МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу
«Data Science»

Тема: «Прогнозирование доходов крестьянских (фермерских) хозяйств»

Слушатель Дашиева Баярма Шагдаровна

Москва, 2023

Содержание

Введение

Актуальность исследования. С целью сохранения сельских территорий и целостности страны необходимо усилить меры по поддержке малого и среднего предпринимательства. В настоящее время требуется проведение подробного анализа больших массивов данных, собираемых Министерством сельского хозяйства России, Росстатом и другими ведомствами.

Актуальность исследований в области анализа данных и искусственного интеллекта подтверждается национальной программой «Цифровая экономика Российской Федерации», ведомственным проектом «Цифровое сельское хозяйство». К 2024 г. в ведомственном проекте «Цифровое сельское хозяйство» планируется достижение роста производительности труда на «цифровых» сельскохозяйственных предприятиях в 2 раза вследствие внедрения цифровых технологий и платформенных решений, цифровой трансформации сельского хозяйства. Компенсировать сокращение потребности трудовых ресурсов вследствие роста производительности труда может общемировой тренд на развитие более трудоемкого органического сельского хозяйства, в том числе в малых формах хозяйствования.

Ведомственный проект «Цифровое сельское хозяйство» предполагает выполнение задач по созданию и внедрению национальной платформы цифрового государственного управления сельским хозяйством «Цифровое сельское хозяйство» (ЦСХ), которая предполагает разработку системы сбора, хранения и обработки данных о ресурсах и результатах сельскохозяйственного производства и разработку системы интеллектуального анализа данных и прогнозирования на основе технологий Advanced Analytics, Data Discovery, Data Mining, Machine Learning и искусственного интеллекта.

Целью выпускной квалификационной работы является прогнозирование доходов крестьянских (фермерских) хозяйств (КФХ).

Задачи выпускной квалификационной работы:

1) Изучить теоретические основы и методы решения поставленной задачи.

2) Провести разведочный анализ данных: нарисовать гистограммы распределения каждой из переменной, диаграммы ящика с усами, попарные графики рассеяния точек; получить среднее, медианное значение; провести анализ и исключение выбросов, проверить наличие пропусков.

3) Провести предобработку данных (удалить шумы, нормализовать данные и т.д.).

4) Обучить несколько моделей для прогноза доходов КФХ (30% данных оставить на тестирование модели). При построении моделей провести поиск гиперпараметров модели с помощью поиска по сетке с перекрестной проверкой с количеством блоков равным 10.

5) Написать нейронную сеть, которая будет прогнозировать доходы КФХ.

6) Разработать приложение с графическим интерфейсом или интерфейсом командной строки, которое будет выдавать прогноз, полученный в задании 4 или 5 (один или два прогноза).

7) Оценить точность модели на тренировочном и тестовом датасете.

8) Создать репозиторий в GitHub / GitLab и разместить код исследования. Оформить файл README.

Объектом исследования являются крестьянские (фермерские) хозяйства Ставропольского края, специализирующиеся на зернопроизводстве.

Предметом исследования – модели, прогнозирующие доходы КФХ

Структура работы. Выпускная квалификационная работа состоит из двух глав: аналитическая часть и практическая часть.

Глава 1 Аналитическая часть

1.1 Постановка задачи

Для автоматизации прогнозирования доходов КФХ необходимо разработать веб-приложение Flask, позволяющее получать прогнозные значения доходов КФХ в зависимости от факторов. С это целью необходимо построить различные модели с использованием методов машинного обучения.

Характеристика датасета

Датасет представлен обезличенными данными из формы № 1-КФХ ведомственной отчетности Минсельхоза России «Информация о производственной деятельности глав крестьянских (фермерских) хозяйств – индивидуальных предпринимателей». В форме № 1-КФХ отражаются сведения о размере доходов КФХ, расходов КФХ, количестве членов КФХ, включая главу КФХ, о численности постоянных наемных работников КФХ, площади земельных участков и объектов природопользования, площади посевной площади, наличии сельскохозяйственной техники, сумме кредитов и займов, полученных хозяйством и др. К недостаткам данной формы, относится то, что не приводятся данные о численности временных и сезонных работников, затратах труда всех работников КФХ, о характере и степени занятости в КФХ: полная или частичная занятость.

Out[4]:

	№	Доходы, руб - за 2019 год, руб.	в том числе: от реализации селькохозяйственной продукции, продуктов её первичной и промышленной переработки - за 2019 год	от оказания услуг - за 2019 год	получено средств государственной поддержки (субсидии, гранты) - за 2019 год	Расходы, тыс. руб - за 2019 год	в том числе: расходы на приобретение основных средств, включая лизинговые платежи \n(стр.231211+ 231212+ 231213+ 231214) - за 2019 год	из них: техника, машины и оборудование - за 2019 год	племенные и продуктивные животные - за 2019 год	земельные участки - за 2019 год	...	Сельскохо: техника налич
0	1	111242000.0	84843000	25959000.0	330000.0	73295000	120000.0	120000.0	NaN	NaN	...	
1	2	4734000.0	3572000	NaN	1162000.0	4181000	35000.0	NaN	35000.0	NaN	...	
2	3	6147000.0	6147000	NaN	NaN	4650000	NaN	NaN	NaN	NaN	...	
3	4	4132000.0	4132000	NaN	NaN	4120000	NaN	NaN	NaN	NaN	...	
4	5	2676000.0	2676000	NaN	NaN	2112000	228000.0	28000.0	NaN	200000.0	...	

5 rows × 56 columns

Рисунок 1 – Данные по крестьянским (фермерским) хозяйствам

Характеристика затруднена из-за отсутствия в форме 1-КФХ информации о размере площади сельскохозяйственных угодий, в т.ч. используемой,

себестоимости произведенной продукции, о количестве временных и сезонных работников, о затраченном времени работы в хозяйстве и вне хозяйства и др. (56 столбцов).

Статистическая обработка первичных данных позволила получить относительные показатели КФХ.

Объем выборки – 1202 крестьянских (фермерских) хозяйств Ставропольского края, в основном специализирующиеся на производстве зерновых и зернобобовых культур, где удельный вес продукции зернопроизводства превышал 50% от общего размера выручки продукции сельского хозяйства (согласно приказу Росстата № 742 от 31.12.2014 (ред. от 04.02.2016) «О Методических указаниях по определению основного вида экономической деятельности хозяйствующих субъектов на основе Общероссийского классификатора видов экономической деятельности (ОКВЭД 2) для формирования сводной официальной статистической информации»).

В качестве входных переменных (факторов) выбраны признаки, характеризующие ресурсы производства: численность работников КФХ (чел.), наличие тракторов (шт.), комбайнов (шт.), общая площадь земли (га) в расчете на одно хозяйство и эффективность производства - урожайность зерновых (ц/га).

В качестве выходной переменной (результативной, целевой) – доходы КФХ в расчете на одно хозяйство.

Численность всех работников КФХ по изначальному датасету (KFH_dataset) определена как сумма членов КФХ и постоянных наемных работников. Среднегодовое число тракторов определено как среднеарифметическая из суммы тракторов на начало года и на конец года. Таким же образом найдено среднегодовое число комбайнов, и среднегодовая общая площадь земли.

Характеристика выборки с точки зрения ее особенностей (выбросы, пропуски и т.д.).

На первоначальном этапе определено количество пропусков в исходном датасете (1202 наблюдения).

Таблица 1 – Количество пропущенных значений в датасете по выбранным показателям

Показатель	**Количество пропусков**
Доходы, руб.	0
Численность постоянных работников, чел	654
Члены КФХ (включая главу КФХ), чел	0
Зерновые и зернобобовые культуры на зерно и семена (кроме рис) - урожайность, ц/га	0
в том числе: тракторы - наличие на начало года	209
в том числе: тракторы - наличие на конец года	186
комбайны - наличие на начало года	507
комбайны - наличие на конец года	494
Земельные участки и объекты природопользования - всего, га - наличие на начало года	10
Земельные участки и объекты природопользования - всего, га - наличие на конец года	0

В таблице приведен результат поиска количества нулей по изучаемым показателям. В представленном датасете видно, что не все КФХ нанимают работников. Только 45,6% КФХ нанимают работников. Не имеют тракторов 209 КФХ, или 17,4% всех КФХ. Только у 695 хозяйств, или у 57,8% хозяйств имеются в наличии комбайны. Все пропуски заменены на нулевые значения, так как это говорит об отсутствии наемных работников, или же об отсутствии сельскохозяйственной техники в хозяйстве.

Далее определим, есть ли выбросы по изучаемым признакам, построив графики boxplot (рисунок 2).

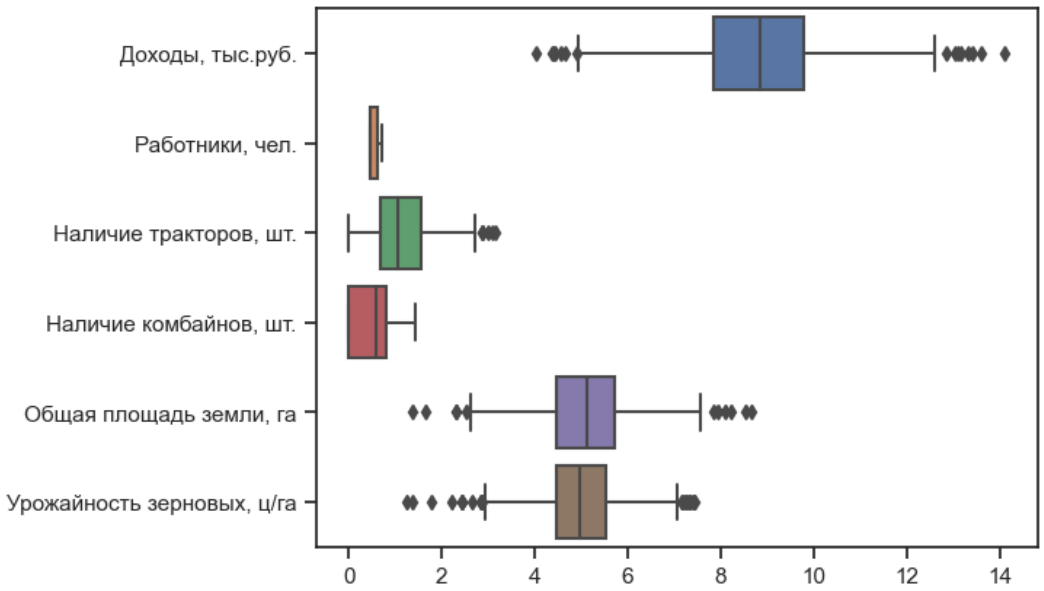


Рисунок 2 – Диаграмма «ящик с усами» по нормированным данным

Из рисунка 2 видно, что практически по всем изучаемым признакам имеются выбросы, которые необходимо будет удалить.

Далее проверим фактические распределения хозяйств по изучаемым признакам на соответствие их нормальному распределению.

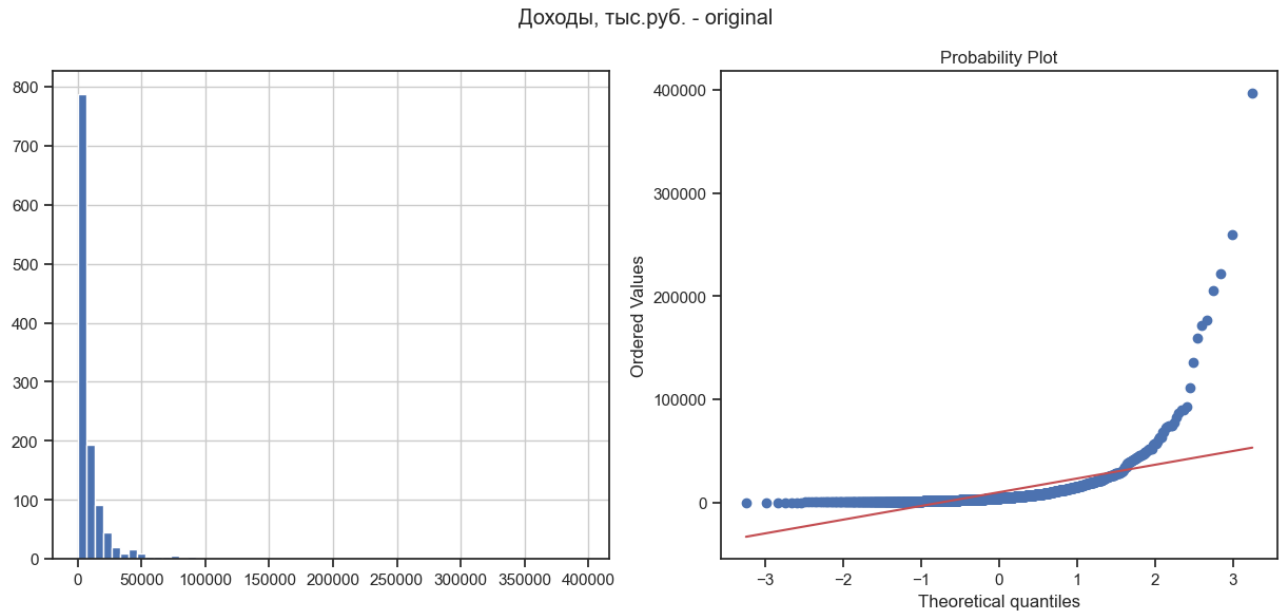


Рисунок 3 – Гистограмма и график Q-Q распределения КФХ по доходам КФХ

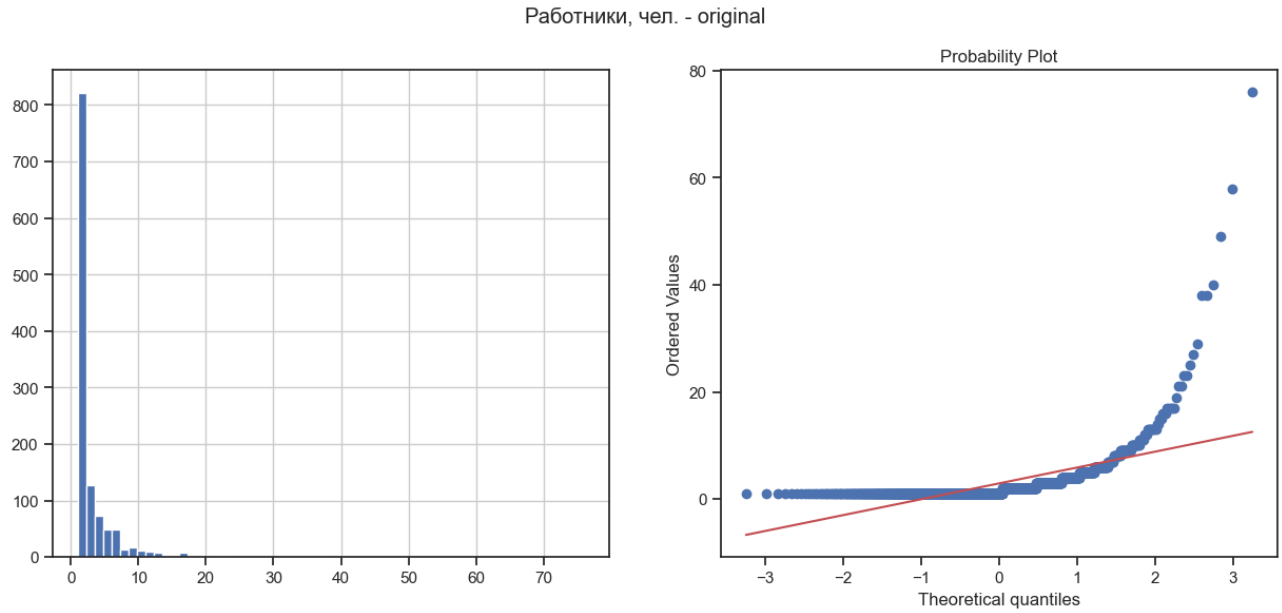


Рисунок 4 – Гистограмма и график Q-Q распределения КФХ по численности работников КФХ

