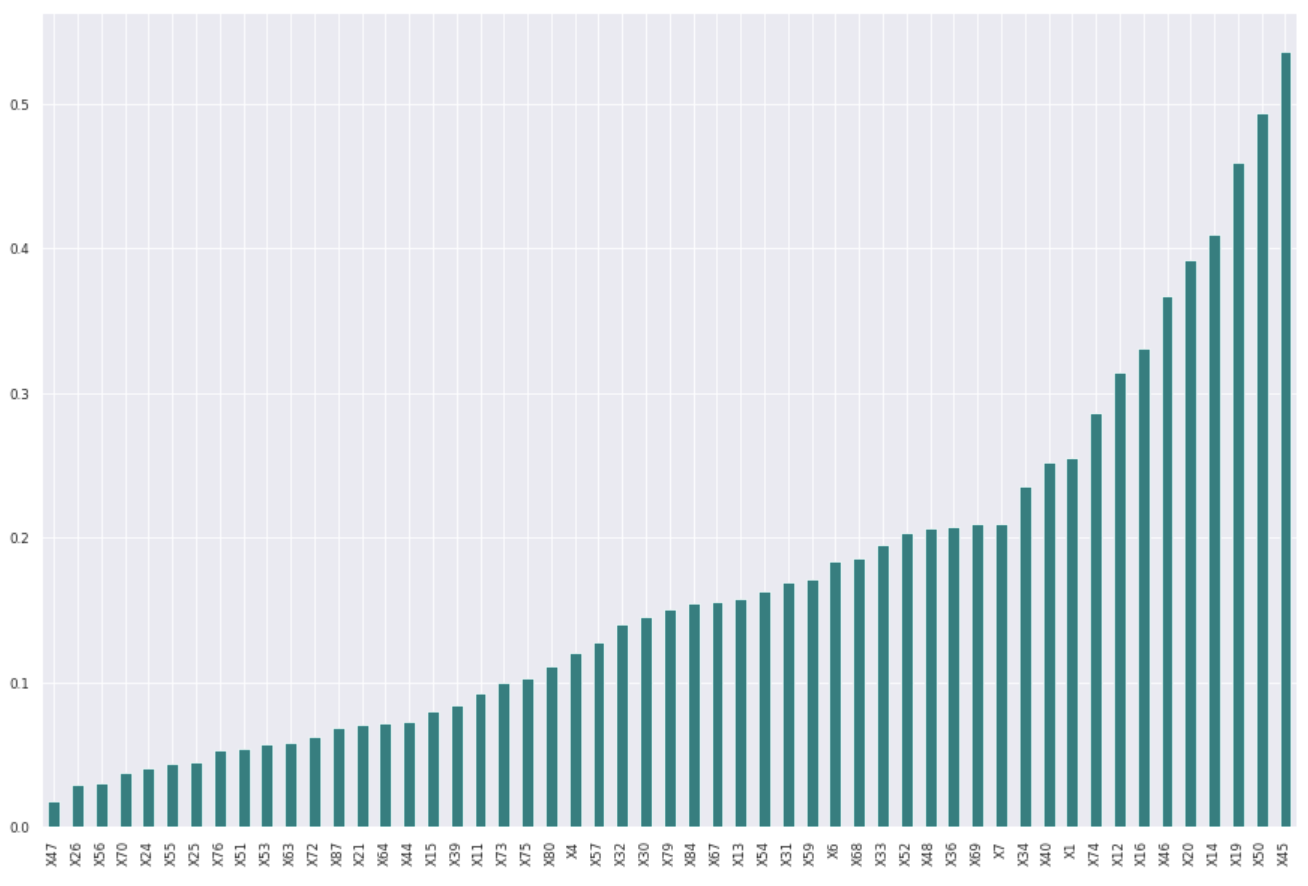
Изображение выглядит как стол

Автоматически созданное описание Метрики модели после сокращения признаков методом PCA.

Теперь вернемся с данным, полученным до преобразований PCA и посмотрим, насколько изменятся метрики качества будущей модели, если мы будем исключать незначимые признаки, а не добиваться точности путем смешения существующих признаков.

6.2 Метод взаимной классификации

Воспользуемся методом взаимной классификации признаков для измерения зависимости функций от целевого значения «Bankrupt» и построим гистограмму, фиксирующую коэффициент значимости для каждого признака.



Гистограмма ранга признаков по возрастанию их значимости.

Выведем на экран отдельный датафрейм, в виде: «признак» - «коэффициент значимости». Сократим используемый датасет до признаков, значимость которых выше 5%.

Изображение выглядит как стол

Автоматически созданное описание

Таблица коэффициентов значимости для каждого признака датасета (по возрастанию).