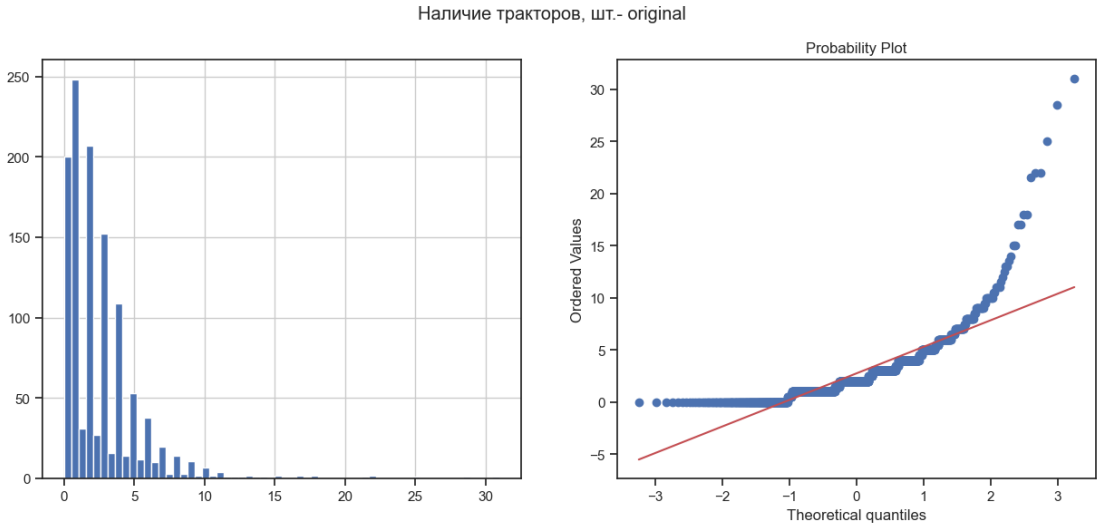


Рисунок 5 – Гистограмма и график Q-Q распределения КФХ по числу тракторов КФХ

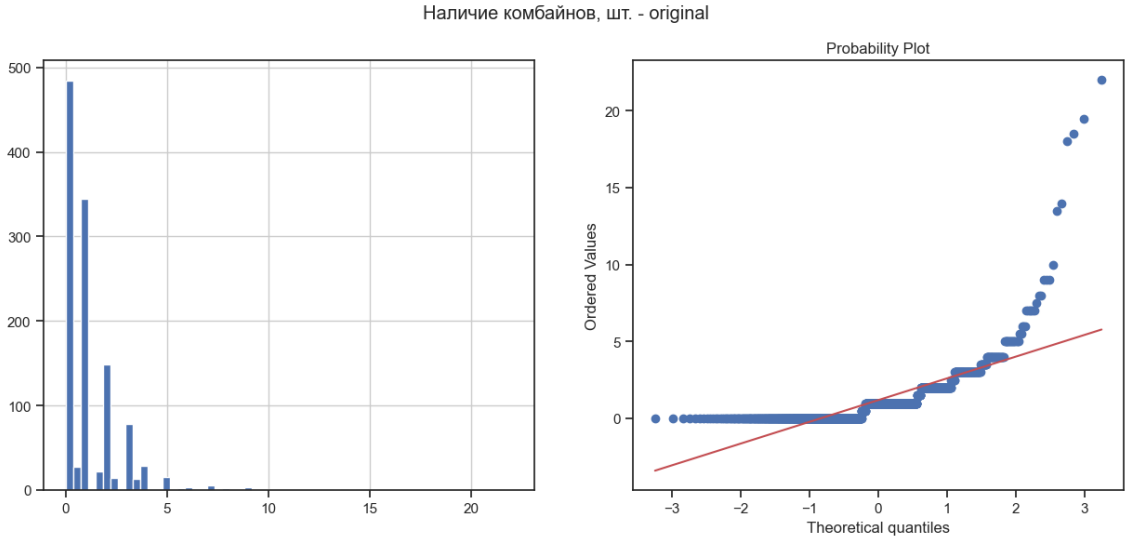


Рисунок 6 – Гистограмма и график Q-Q распределения КФХ по числу комбайнов КФХ

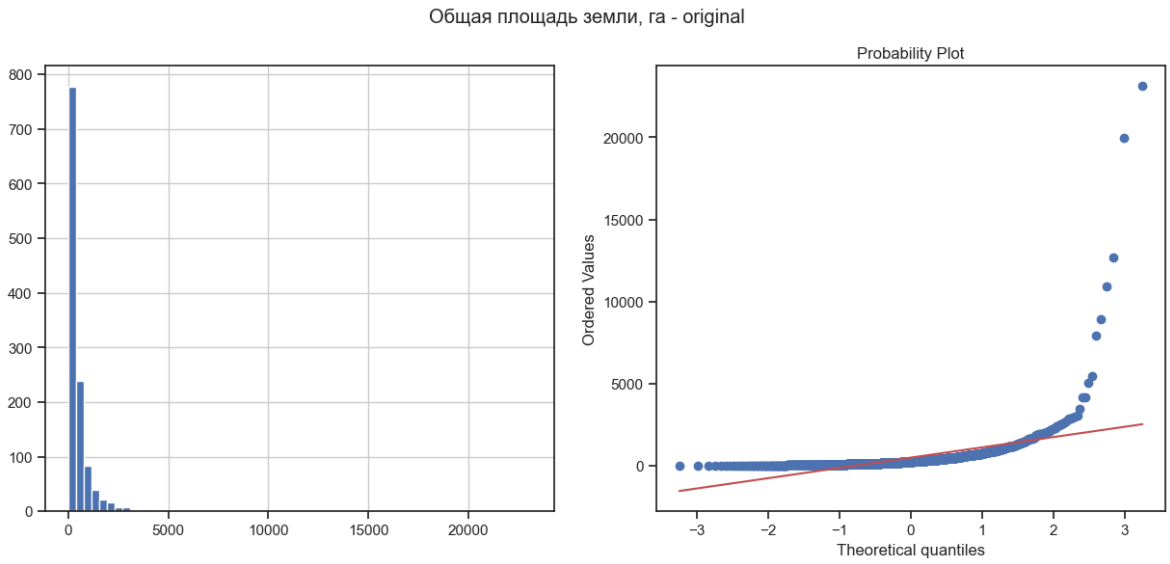


Рисунок 7 – Гистограмма и график Q-Q распределения КФХ по общей площади земли

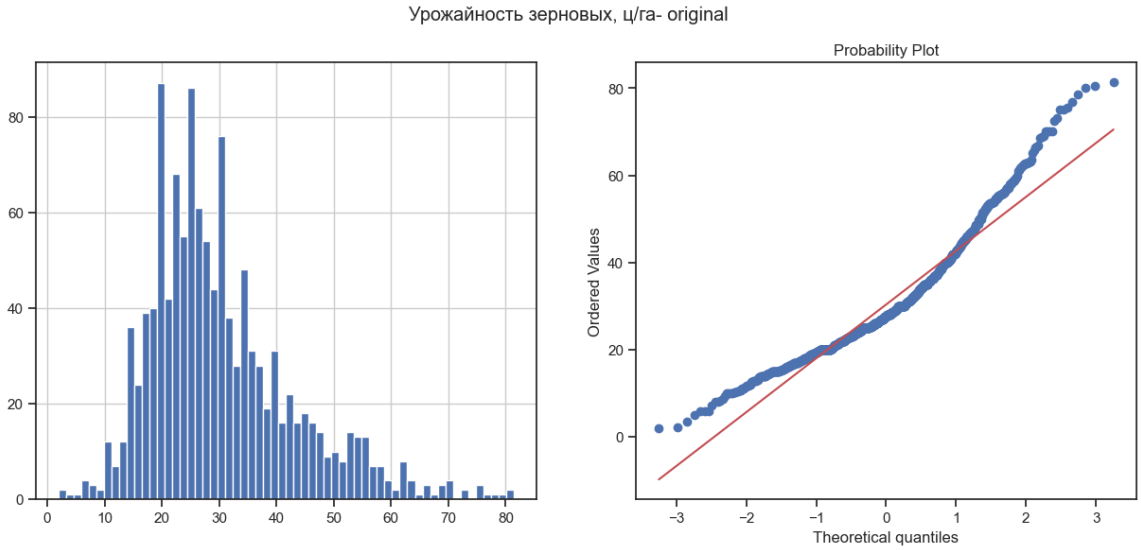


Рисунок 8 – Гистограмма и график Q-Q распределения КФХ по урожайности зерновых

Гистограммы и графики Q-Q показали, что распределения КФХ по всем признакам отличаются от нормального. Следовательно, нужно будет провести нормировку данных.

1.2 Описание используемых методов

Множественная линейная регрессия по МНК

Для учета влияния на зависимую переменную комплекса факторов в случае линейной связи используется модель множественной линейной регрессии:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i,$$

где i меняется от 1 до n, а n – это число наблюдений;

p – число переменных, включенных в модель.

Выборочной оценкой этой модели является уравнение:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i$$

Параметры при независимых переменных называют коэффициентами чистой регрессии, каждый из которых показывает, на сколько изменится зависимая переменная, если анализируемый фактор изменится на одну единицу своего измерения, при условии, что другие факторы останутся зафиксированными на среднем уровне.

Для оценки параметров используется метод наименьших квадратов, который в условиях множественной регрессии требует отсутствия коррелированности (коллинеарности или мультиколлинеарности в случае большого число связанных переменных) независимых переменных. Включение в модель регрессии переменных прямо или обратно пропорциональных другой или другим переменным является избыточным.

Для оценки параметров, в случае выполнения требований об отсутствии автокорреляции и гетероскедастичности остатков, а также мультиколлинеарности факторов, может быть использован метод наименьших квадратов, условие минимизации в матричной форме можно записать в виде:

$$S = \sum_{i=1}^{n} (\tilde{y}_i - y_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = (Y - Xb)'(Y - Xb) \to \min,$$

решением будет вектор оценок параметров:

$\beta = (X'X)^{-1} X'Y$.

Требование отсутствия мультиколлинеарности факторов связано с необходимостью получения обратной матрицы для оценок МНК. Мультиколлинеарность – высокая взаимная коррелированность объясняющих переменных – приводит к тому, что матрица оказывается вырожденной (случай функциональной (явной) мультиколлинераности), так как содержит линейно зависимые векторы-столбцы, а, следовательно, ее определитель равен нулю и не существует обратной матрицы , метод наименьших квадратов не позволяет в этом случае найти вектор оценок параметров уравнения регрессии. На практике чаще встречается стохастическая (неявная) форма мультиколлинеарности, и, хотя матрица в этом случае неособенная, ее определитель очень мал, что приводит к значительным ошибкам оцениваемых параметров.

Поэтому следствием мультиколлинеарности могут являтся:

1) незначимость большинства или всех оценок при независимых переменных множественной регрессии по t-критерию при значимости уравнения в целом по F-критерию;

2) при незначительном изменении исходных данных (увеличении (уменьшении) числа наблюдений) оценки существенно изменяются;

3) трудность (невозможность) интерпретации параметров регрессии с экономической точки зрения.

В случае коллинеарности факторов применяются модели ридж-регрессии.

Для оценки моделей с гетероскедастичными остатками переходят к использованию взвешенного метода наименьших квадратов.

Ридж-регрессия

Одним из способов устранения мультиколлинеарности между факторами является построение «ридж-регрессии»:

$$B = (X'X+\lambda \mathrm{E}_{p+1})^{-1}X'Y,$$

где λ – некоторое положительное число, «гребень», E_{p+1} – единичная матрица размера p+1 (число параметров с учетом условного начала) на p+1. Оценки гребневой регрессии получаются смещенными, но добавление λ увеличивает определитель матрицы $X'X$: $X'X + \lambda \mathrm{E}_{p+1}$, он уже не будет равен нулю, и уменьшает ошибки параметров регрессии.

Лассо-регрессия

Регрессия по методу наименьших квадратов (МНК) часто может стать неустойчивой, то есть сильно зависящей от обучающих данных, что обычно является проявлением тенденции к переобучению. Избежать такого переобучения помогает регуляризация - общий метод, заключающийся в наложении дополнительных ограничений на искомые параметры, которые могут предотвратить излишнюю сложность модели. Смысл процедуры заключается в “стягивании” в ходе настройки вектора коэффициентов β таким образом, чтобы они в среднем оказались несколько меньше по абсолютной величине, чем это было бы при оптимизации по МНК.

Метод регрессии “лассо” (LASSO, Least Absolute Shrinkage and Selection Operator) заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение. При этом достигается некоторый компромисс между

ошибкой регрессии и размерностью используемого признакового пространства, выраженного суммой абсолютных значений коэффициентов |β|. В ходе минимизации некоторые коэффициенты становятся равными нулю, что, собственно, и определяет отбор информативных признаков.

При значении параметра регуляризации λ=0 лассо-регрессия сводится к обычному методу наименьших квадратов, а при увеличении λ формируемая модель становится все более “лаконичной”, пока не превратится в нуль-модель. Оптимальная величина λ находится с использованием перекрестной проверки, т.е. ей соответствует минимальная ошибка прогноза на наблюдениях, не участвовавших в построении самой модели.

Метод k-ближайших соседей

Метод k-ближайших соседей является одним из наиболее простых методов. Значение целевого признака определяется на основе значений целевых признаков тех объектов, которые находятся ближе всего к искомому объекту в пространстве признаков. Исторически является одним из наиболее известных методов. В терминологии Data Mining рассматривался как основной алгоритм поиска по прецедентам. Метод может использоваться как для классификации, так и для регрессии. Как правило, метод k-NN показывает худшее качество, по сравнению с другими, более сложными методами.

Достоинства: Простота и универсальность метода. Возможность использования для классификации и регрессии. Возможность использования для обучения без учителя.

Недостатки. Метод в целом считается не очень точным. Зависимость от гиперпараметра K. При большом количестве точек перебор и вычисление расстояния занимают много времени. Для этого используются различные методы ускорения алгоритма.

Дерево решений

Фактически, алгоритм построения обучающего дерева, детально описанный здесь, сводится к нескольким пунктам:

Для текущего выбранного признака (колонки) из N признаков построить все варианты ветвления (разбиения) по значениям (для категориальных признаков) или по диапазонам значений (для числовых признаков). При этом будет сформировано K поддеревьев (где K - число ветвлений). Каждое поддерево содержит подвыборку, которая включает только строки выборки, соответствующие результатам ветвления. В каждом поддереве расположена: или выборка, содержащая N-1 признак, если признак, для которого строится ветвление, полностью пропадает в результате ветвления; или выборка, содержащая N признаков, если признак, для которого строится ветвление, не пропадает полностью в результате ветвления, но при этом число строк в выборке уменьшается. Если подвыборке соответствует единственное значение целевого признака, то в дерево добавляется терминальный лист, который соответствует предсказанному значению. Если в подвыборке больше одного значения целевого признака, то предыдущие пункты выполняются рекурсивно для подвыборки.

Преимущества деревьев решений. Работают по принципу "белого ящика". Логика построенного дерева хорошо отображается на исследуемую предметную область. Можно визуализировать алгоритм в виде дерева. Требуют мало данных для обучения. Работает с числовыми и категориальными признаками.

Недостатки деревьев решений. Могут переобучаться. Для борьбы с переобучением используется регулирование глубины дерева или "стрижка" дерева. Очень сильно зависят от набора данных в обучающей выборке. Появление одного нового примера может полностью перестроить весь каскад условий. Однако, на этом недостатке построено использование дерева в ансамблевых классификаторах, поэтому в ансамблях на основе деревьев данный недостаток можно рассматривать как достоинство.

Случайный лес

Алгоритм случайного леса сочетает в себе две основные идеи: метод бэггинга, предложенный Лео Брейманом, и метод случайных подпространств, предложенный Tin Kam Ho. Случайный лес можно рассматривать как алгоритмом бэггинга над решающими деревьями. Но при этом каждое

решающее дерево строится на случайно выбранном подмножестве признаков. Эта особенность называется "feature bagging" и основана на методе случайных подпространств. Метод случайных подпространств позволяет снизить коррелированность между деревьями и избежать переобучения. Базовые алгоритмы обучаются на случайно выбранных подмножествах признаков. Ансамбль моделей, использующих метод случайного подпространства, можно построить, используя следующий алгоритм: Чтобы применить модель ансамбля к тестовой выборке, объединяются результаты отдельных моделей или мажоритарным голосованием или более сложными способами.

Деревья очень чувствительны к выбросам и изменению данных. Незначительное изменение обучающей выборки, удаление или добавление признаков может сильно изменить вид решающего дерева. Это является недостатком с точки зрения построения отдельного решающего дерева. Это же является преимуществом при использовании деревьев в ансамблевой модели, так как с помощью незначительного изменения обучающих данных или набора используемых признаков можно построить очень разные решающие деревья. Различные решающие деревья дадут хорошую "дисперсию" решений при объединении их в ансамбль.

Нейронная сеть

Достоинства:

1. Гибкость: большое число параметров модели позволяет подбирать архитектуру, наиболее эффективно работающую с конкретным набором данных.

2. Универсальность: нейронные сети способны решать множество разнотипных задач, в том числе задачи регрессии, классификации, кластеризации.

3. Множество моделей: существует большое число различных архитектур нейронных сетей, выбор наиболее подходящей из них может привести к успешному решению поставленной задачи, при безуспешных попытках ее решить, используя какие-либо другие методы.

Недостатки:

1. Черный ящик: человек не до конца понимает, как работает нейросеть и какие именно данные из предоставленного объёма используются для принятия решения.

2. Нестабильность: верное решение задачи не гарантируется, так как результаты нейронной сети зависят от установленных параметров.

3. Сложность: большое число параметров осложняет подбор оптимальной архитектуры сети для решения конкретной задачи.

4. Техническая зависимость: процесс обучения нейронной сети может занимать достаточно много времени и зависит от технических характеристик компьютера.

1.3 Разведочный анализ данных

Вначале сделаем просмотр полученного нами датафрейма. Выведем первые пять строк:

РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

```
In [12]: df_pok.head()
```

Out[12]:

	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
0	111242.0	49.0	4.0	4.0	1854.0	52.2
1	4734.0	1.0	0.0	0.0	762.0	30.3
2	6147.0	2.0	1.0	0.0	450.0	25.2
3	4132.0	4.0	4.0	3.0	861.0	28.8
4	2676.0	3.0	1.0	0.0	120.0	28.6

Рисунок 10 – Показатели, характеризующие наличие ресурсов и результаты деятельности КФХ по первым пяти наблюдениям

Выведем информацию о изучаемом наборе данных:

```
In [18]: df_pok.info ()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1202 entries, 0 to 1201
Data columns (total 6 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Доходы, тыс.руб.         1202 non-null   float64
 1   Работники, чел.          1202 non-null   float64
 2   Наличие тракторов, шт.   1202 non-null   float64
 3   Наличие комбайнов, шт.   1202 non-null   float64
 4   Общая площадь земли, га  1202 non-null   float64
 5   Урожайность зерновых, га 1202 non-null   float64
dtypes: float64(6)
memory usage: 56.5 KB
```

Рисунок 11 – Информация о наборе данных

Из рисунка 11 видно, что у всех признаков тип данных – числа с плавающей точкой. Пропущенные значения отсутствуют, так как пропущенные значения были заменены на нули.

Приведем описательную статистику по изучаемому датафрейму (рисунок 12). По результатам вывода описательной статистики мы видим, что масштаб изучаемых признаков различен. Диапазон изменений значений признаков велик, так например, по величине доходов КФХ видно, что минимальные значения доходов составили 50 тыс. руб., а максимальные 396570 тыс. руб., тогда как средняя величина доходов составила 9871,5 тыс. руб., а медианная средняя 4153,0 тыс. руб. Такие различия в величине среднеарифметической и медианной средней говорят об асимметричности ряда распределения КФХ по доходам. Таким же образом можно проанализировать каждый признак в наборе данных.

```
In [16]: df_pok.describe ()
```

Out[16]:

	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
count	1202.0	1202.0	1202.0	1202.0	1202.0	1202.0
mean	9871.5	2.9	2.8	1.2	504.9	30.4
std	22013.3	4.7	3.0	1.8	1167.7	12.8
min	50.0	1.0	0.0	0.0	3.0	2.0
25%	1717.5	1.0	1.0	0.0	120.0	21.5
50%	4153.0	1.0	2.0	1.0	259.4	27.7
75%	9716.2	3.0	4.0	2.0	539.8	36.4
max	396570.0	76.0	31.0	22.0	23118.5	81.3

Рисунок 12 – Описательная статистика

В качестве целевой переменной (результативной) выбран показатель «Доходы, тыс. руб.», так как данный показатель отражает результаты деятельности крестьянских (фермерских) хозяйств.

В качестве факторных признаков выбраны показатели, характеризующие ресурсы производства – площадь земли, число всех работников, наличие сельскохозяйственной техники, а также показатель урожайности зерновых и зернобобовых, так как данные представлены зерновыми хозяйствами, и с помощью данного показателя можно также прогнозировать доходы КФХ.

Далее рассмотрим как связаны изучаемые факторные признаки с целевой переменной и как они связаны между собой. Наглядно можно увидеть характер зависимостей, построив попарные диаграммы рассеяния (рисунок 13). Из рисунка 13 видно, что направление связи между факторными признаками и результативным прямое (поле корреляции направлено вверх), но имеются выбросы. Визуально посмотрев на диаграммы рассеяния урожайности зерновых и зернобобовых и доходов КФХ, можно предположить, что сила связи между ними небольшая. В то же время можно заметить о заметной силе связи между самими факторами.

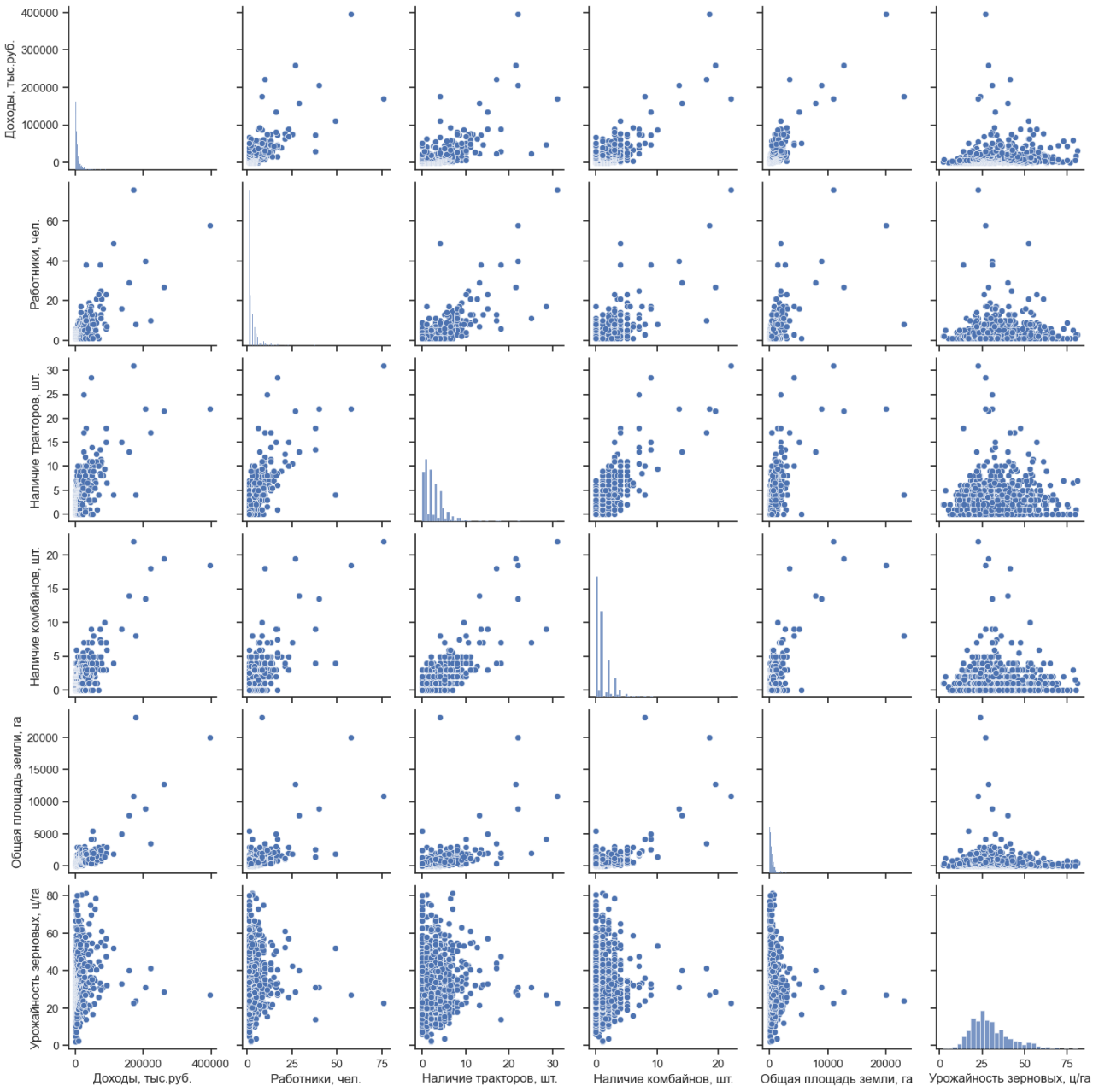


Рисунок 13 – Попарные диаграммы рассеяния

Рассмотрим гистограммы распределения крестьянских (фермерских) хозяйств по всем признакам (рисунок 14).

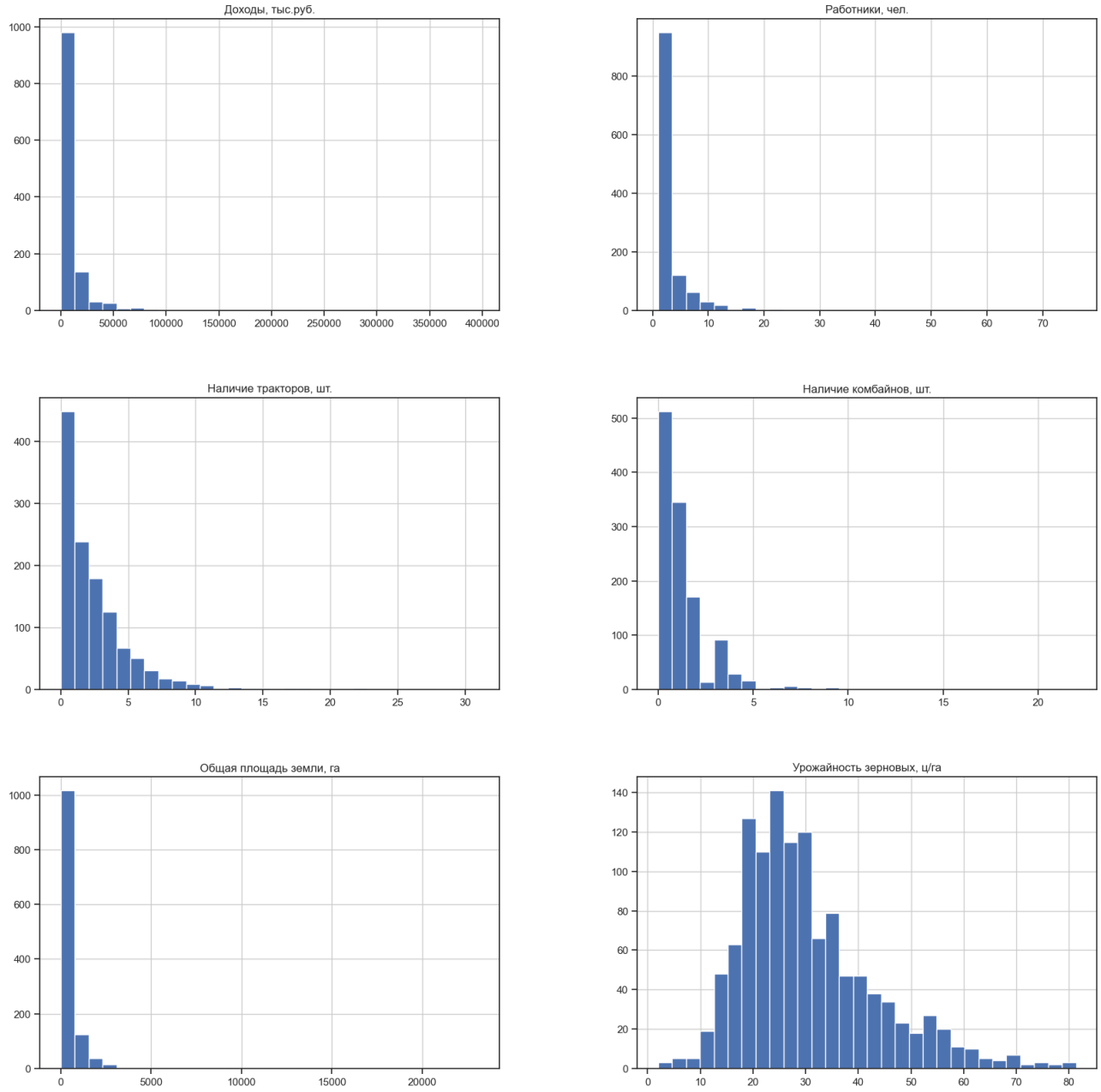


Рисунок 14 – Гистограммы по каждому признаку

Рассмотрим матрицу парных коэффициентов корреляции для определения силы связи между изучаемыми признаками. Признаки: численность работников, общая площадь земли и число тракторов и комбайнов довольно тесно связаны с результативным признаком (рисунок 15). Урожайность зерновых и зернобобовых с доходами КФХ связаны очень слабо. Поэтому включение данного факторного признака не приведет к хорошему качеству модели.

Out[20]:

	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
Доходы, тыс.руб.	1.000	0.763	0.655	0.769	0.829	0.183
Работники, чел.	0.763	1.000	0.700	0.692	0.634	0.110
Наличие тракторов, шт.	0.655	0.700	1.000	0.776	0.532	0.113
Наличие комбайнов, шт.	0.769	0.692	0.776	1.000	0.671	0.042
Общая площадь земли, га	0.829	0.634	0.532	0.671	1.000	-0.018
Урожайность зерновых, ц/га	0.183	0.110	0.113	0.042	-0.018	1.000

Рисунок 15 – Матрица парных коэффициентов корреляции

По тепловой карте удобно изучать силу связи между двумя признаками. Недостаток тепловой карты в том, что по ней трудно проследить тесные зависимости между тремя и более признаками одновременно. Для обнаружения мультиколлинеарности факторов проанализируем корреляционную матрицу факторов. Уже наличие больших по модулю (выше 0,7-0,8) значений коэффициентов парной корреляции свидетельствует о возможных проблемах с качеством получаемых оценок.

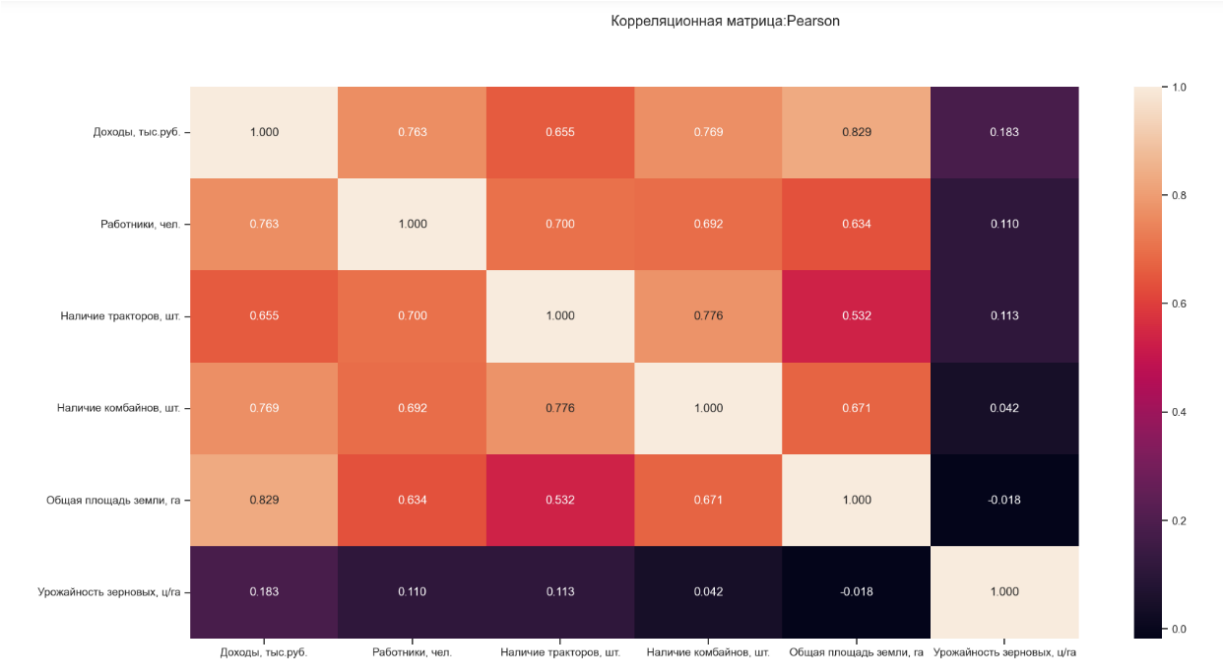


Рисунок 16 – Тепловая карта

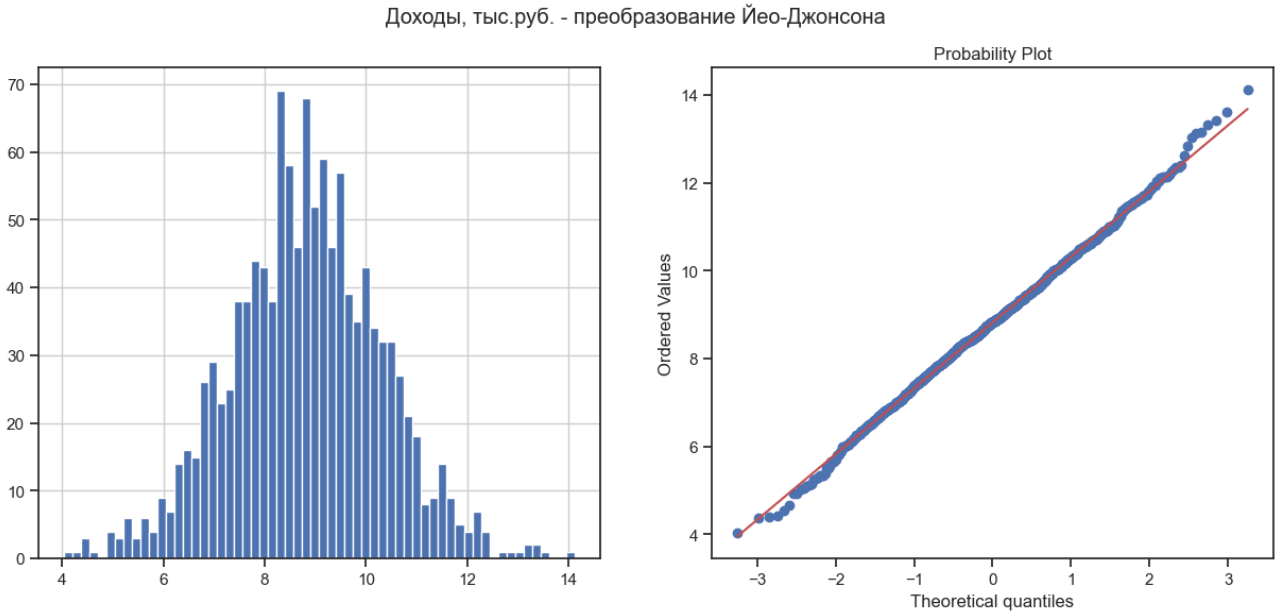
Имеется тесная связь между числом тракторов и комбайнов (r=0,776), числом тракторов и численностью работников (r=0,700). Но данная сила связи проявляется по исходному количеству наблюдений вместе с выбросами. Далее проведя предобработку данных будут рассмотрены новые значения парных коэффициентов корреляции без выбросов, имеющихся в совокупности.

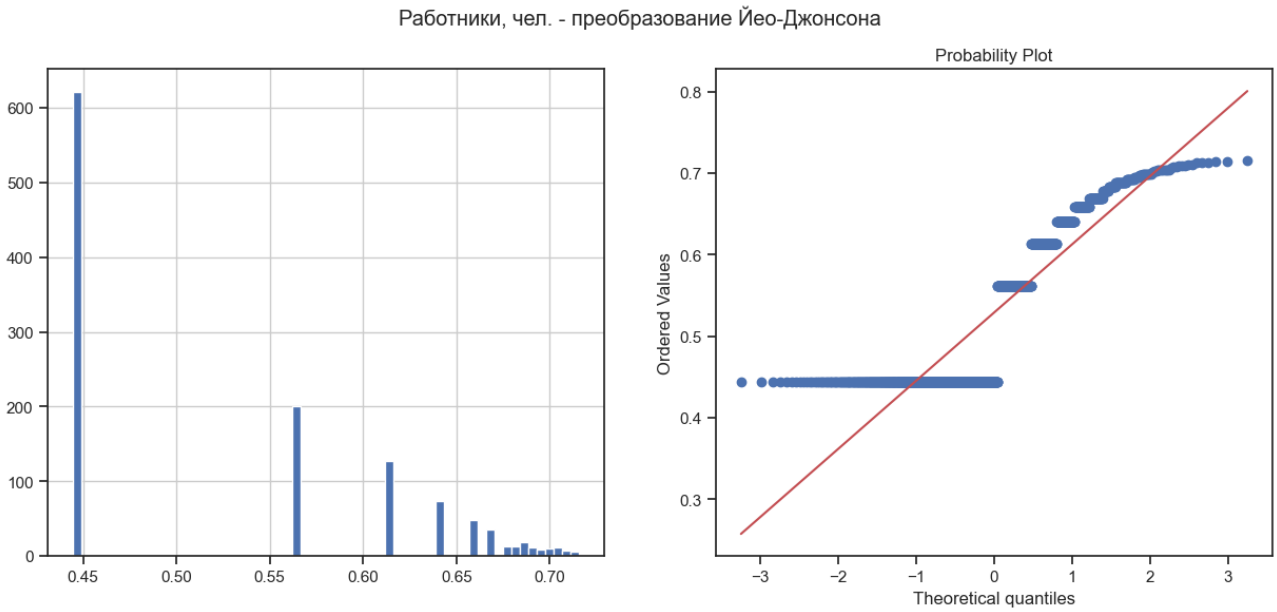
Глава 2 Практическая часть

2.1 Предобработка данных

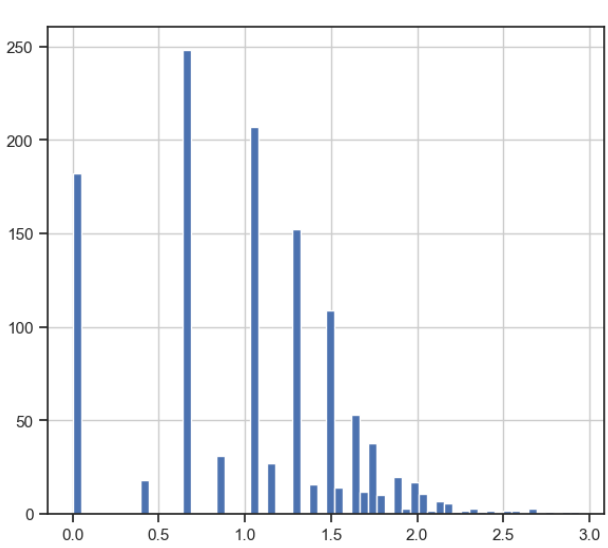
Нормирование данных

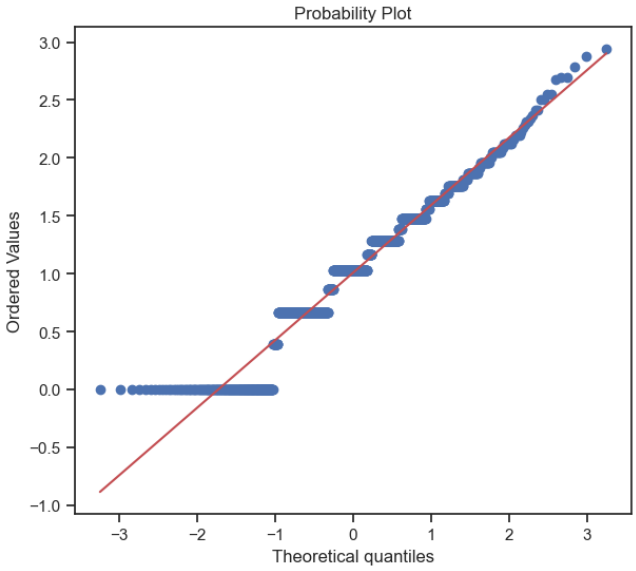
Нормирование данных произведено по преобразованию Йео-Джонсона. По другим методам преобразования приведение к нормальному распределению (логарифмическое преобразование, обратное преобразование, возведение в степень, преобразование по корню квадратному) показало худшие результаты (Приложение А). Преобразование Бокса-Кокса требовало отсутствие нулевых значений.