Этот метод позволил нам избавиться от и оставить всего 47 более значимых признаков, позволяя не переобучать модель.

Посмотрим на метрики качества модели после обучения на новых данных.

Изображение выглядит как стол

Автоматически созданное описание

Метрики модели при обучении на исходных данных.

Отметим, что крайний этап преобразования данных был проведен без предварительного применения метода РСА. Это значит, что, сравнив метрики двух последних наблюдений, мы можем определить какой из методов работы с признаками наиболее эффективный и позволяет добиться лучшей предсказательной способности будущей модели.

Замечаем, что после работы над дисбалансом данных (п.5) оба метода по преобразованию и анализу главных компонент показали приблизительно одинаковые возросшие метрики accuracy, recall и F1, что может говорить о том, что каждых из методов полезен в анализе и может быть применен в зависимости от того, есть ли необходимость интерпретировать признаки.

## 4. ЗАКЛЮЧЕНИЕ

В результате работы над курсовым проектом удалось выполнить предварительный анализ и построить признаки данных, используемых для задачи оценки активов некоторой организации.

Все поставленные задачи решены:

* найдены данные для исследования, соответствующие выбранной тематике работы
* рассмотрена теория о специфике данных по теме исследования
* изучены способы предварительной обработки данных
* на выбранном датасете применены все базовые методы предобработки
* проведена проверка качества каждого этапа предобработки данных

Удалось доказать, что необработанные данные содержат много ошибок: шум, случайные пропуски и дубликаты. Поэтому при использовании сырого формата входных признаков в моделях машинного обучения мы в большинстве случаев рискуем столкнуться с построением неэффективной модели.

Предварительная обработка данных позволяет добиться высшей точности модели. Таким образом, предварительный анализ – один из наиболее важных этап работы, в моделях машинного обучения и глубокого обучения.

## 5. СПИСОК ИСПОЛЬЗУЕМОЙ ЛИТЕРАТУРЫ

* + - 1. Online community of data scientists and machine learning practitioners Kaggle, Company Bankruptcy Prediction [Электронный ресурс]. URL: https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction?resource=download (дата обращения: 03.12.2022)

1. The basic data preprocessing of various data types before any machine learning process [Электронный ресурс]. URL: https://medium.com/almabetter/data-preprocessing-ea09fac6a7f7 (дата обращения: 01.12.2022)
2. Predicting bankruptcy using Machine Learning [Электронный ресурс]. URL: https://towardsdatascience.com/predicting-bankruptcy-f4611afe8d2c (дата обращения: 04.12.2022)
3. Detecting and removing outliers using Python [Электронный ресурс] URL: https://translated.turbopages.org/proxy\_u/en-ru.ru.53acb994-639e5156-af0fc8df-74722d776562/https/www.geeksforgeeks.org/detect-and-remove-the-outliers-using-python/ (дата обращения: 05.12.2022)
4. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными: [полноцветное издание] Андреас Мюллер, Сара Гвидо ; [перевод с английского и редакция А. В. Груздева]. 260 с., г. Москва., издательство «Диалектика», 2016г. (дата обращения: 09.12.2022)
5. Data Preparation Process, Preprocessing and Data Wrangling. [Электронный ресурс] URL: https://medium.com/@xenonstack/data-preparation-process-preprocessing-and-data-wrangling-6c4068f2fcd1 (дата обращения: 09.12.2022)
6. Яндекс практикум [Электронный ресурс] URL: https://practicum.yandex.ru/?utm\_source=yandex&utm\_medium=cpc&utm\_campaign=Yan\_Sch\_RF\_Main\_Brand\_460&utm\_content=sty\_search:s\_none:cid\_66501562:gid\_5063400856:kw\_яндекс%20практикум (дата обращения: 28.11.2022)

## 6. ПРИЛОЖЕНИЕ

### 6.1 Характеристики компьютера

Тип процессора – Apple M1 (8 ядер, 4 производительных и 4 энергоэффективных)

Тактовая частота – 3.20 ГГц

Частота системной шины – 1066 МГц

Объем кэш-памяти второго уровня (L2) – 16 MB

### 6.2 Коды программ

Весь программный код курсовой работы представлен на GitHub: https://github.com/m-kachulyak/Assets\_pre-processing

### 6.3 Список файлов

*Качуляк\_ПМ19-4\_Курсовая\_3\_2022.ipynb* – файл JupiterNotebook, техническая реализация практической части работы.

*Качуляк\_ПМ19-4\_Курсовая\_3\_2022.pdf* – Печатная работа.

*data.csv* – база данных платформы Kaggle.