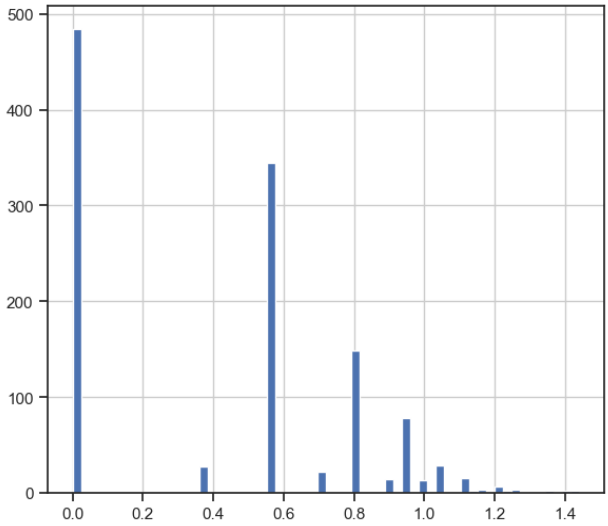


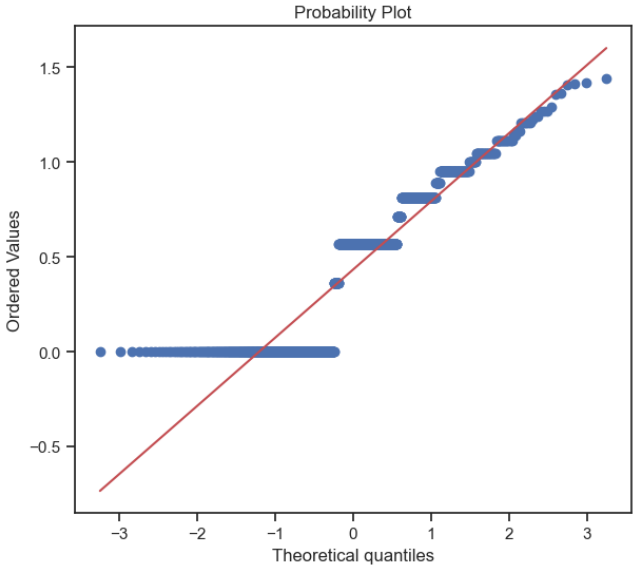
Наличие тракторов, шт. - преобразование Йео-Джонсона



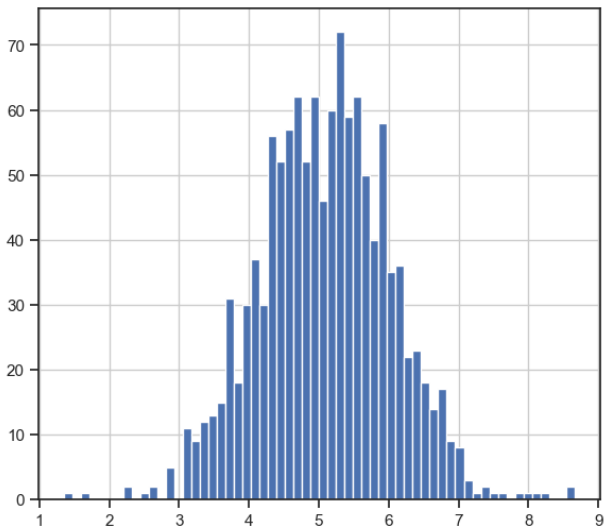


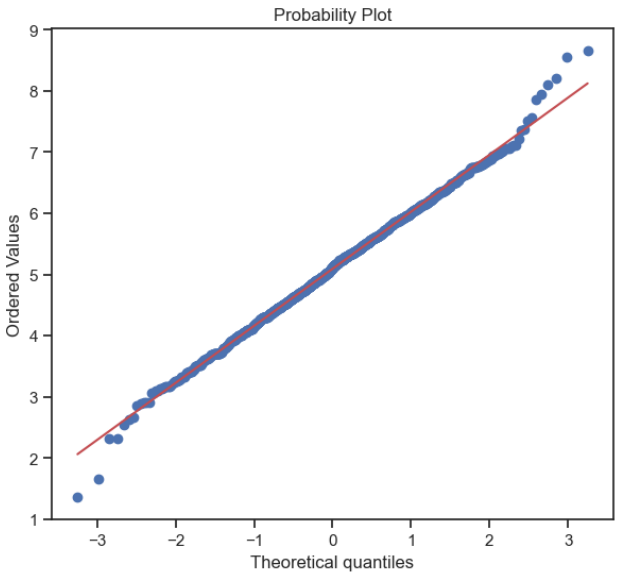
Наличие комбайнов, шт. - преобразование Йео-Джонсона





Общая площадь земли, га - преобразование Йео-Джонсона





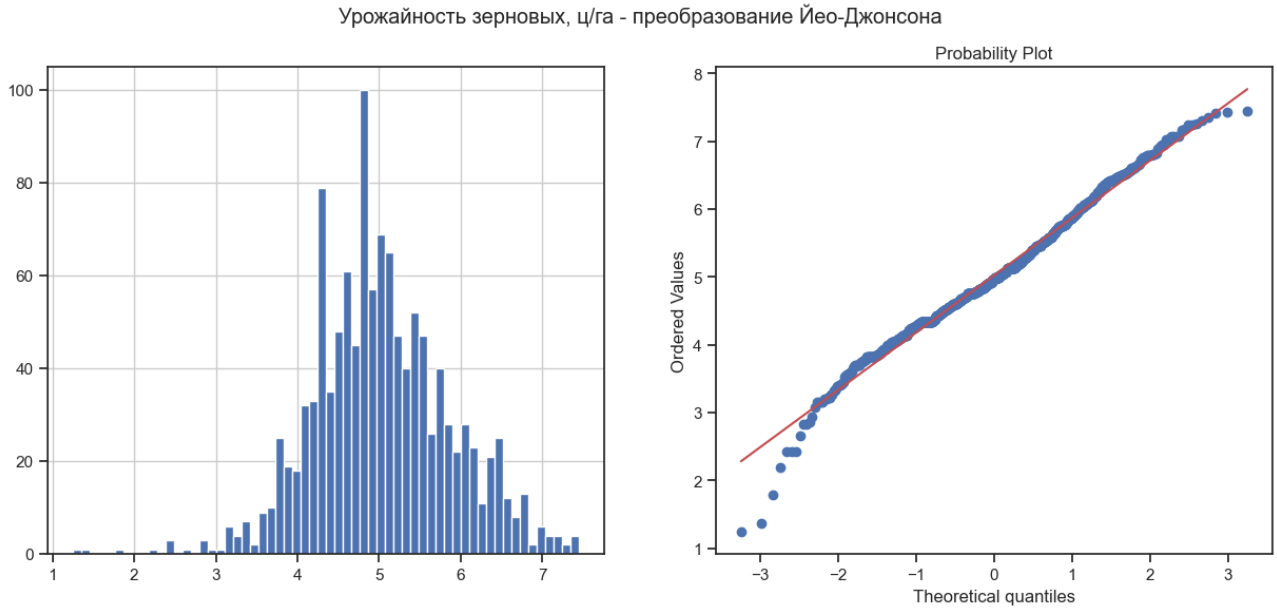


Рисунок 17 – Нормирование данных по преобразованию Йео-Джонсона

На рисунке 18 приведены данные, преобразованные по Йео-Джонсону:

```
In [56]: df_pok_norm.head()
```

Out[56]:

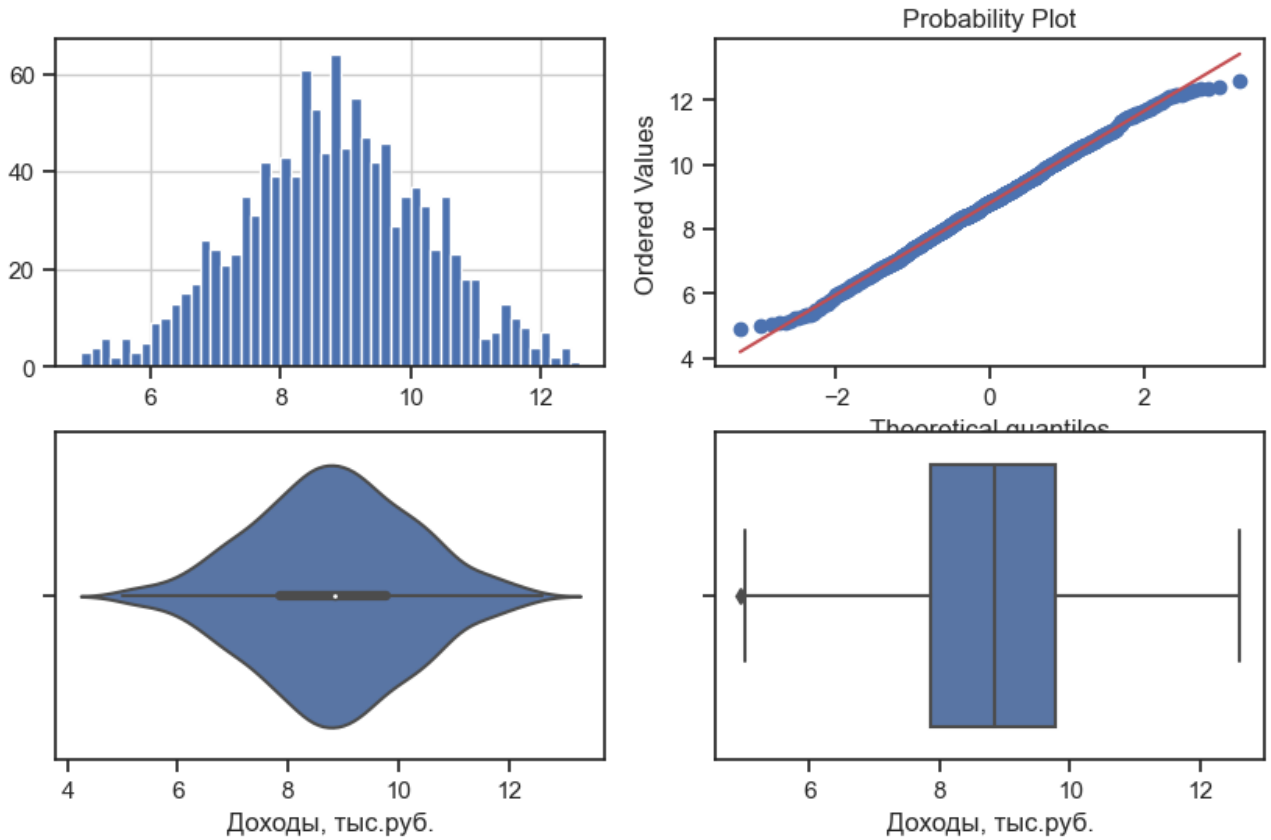
	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
0	12.608	0.714	1.541	1.044	6.725	6.350
1	8.979	0.444	-0.000	-0.000	6.009	5.152
2	9.273	0.562	0.680	-0.000	5.576	4.781
3	8.826	0.641	1.541	0.951	6.108	5.050
4	8.340	0.614	0.680	-0.000	4.462	5.034

Рисунок 18 – Вывод нормированных данных по всем признакам по первым пяти наблюдениям

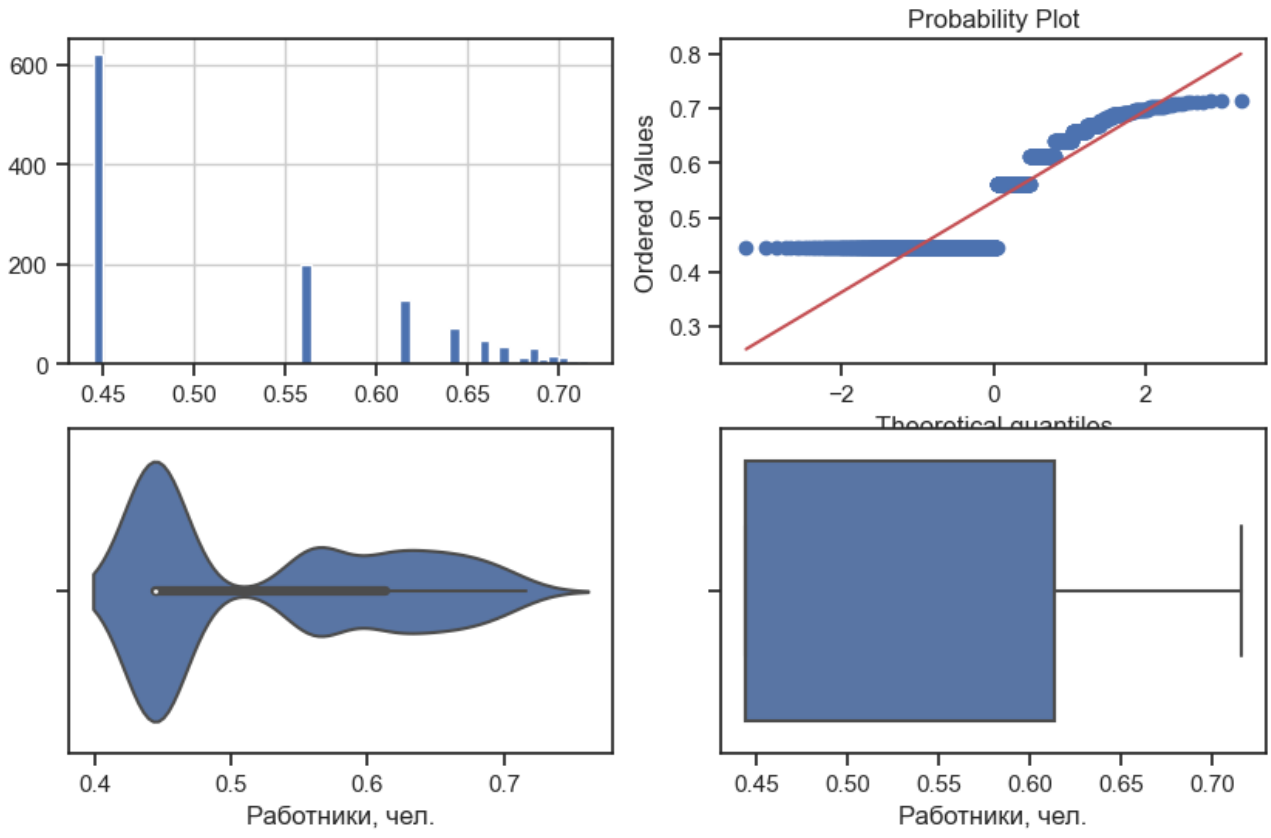
Обработка выбросов

Для удаления выбросов был выбран метод по межквартильному размаху, так как наблюдается скошенность в рядах распределения. Другие методы удаления выбросов показаны в приложении Б. На рисунке представлены графики «ящики с усами» после удаления выбросов. Можно сравнить с рисунком 2, в котором приведены Выбросы удалялись с использованием межквартильного размаха графики «ящики с усами» до удаления выбросов. Сравнив, два рисунка, можно заметить, что выбросы были удалены.

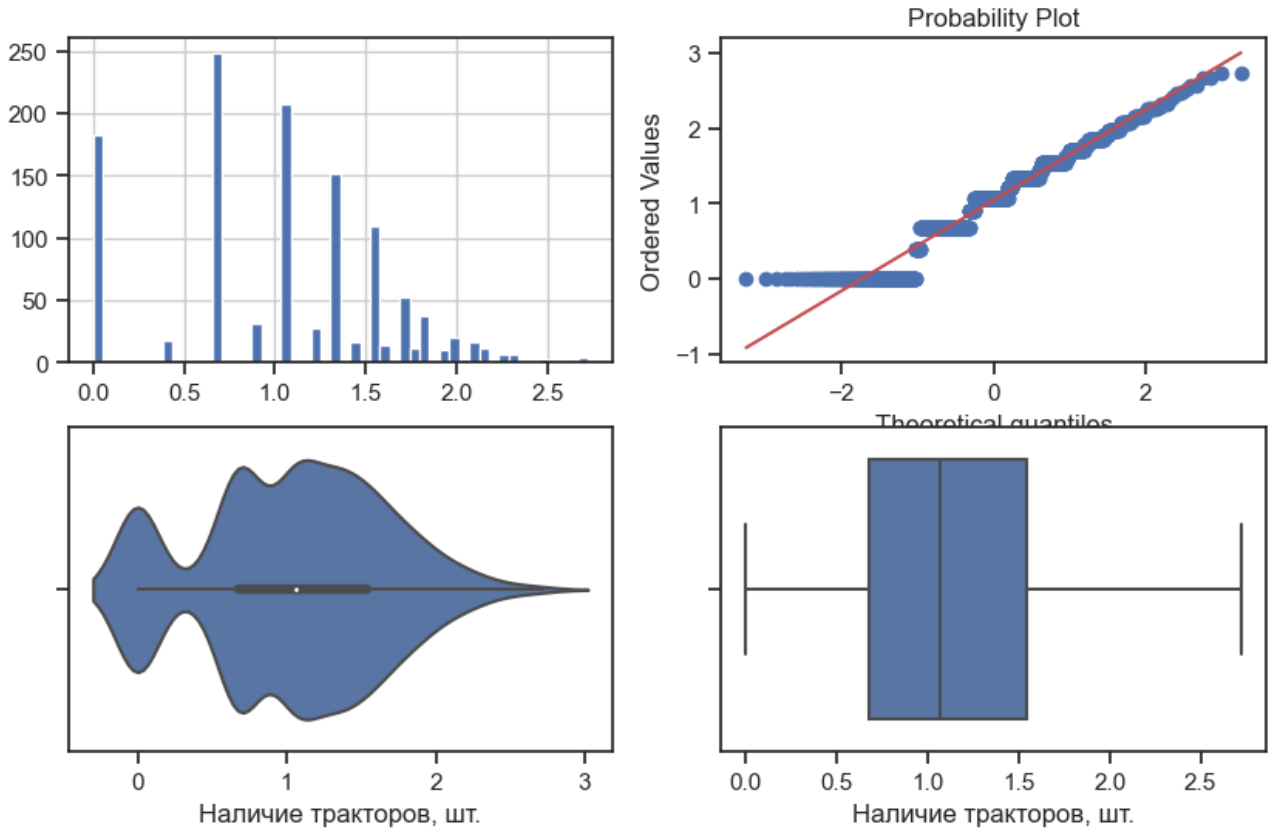
Поле-Доходы, тыс.руб., метод-OutlierBoundaryType.IQR, строк-1181



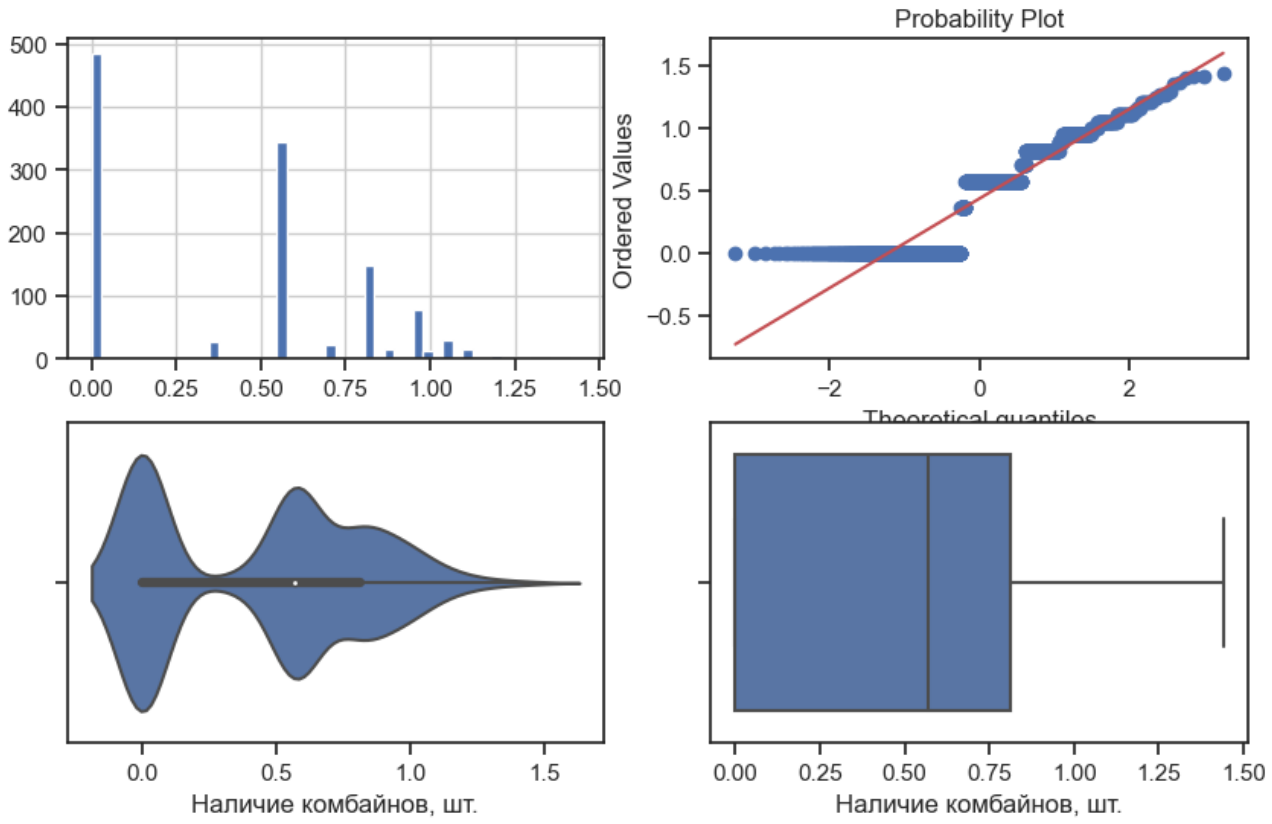
Поле-Работники, чел., метод-OutlierBoundaryType.IQR, строк-1181



Поле-Наличие тракторов, шт., метод-OutlierBoundaryType.IQR, строк-1181



Поле-Наличие комбайнов, шт., метод-OutlierBoundaryType.IQR, строк-1181



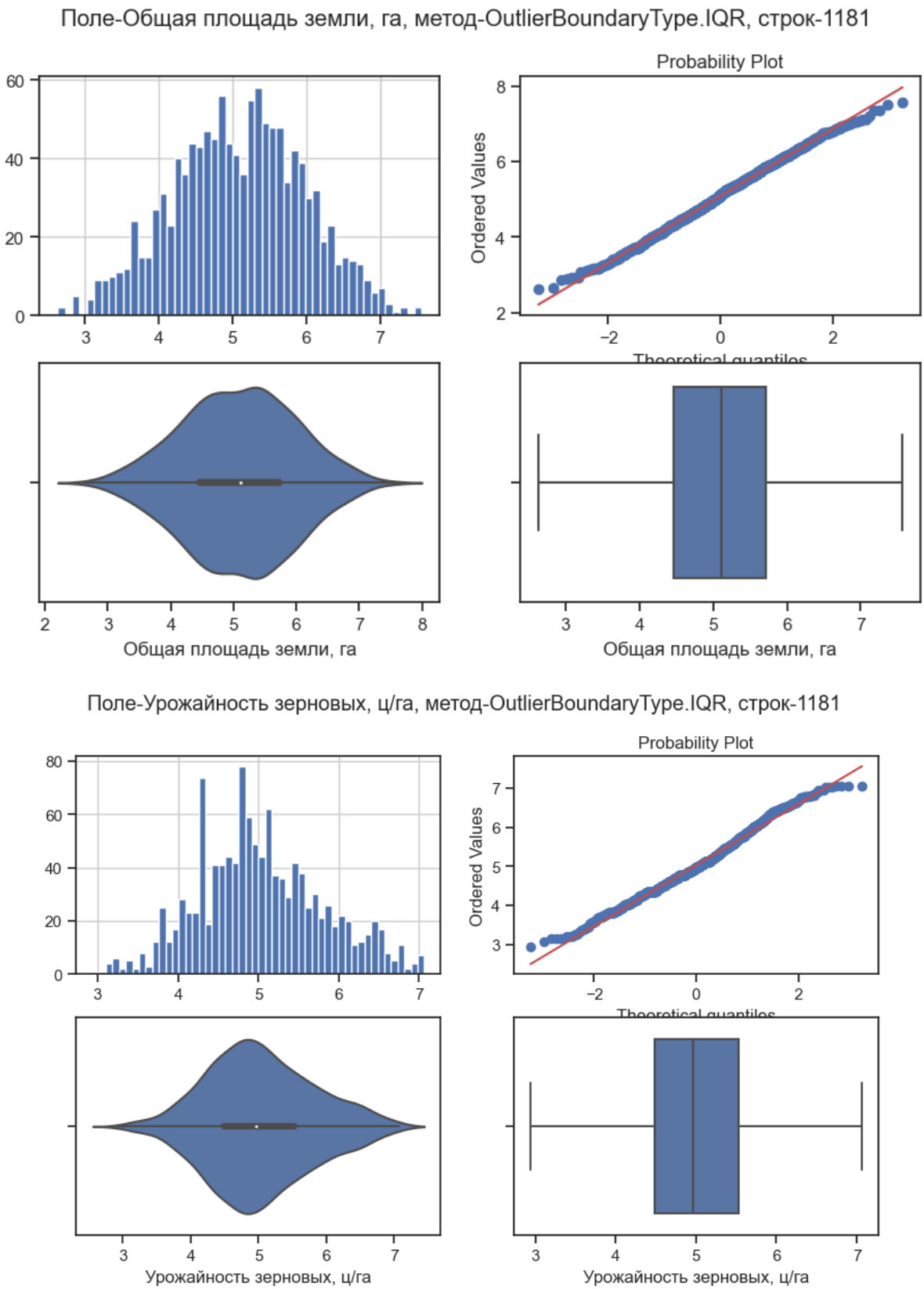


Рисунок 17 – Гистограмма, график Q-Q plot, график «ящик с усами», скрипичная диаграмма по нормированным данным после удаления выбросов

В таблице приведена описательная статистика данных после преобразования по Йео-Джонсону после удаления выбросов.

Out[80]:

	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
count	1181.000	1181.000	1181.000	1181.000	1181.000	1181.000
mean	8.826	0.530	1.058	0.437	5.101	5.032
std	1.477	0.094	0.626	0.391	0.924	0.783
min	4.361	0.444	-0.000	-0.000	1.357	2.938
25%	7.862	0.444	0.680	0.000	4.469	4.488
50%	8.841	0.444	1.066	0.570	5.120	4.968
75%	9.792	0.614	1.541	0.811	5.726	5.532
max	14.114	0.716	3.159	1.440	8.655	7.067

Рисунок 18 – Описательная статистика датафрейма по нормированным данным после удаления выбросов

Масштабирование данных производилось с использованием Z-оценки, где среднее значение равно 0, а дисперсия 1. Метод реализован с использованием класса StandardScaler. Результаты стандартизации приведены на рисунке 19.

```
In [85]: df_pok_scaled.describe()
```

Out[85]:

	Доходы, тыс.руб.	Работники, чел.	Наличие тракторов, шт.	Наличие комбайнов, шт.	Общая площадь земли, га	Урожайность зерновых, ц/га
count	1181.000	1181.000	1181.000	1181.000	1181.000	1181.000
mean	0.000	-0.000	0.000	0.000	-0.000	0.000
std	1.000	1.000	1.000	1.000	1.000	1.000
min	-3.025	-0.904	-1.692	-1.117	-4.054	-2.675
25%	-0.653	-0.904	-0.604	-1.117	-0.685	-0.695
50%	0.010	-0.904	0.013	0.342	0.020	-0.082
75%	0.655	0.892	0.772	0.959	0.677	0.640
max	3.583	1.978	3.359	2.569	3.848	2.600

Рисунок 19 – Описательная статистика датафрейма по стандартизованным данным

Далее рассмотрены плотности распределения хозяйств по изучаемым признакам до и после масштабирования.

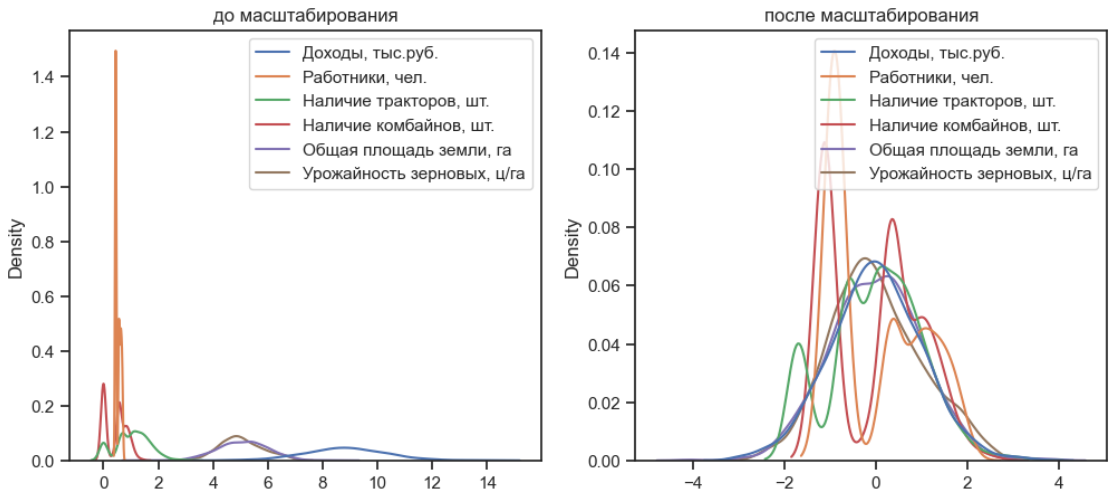


Рисунок 20 – Плотности распределения хозяйств по изучаемым признакам до и после масштабирования.

Отбор факторов

Отбор факторов произведен на основе матрицы парных коэффициентов корреляции по данным после удаления выбросов. Так как связь между доходами КФХ и урожайностью зерновых близкая к слабой, то последний признак решено не добавлять в модель регрессии. Два факторных признака: наличие тракторов и комбайнов - тесно связаны между собой (r=0,692) – получено значение, близкое к 0,7. Это говорит о наличии мультиколлинеарности, поэтому один их них нужно исключить из модели. Решено удалить число комбайнов, так как сила связи между числом комбайнов и доходами слабее, и хозяйств, имеющих комбайны, встречается в совокупности гораздо реже, нежели хозяйств, не имеющих трактора.

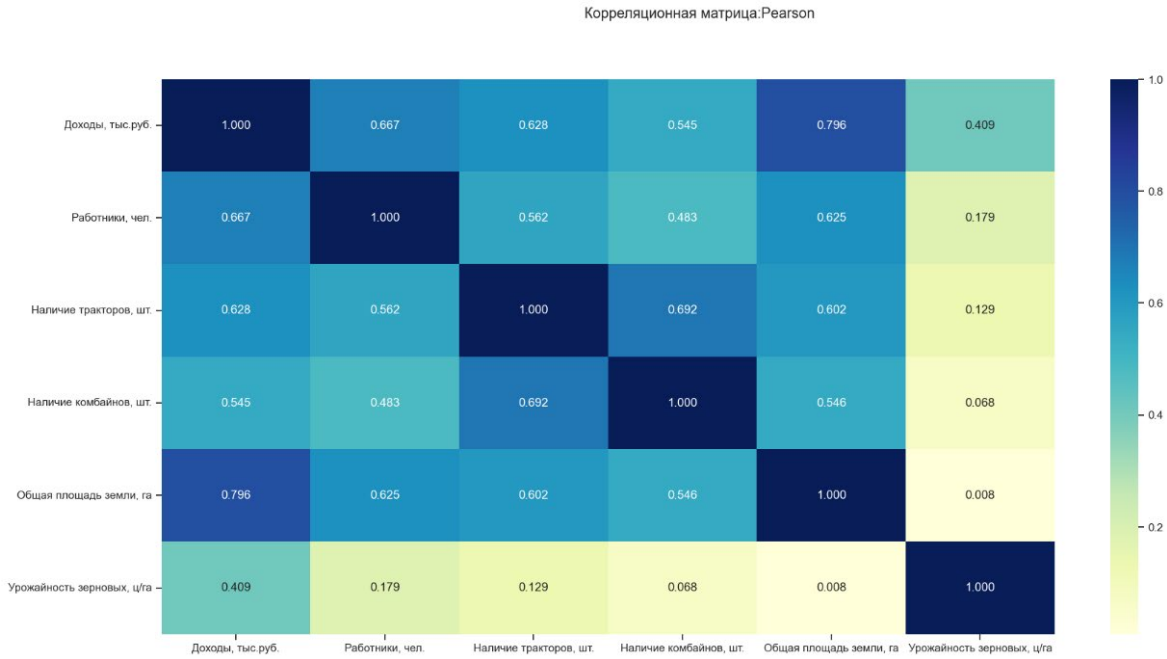


Рисунок 21 – Тепловая карта

Также построена диаграмма на основе взаимной информации, показывающей силу связи каждого факторного признака с результативным, где 2 – это площадь земли, 0- количество работников и 1 – число тракторов.

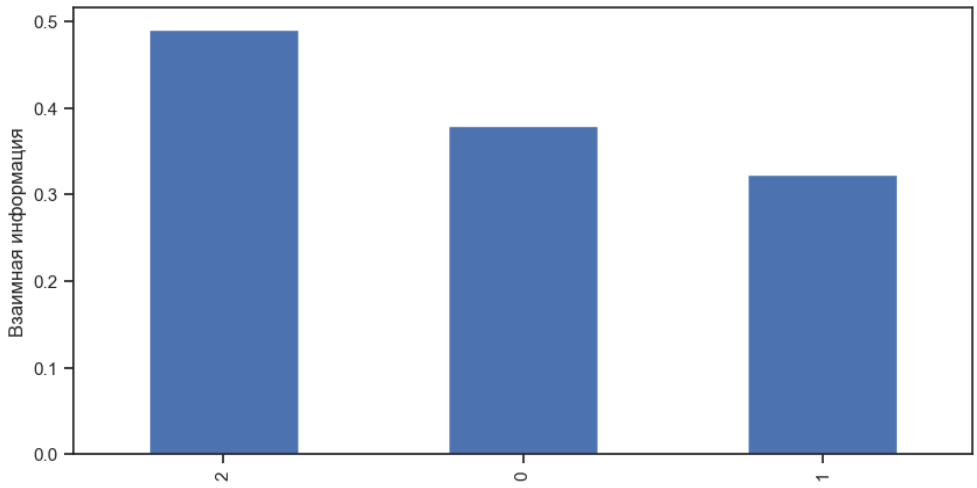


Рисунок 22 – Взаимная информация

2.2 Разработка и обучение модели

При построении моделей различными методами совокупность наблюдений разбита на две части: 30% наблюдений приходится на тестирование моделей, 70% - на обучение моделей.

```
In [112]: # Разделим выборку на обучающую и тестовую
          X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=1)
```

Поиск гиперпараметров моделей производился с помощью поиска по сетке с перекрестной проверкой (GridSearchCV), количество блоков выбрано10.

Для прогнозирования доходов КФХ было обучено несколько моделей машинного обучения.

Вначале реализована множественная линейная регрессия, построенная с помощью библиотек statsmodels и sklearn. Получены следующие результаты по методу наименьших квадратов (рисунок 23).

Скорректированный коэффициент детерминации R^2=0,706 показывает, что 70,6% вариации доходов КФХ объясняется изменением трех факторов, включенных в модель, а остальные 29,4% - это влияние других неучтенных факторов. Модель в целом статистически значима. Все параметры, кроме условного начала оказались статистически значимыми.

Out[121]:

OLS Regression Results

Dep. Variable:	Доходы, тыс.руб.	**R-squared:**	0.707
Model:	OLS	**Adj. R-squared:**	0.706
Method:	Least Squares	**F-statistic:**	660.2
Date:	Tue, 25 Apr 2023	**Prob (F-statistic):**	2.25e-218
Time:	19:06:49	**Log-Likelihood:**	-671.64
No. Observations:	826	**AIC:**	1351.
Df Residuals:	822	**BIC:**	1370.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0154	0.019	0.807	0.420	-0.022	0.053
Работники, чел.	0.2216	0.026	8.599	0.000	0.171	0.272
Наличие тракторов, шт.	0.1627	0.024	6.661	0.000	0.115	0.211
Общая площадь земли, га	0.5706	0.027	21.203	0.000	0.518	0.623

Omnibus:	11.146	**Durbin-Watson:**	2.113
Prob(Omnibus):	0.004	**Jarque-Bera (JB):**	12.636
Skew:	-0.206	**Prob(JB):**	0.00180
Kurtosis:	3.445	**Cond. No.**	2.52

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Рисунок 23 – Результаты регрессии

Далее реализованы ридж и лассо-регрессии, которые показали одинаковый результат по качеству модели с множественной линейной регрессий по обычному МНК. Поиск гиперпараметров производился следующим образом (рисунок 24):

РИДЖ-РЕГРЕССИЯ

In [127]:
```
ridge = Ridge()
param_grid_ridge = [{'alpha': [1e-15, 1e-10, 1e-8, 1e-4, 1e-3, 1e-2, 1e-1, 0, 1, 5, 10 ],
                     'solver': ['svd', 'cholesky', 'lsqr', 'sparse_cg', 'sag', 'saga']}]
GSCV_ridge = GridSearchCV(estimator=ridge, param_grid=param_grid_ridge, cv=10, verbose=2)
GSCV_ridge.fit(X_train, y_train)
GSCV_ridge.best_params_
```

ЛАССО-РЕГРЕССИЯ

In [132]:
```
lasso = Lasso()
param_grid_lasso = { 'alpha': [1e-15, 1e-10, 1e-8, 1e-4, 1e-3, 1e-2, 1e-1, 0, 1, 5, 10 ]}
GSCV_lasso = GridSearchCV(estimator=lasso, param_grid=param_grid_lasso, cv=10, verbose=2)
GSCV_lasso.fit(X_train, y_train)
GSCV_lasso.best_params
```

Рисунок 24 – GSCV для ридж-регрессии и лассо-регрессии

Оптимальное значение гиперпараметра для альфа по ридж-регрессии - единица, а по лассо-регрессии – 0,001.

Далее применен метод K-ближайших соседей. Оптимальное значение 'n_neighbors': 13. Качество модели получилось несколько хуже по сравнению с множественной линейной регрессией по МНК.

МЕТОД К-БЛИЖАЙШИХ СОСЕДЕЙ

```
In [137]: knn = KNeighborsRegressor()
          param_grid = {'n_neighbors': [1, 2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20]}
          GSCV_knn = GridSearchCV(estimator=knn, param_grid=param_grid, cv=10, verbose=2)
          GSCV_knn.fit(X_train, y_train)
          GSCV_knn.best_params_
```

Рисунок 25 – GSCV для K-ближайших соседей

Затем построено дерево решений. Произведен поиск гиперпарметров. Оптимальное значение гиперпараметров: {'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 2}. Схема дерева решений приведена в Приложении В. модели Дерево решений по качеству модели получено самой низкой.

ДЕРЕВО РЕШЕНИЙ

```
In [142]: decision_tree = DecisionTreeRegressor (random_state = 42)
          param_grid = { 'max_features': ['auto', 'sqrt', 'log2'],
                         'max_depth' : [3,4,5,6],
                         'min_samples_leaf': [1,2,3] }
          GSCV_dt = GridSearchCV(estimator=decision_tree, param_grid=param_grid, cv=10, verbose=2)
          GSCV_dt.fit(X_train, y_train)
          GSCV_dt.best_params_
```

Рисунок 26 – GSCV для Дерева решений

Далее построена модель случайного леса.

СЛУЧАЙНЫЙ ЛЕС

```
In [149]: random_forest = RandomForestRegressor(random_state = 42)
          param_grid = { 'n_estimators': [100, 200, 300, 400, 500],
                         'max_features': ['auto', 'sqrt'],
                         'max_depth' : [4,5,6,7,8],
                         'criterion' :['squared_error']}
          GSCV_rf = GridSearchCV(estimator=random_forest, param_grid=param_grid, cv=10, verbose=2)
          GSCV_rf.fit(X_train, y_train)
          GSCV_rf.best_params_
```

Рисунок 27 – GSCV для Случайного леса

Оптимальные гиперпараметры для случайного леса: {'criterion': 'squared_error', 'max_depth': 5, 'max_features': 'auto', 'n_estimators': 500}. Тестовые и прогнозные значения по тестовым данным показаны на графике в приложении Г. Качество модели по случайному лесу чуть лучше, чем по дереву решений и по множественной линейной регрессией по МНК.

Градиентный бустинг показал хорошие результаты качества модели, как случайный лес.

2.3 Тестирование модели

Качество построенных моделей определялось с помощью следующих метрик: коэффициент детерминации, средняя квадратическая ошибка и средняя абсолютная ошибка. Ошибки каждой модели на тренировочной и тестирующей части выборки показаны на рисунке 28.

Рисунок 28 – Сравнение моделей по MAE, MSE и R^2 по обучающей и тестовой выборкам

По тестовой выборке самое лучшее качество показали модели, построенные по нейронным сетям, хотя различия с моделями случайного леса и градиентного бустинга незначительны. Наименьшие ошибки MAE и MSE у моделей нейронных сетей.

2.4 Нейронная сеть

В работе выбрано построение полносвязной нейронной сети. Построение нейронной сети было проведено с помощью библиотек sklearn (MLPRegressor) и tensorflow (keras.Sequential).

Поскольку исходные данные были ранее нормализованы в процессе предобработки, то на вход нейросеть поданы нормализованные значения от 0 до 1. Поэтому дополнительная нормализация данных не проводилась.

При построении нейронной сети MLPRegressor был произведен поиск по сетке, где было предложено оптимальное число нейронов на каждом слое.

При построении нейронной сети с помощью tensorflow (keras.Sequential) было использовано два полносвязных скрытых слоя Dense с различным количеством нейронов, без дополнительного слоя дропаут и с добавлением дропаут, с функциями активации relu, tanh, sigmoid. В процессе обучения была предпринята попытка минимизировать потери функции, параметры обновлялись для повышения точности.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_6 (Dense)             (None, 90)                360

 dense_7 (Dense)             (None, 80)                7280

 dense_8 (Dense)             (None, 1)                 81

=================================================================
Total params: 7,721
Trainable params: 7,721
Non-trainable params: 0
_________________________________________________________________
```

Рисунок 29 – Архитектура нейронной сети

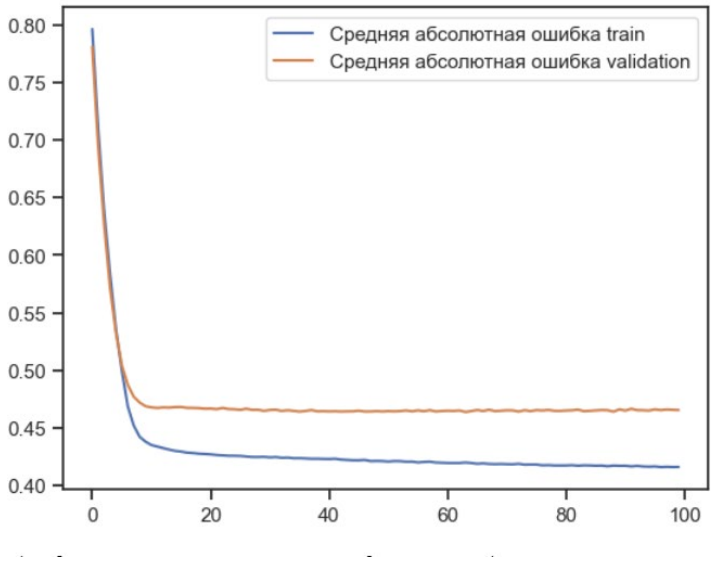


Рисунок 30 – График функции потерь

Таким образом, можно сделать вывод, что по ошибкам MAE, MSE и коэффициенту детерминации R^2 нейросеть показала лучший результат, чем любая из моделей регрессии.

2.5 Разработка приложения

С помощью фреймворка Flask было разработано одностраничное пользовательское веб-приложение, прогнозирующее доходы крестьянских (фермерских) хозяйств на основе нейронной сети.

Для запуска приложения пользователь должен перейти по ссылке на сайт: http://127.0.0.1:5000/ .

Flask-приложение представляет собой форму, состоящую из трех входов, куда вводятся значения трех параметров: численность работников, чел., число тракторов, шт., площадь земли, га. Введенные значения должны быть больше или равны 0, в противном случае появится ошибка «ОШИБКА! Введенные значения должны быть больше или равны 0».

После этого нужно нажать на кнопку «Submit», и модель выдаст прогнозное значение доходов КФХ при заданных параметрах.

2.6 Создание удаленного репозитория и загрузка результатов работы на него

На github.com был создан репозиторий: https://github.com/BayarmaDashieva/KFH

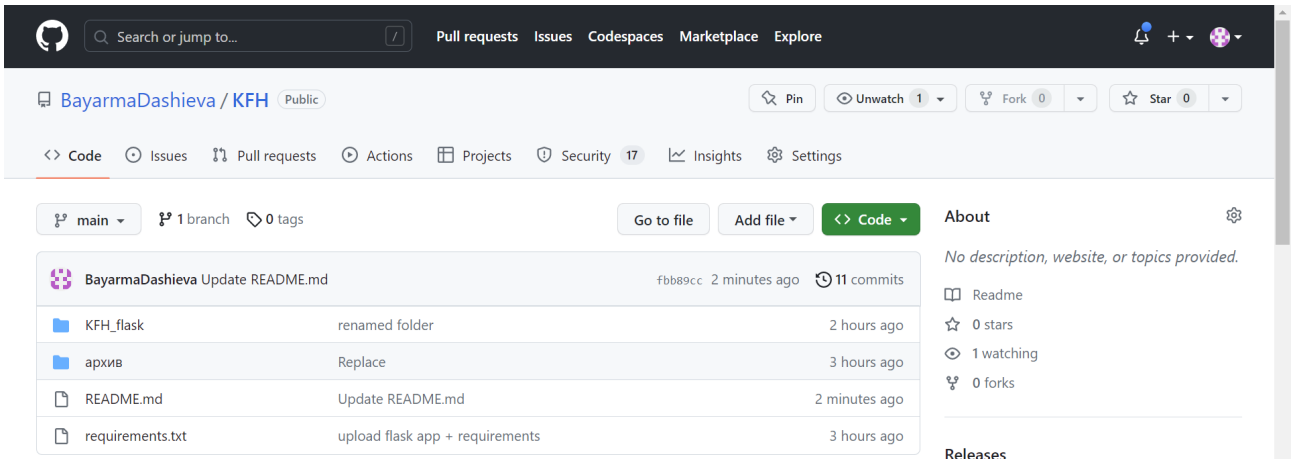


Рисунок 31 – Репозиторий на Github.com

Заключение

В первой главе выпускной квалификационной работы «Аналитическая часть» изучены теоретические основы, приведена характеристика датасета, проведен разведочный анализ данных.

Во второй главе работы «Практическая часть» проведена преодобработка данных, разработка и обучение различных моделей, построение нейронной сети.

По итогам работы разработано приложение с графическим интерфейсом, выдающая прогноз доходов КФХ.

На веб-сервисе Github.com создан репозиторий, где размещены все материалы выпускной квалификационной работы

Таким образом, все поставленные задачи в работе решены, цель достигнута.

По итогам выпускной квалификационной работы можно сделать следующие выводы: построены модели регрессии с использованием различных методов, в результате которых получено, что наилучшие результаты по качетсу модели показывают нейронные сети.

Разработанное веб-приложение могут применять крестьянские (фермерские) хозяйства для прогнозирования их доходов в зависимости от таких существенных факторов, как численность работников, площадь земли и наличие тракторов. А также построенные модели регрессии могут быть использованы при разработке мер аграрной политики для развития малого предпринимательства.

Библиографический список

1. Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).
2. Бринк Х. Машинное обучение / Х. Бринк, Дж. Ричардс, М. Феверолф. — пер. с англ. Рузмайкина И. — Санкт-Петербург: Питер, 2017. — 336 с.
3. Брюс, П., Брюс. Э. Разведочный анализ данных // Практическая статистика для специалистов Data Science. — СПб.: БХВ-Петербург, 2018. — С. 19—58. — 304 с.
4. Воронцов К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О.Б.Лупанова. 2004. Вып. 13 С. 5 – 36.
5. Воронина В. В., Михеев А. В., Ярушкина Н. Г., Святов К. В. // Теория и практика машинного обучения : учебное пособие /В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. –Ульяновск : УлГТУ, 2017. – 291 с.
6. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей / А.Н. Горбань // Сиб. журн. вычисл. математики. – 1998. – Т. 1, № 1. – 21 с.
7. Грас, Джоэл. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил
8. Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. Пер. с англ. - СпБ.: ООО "Альфа-книга': 2018. - 688 с.: ил
9. Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. СПб. : Питер, 2023. – 576 с.
10. Прикладной системный анализ : учебное пособие / Ф.П. Тарасенко. — М. : КНОРУС, 2010. — 224 с.
11. Рассел С., Норвиг П. Искусственный интеллект: современный подход, 2-е изд.. : Пер. с англ. - М. : Издательский дом ‘‘Вильямс’’, 2007. - 1408 с.
12. Статистическая обработка данных, планирование эксперимента и случайные процессы : учебное пособие для вузов / Берикашвили В. Ш., Оськин С. П. - 2-е изд., испр. и доп. - М. : Юрайт, 2021. - 163 с.
13. Репозиторий курсов по машинному обучению. COURSE_TMO, MMO/ Ugapanyuk. – URL: https://github.com/ugapanyuk/ml_course_2022
14. Шитиков В.К., Мастицкий С.Э. (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 351 с. − Электронная книга,

адрес доступа: https://github.com/ranalytics/data-mining

15. Box G.E.P., Jenkins G.M., Reinsel G.C., Ljung G.M. Time Series Analysis: Forecasting and Control - 5th Edition. — Wiley, 2015. — 712 p. — ISBN: 978-1-118-67502-1.

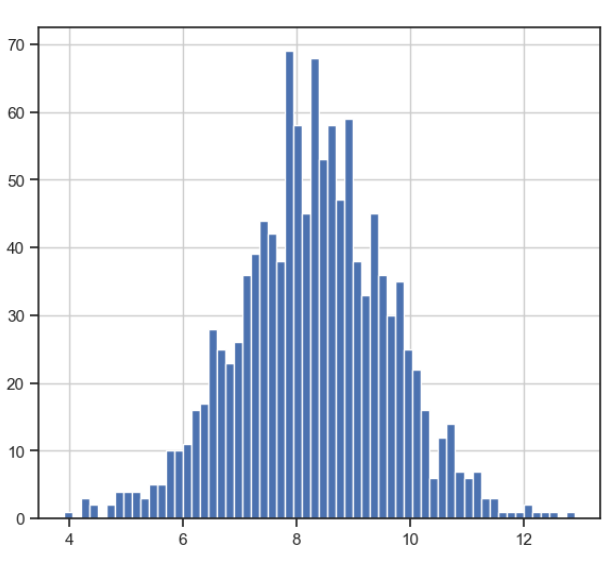
16. Nisbet R., Elder J., Miner G. Handbook of Statistical Analysis and Data Mining Applications. - Academic Press, 2009. — 864 p. — ISBN: 0123747651

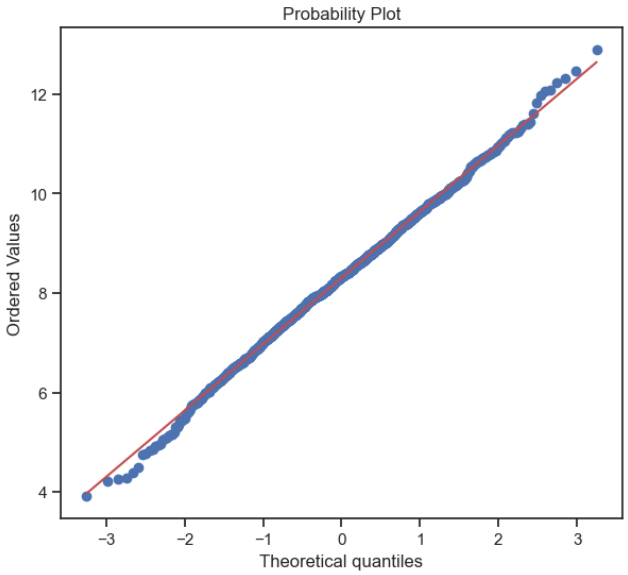
Приложения

Приложение А

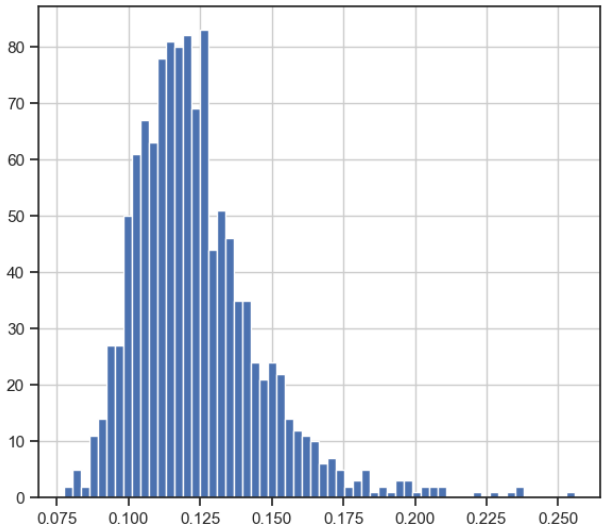
Нормирование данных

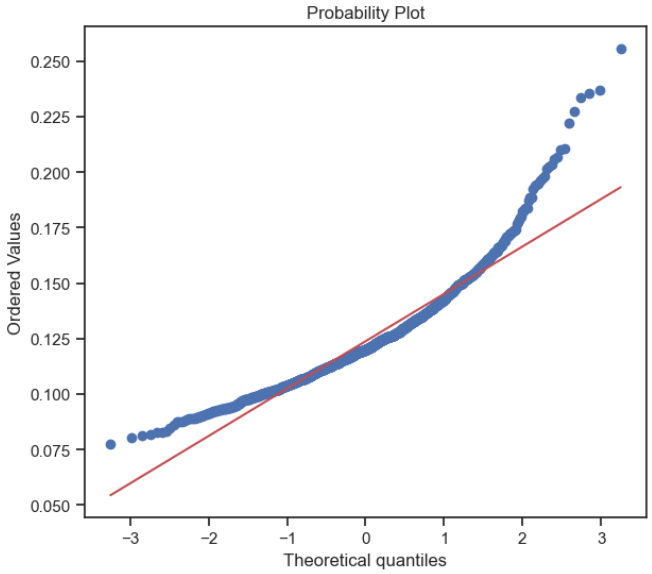
Доходы, тыс.руб. логарифмическое преобразование



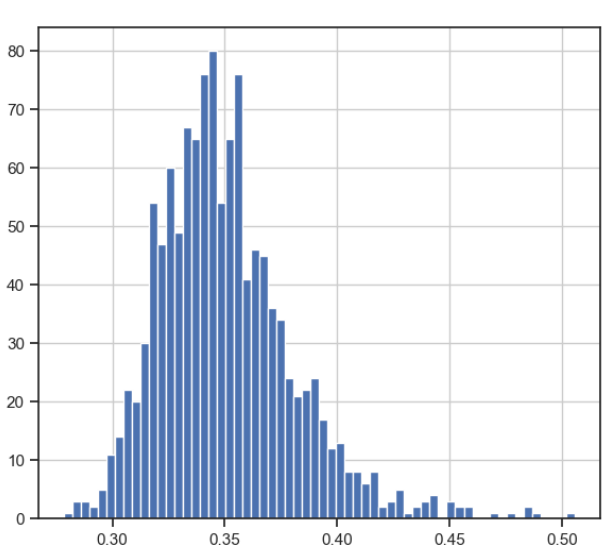


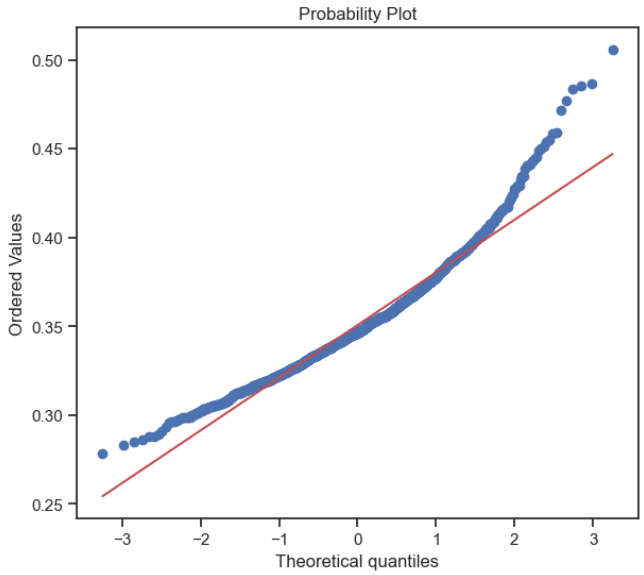
Доходы, тыс.руб. - обратное преобразование



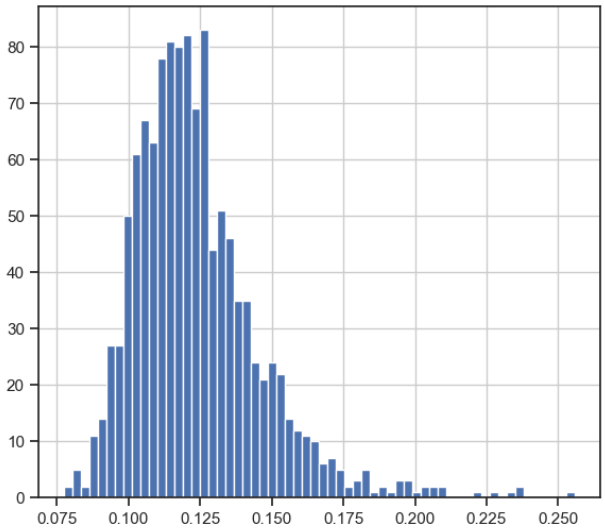


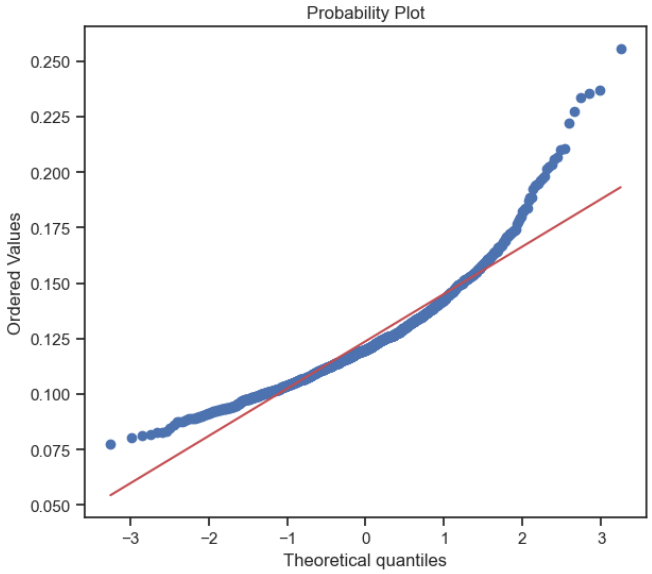
Доходы, тыс.руб. - корень квадратный



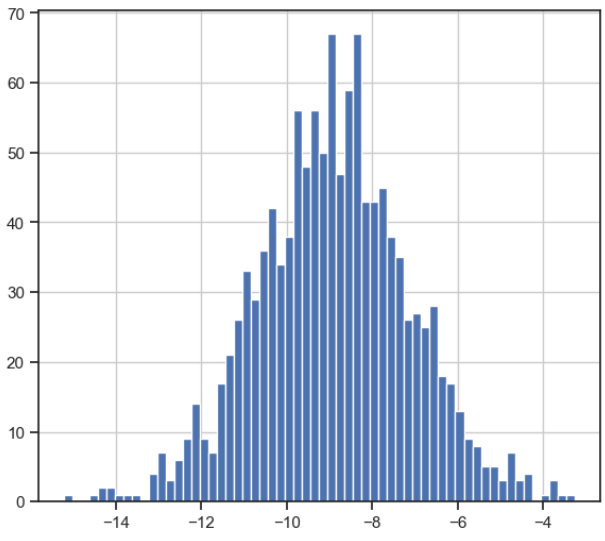


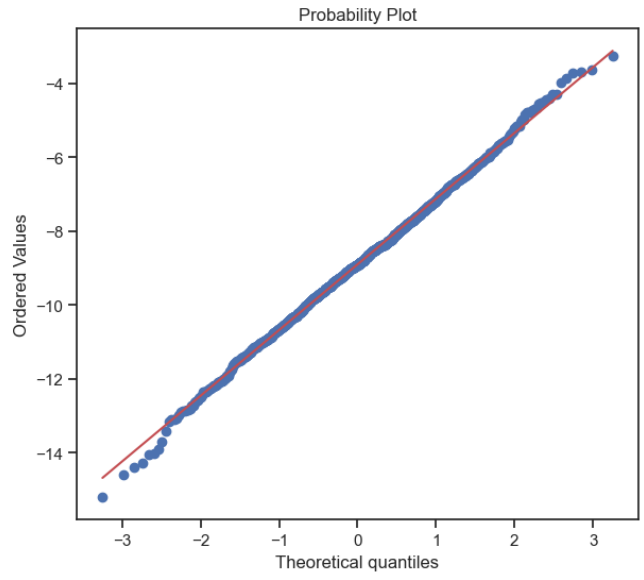
Доходы, тыс.руб. - возведение в степень





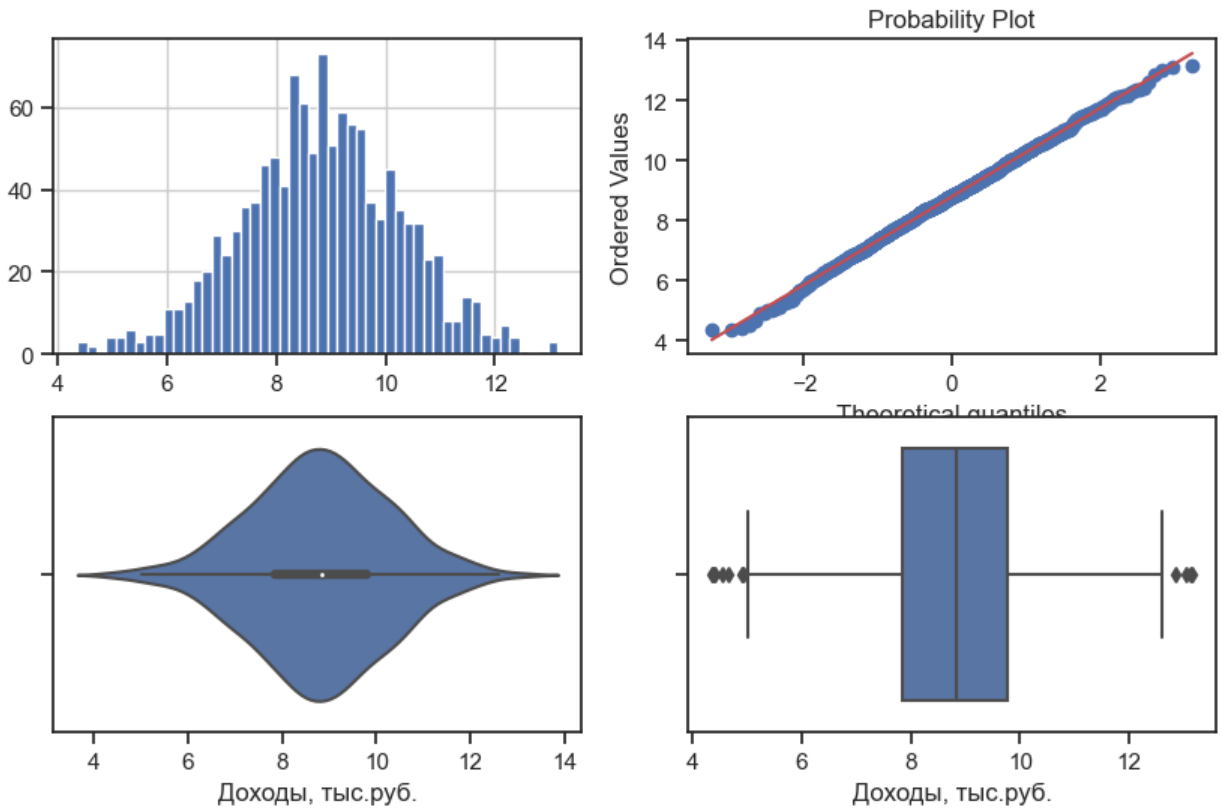
Доходы, тыс.руб. - преобразование Бокса-Кокса



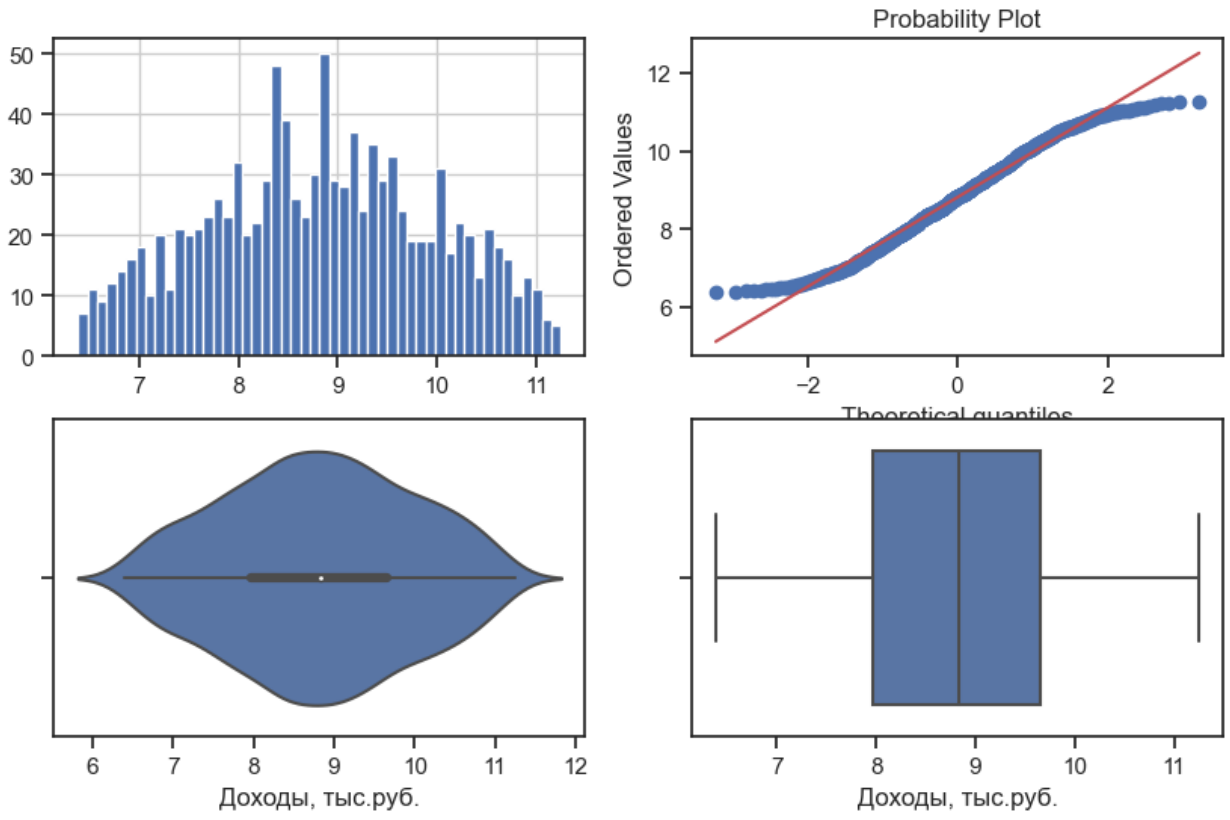


Приложение Б

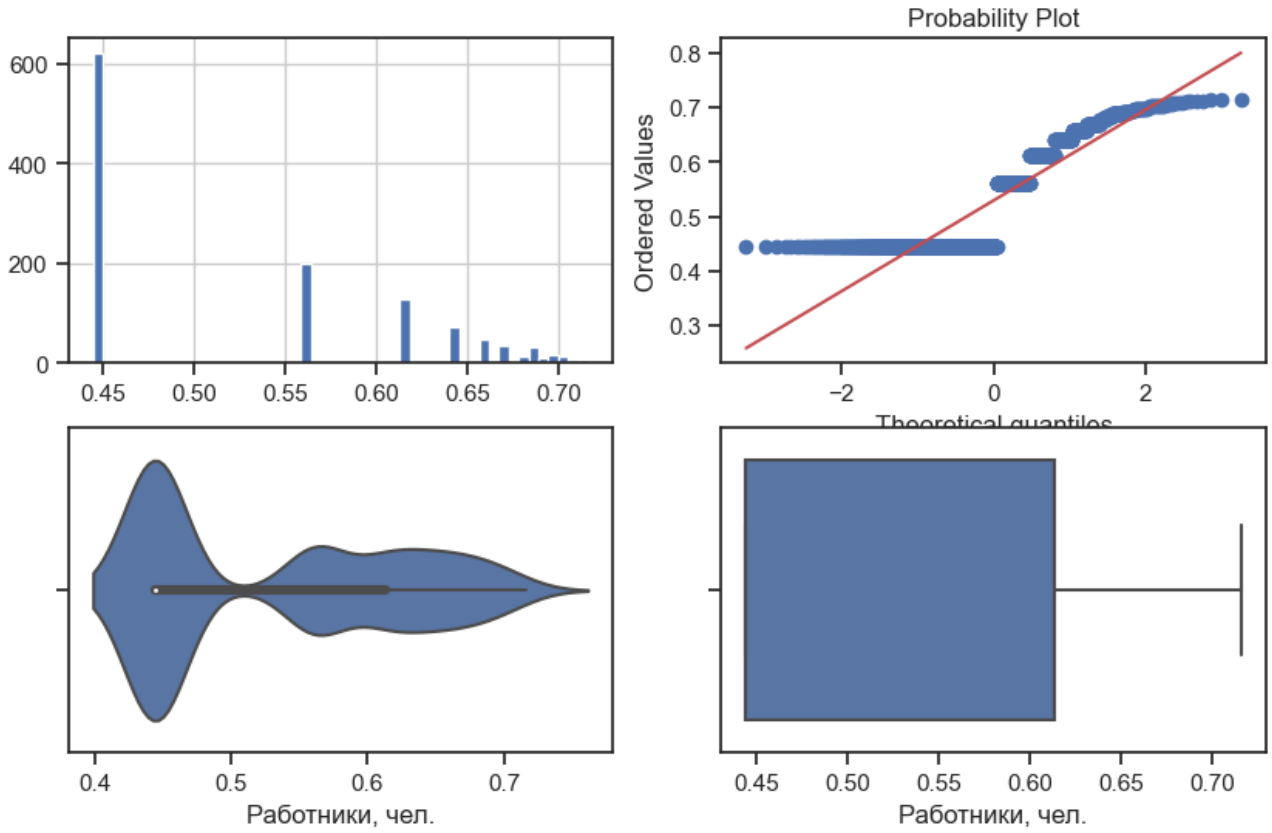
Поле-Доходы, тыс.руб., метод-OutlierBoundaryType.SIGMA, строк-1181



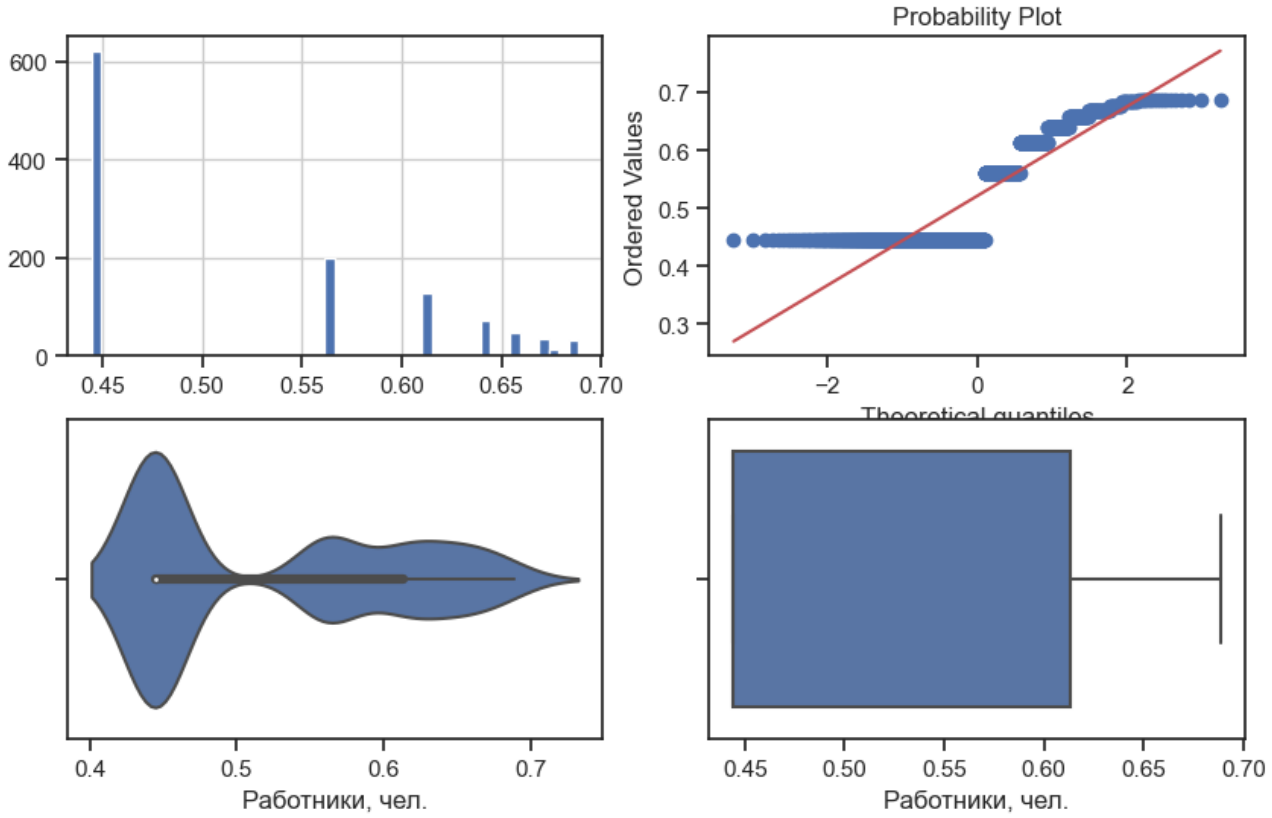
Поле-Доходы, тыс.руб., метод-OutlierBoundaryType.QUANTILE, строк-1181



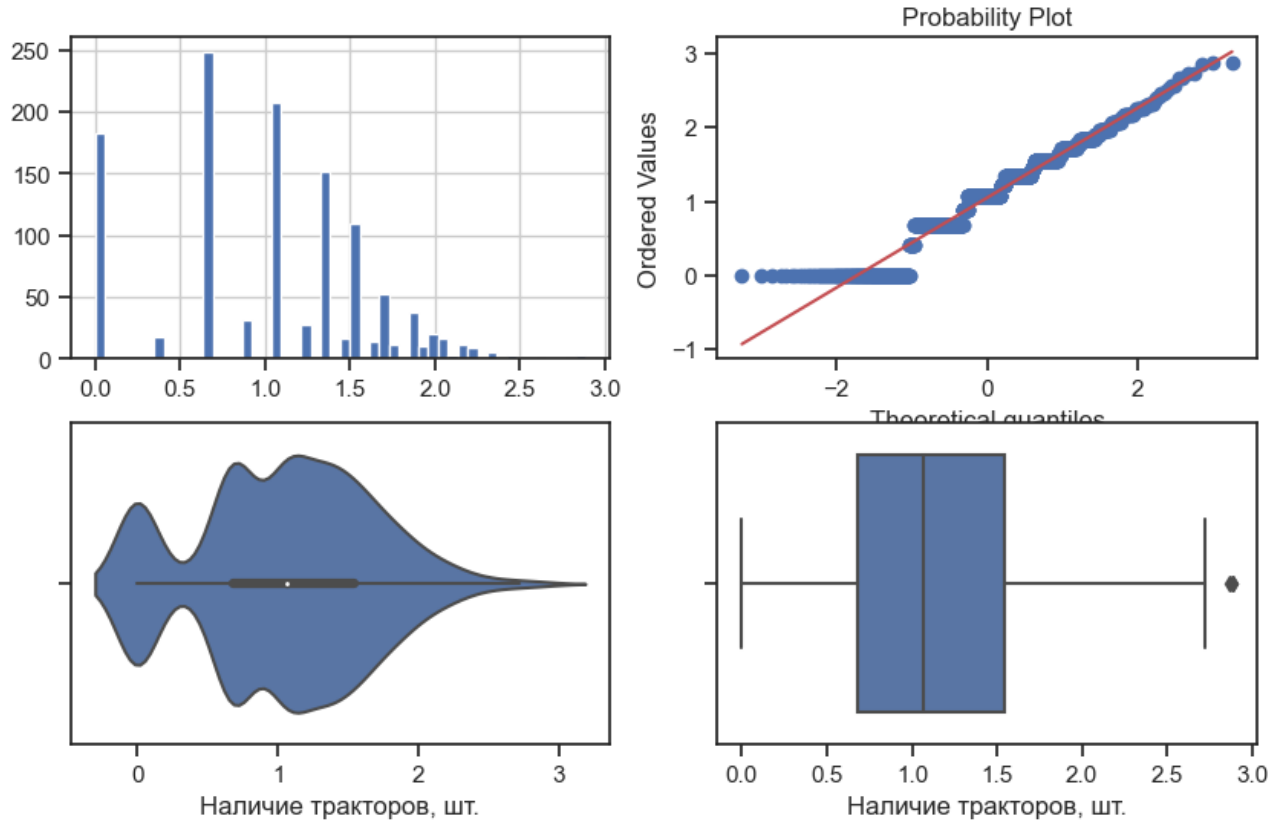
Поле-Работники, чел., метод-OutlierBoundaryType.SIGMA, строк-1181



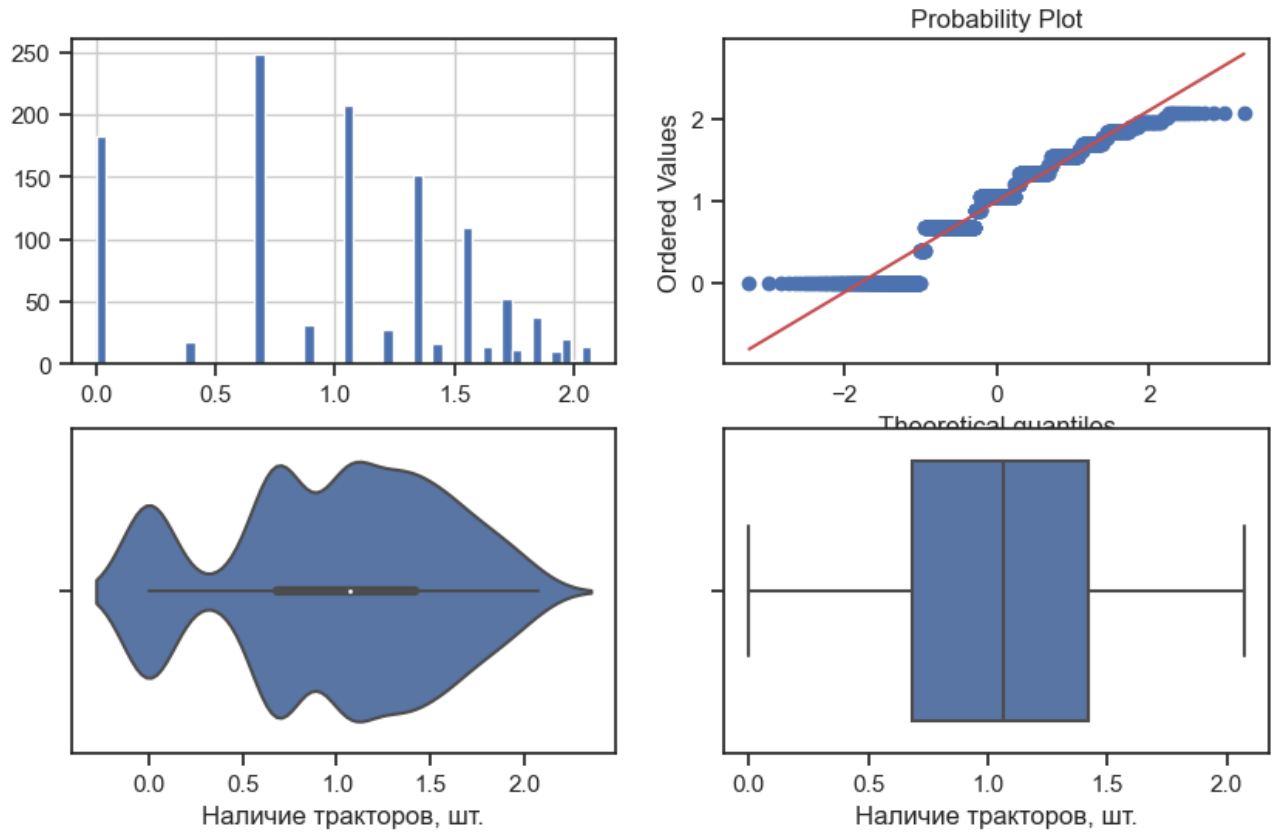
Поле-Работники, чел., метод-OutlierBoundaryType.QUANTILE, строк-1181



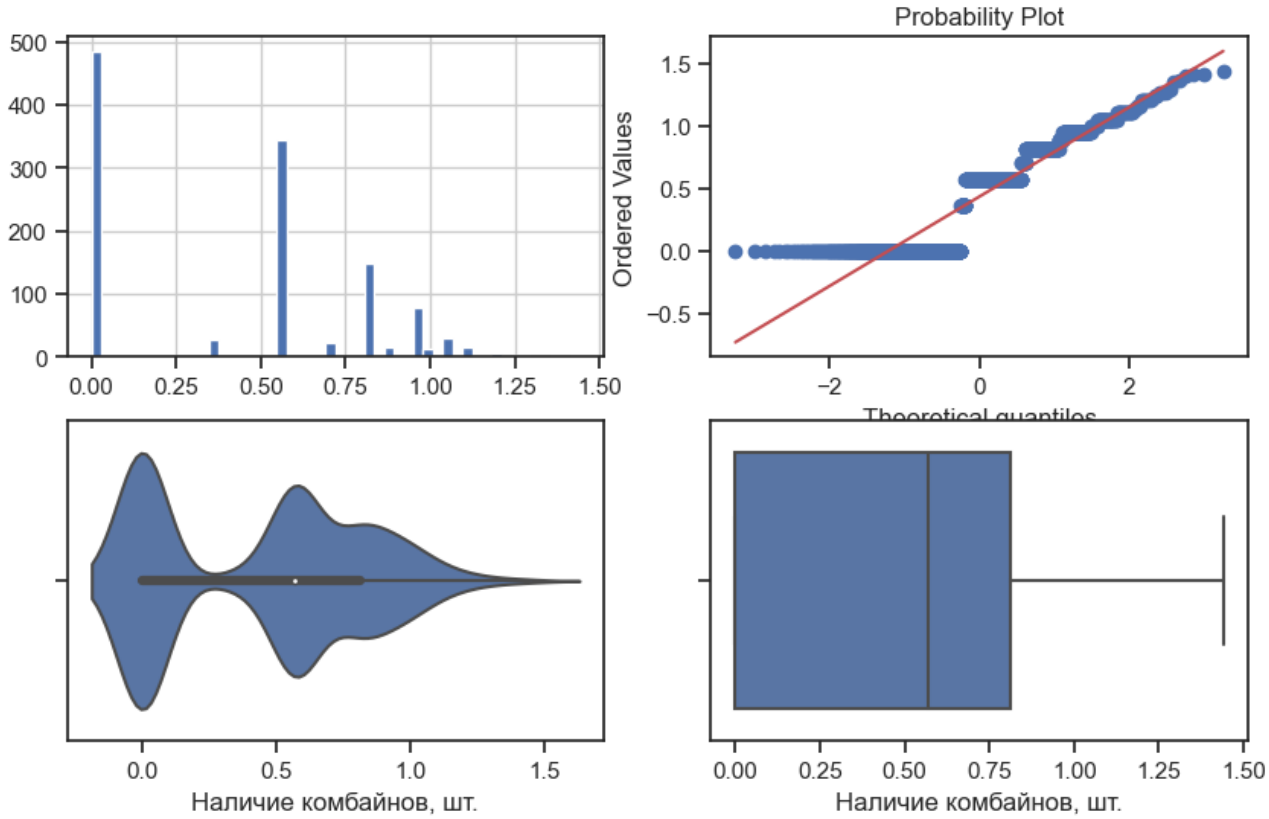
Поле-Наличие тракторов, шт., метод-OutlierBoundaryType.SIGMA, строк-1181



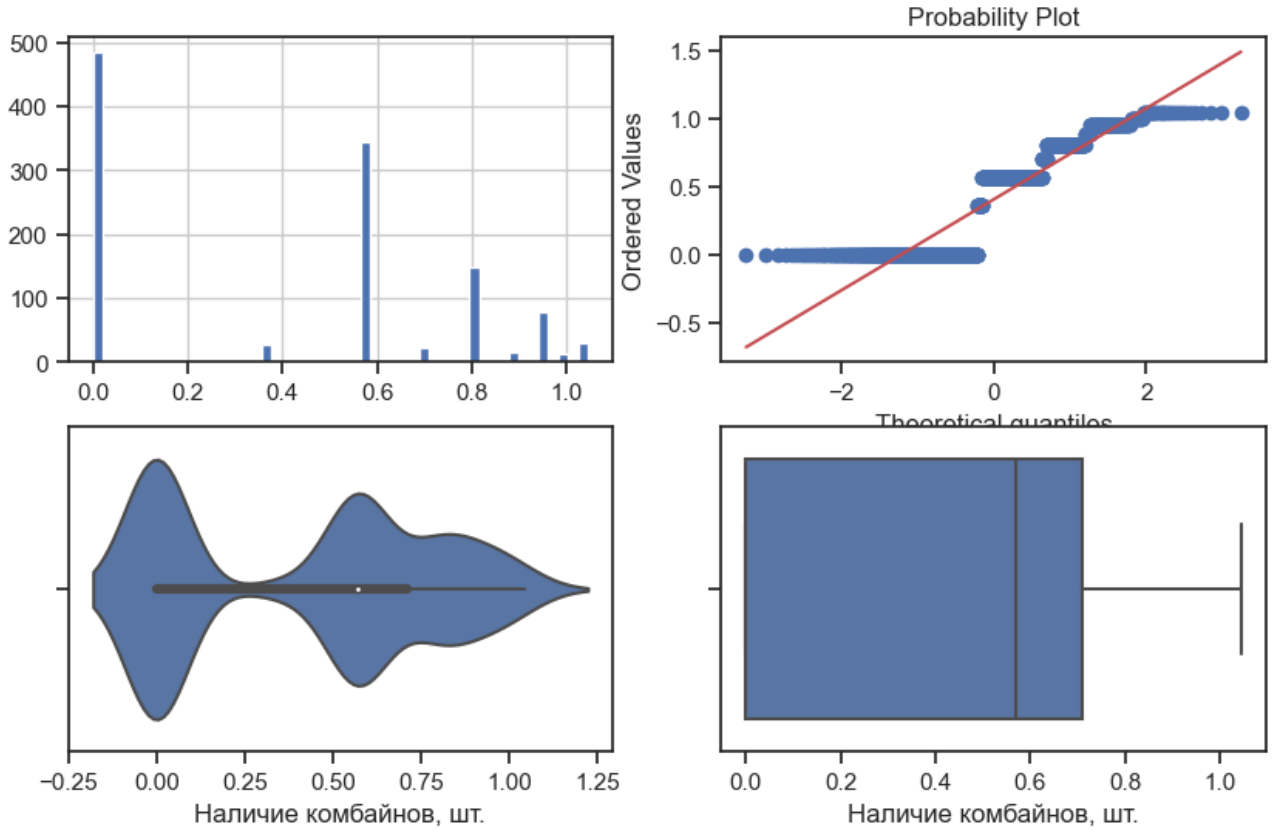
Поле-Наличие тракторов, шт., метод-OutlierBoundaryType.QUANTILE, строк-1181



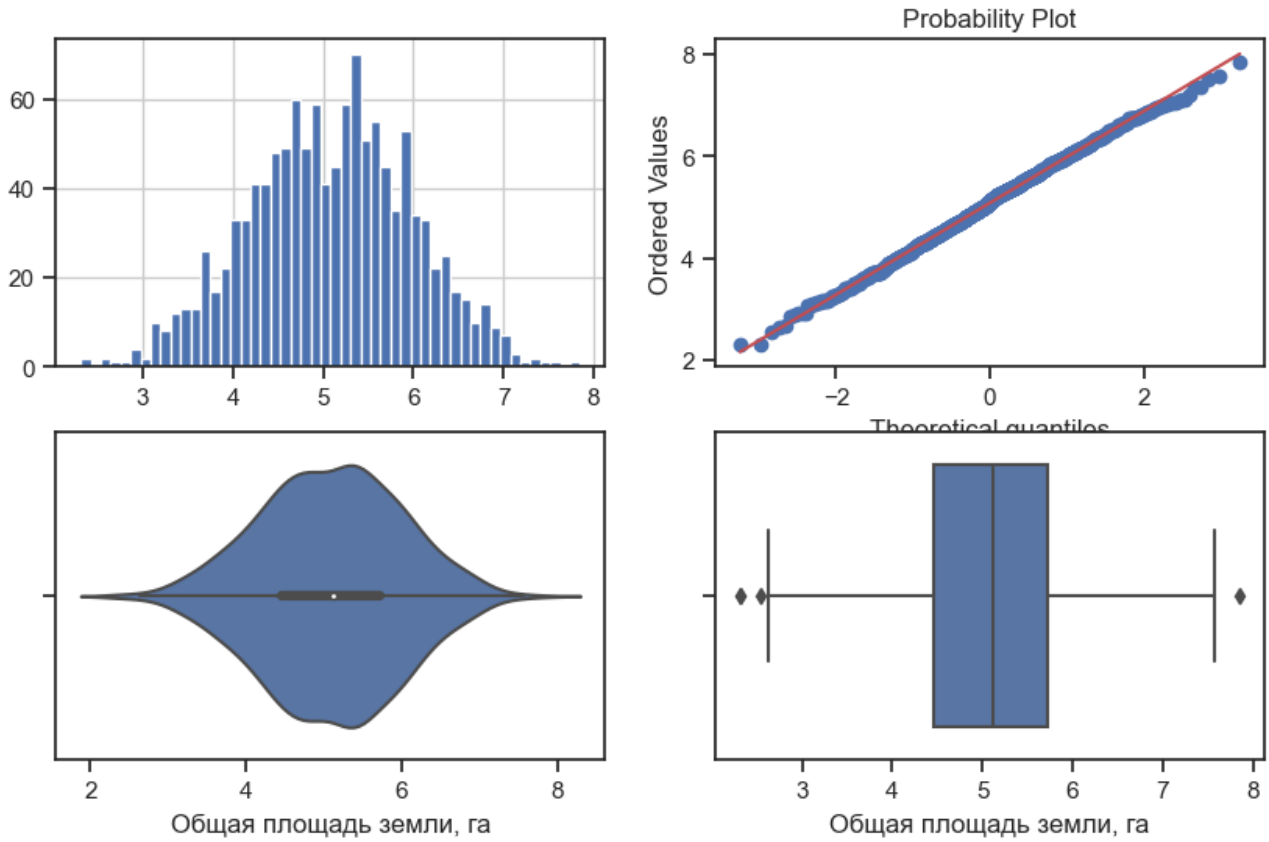
Поле-Наличие комбайнов, шт., метод-OutlierBoundaryType.SIGMA, строк-1181



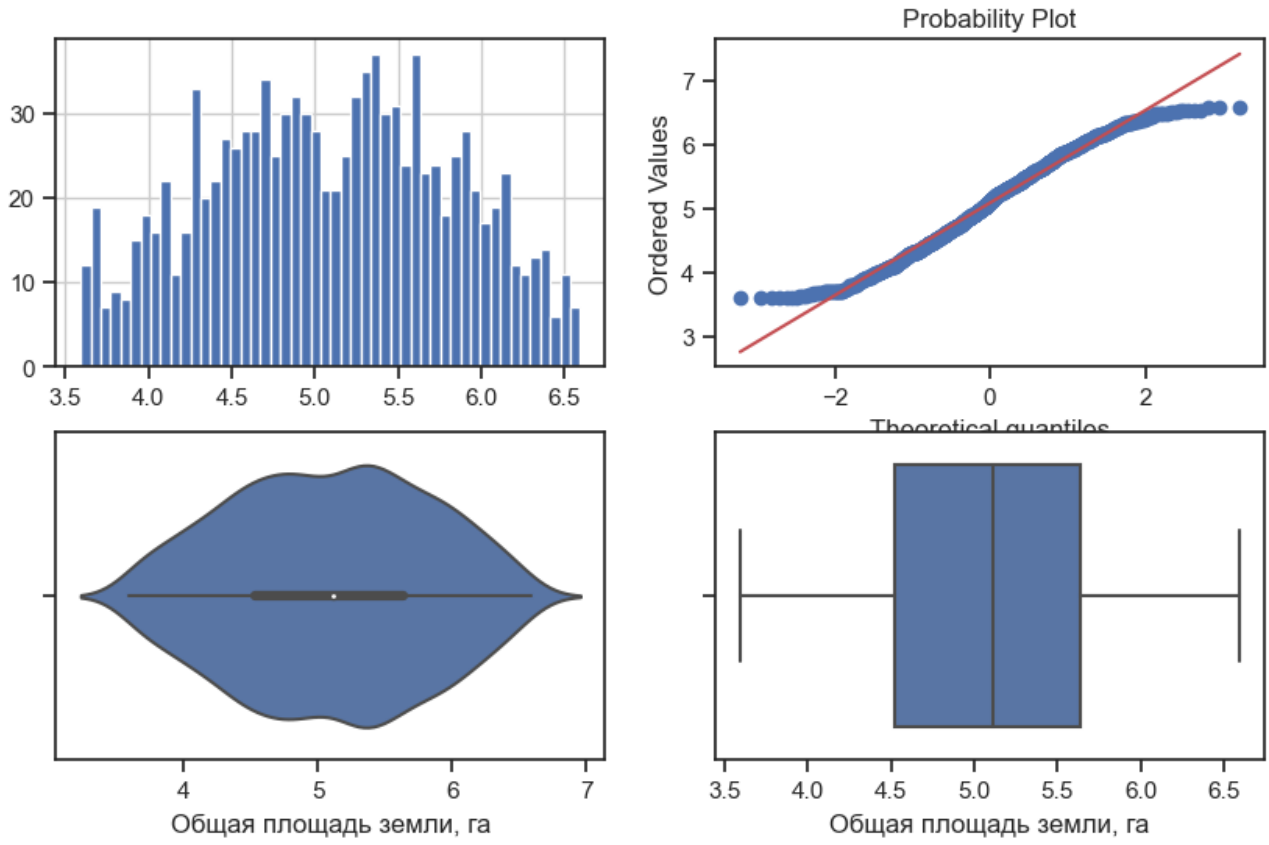
Поле-Наличие комбайнов, шт., метод-OutlierBoundaryType.QUANTILE, строк-1181



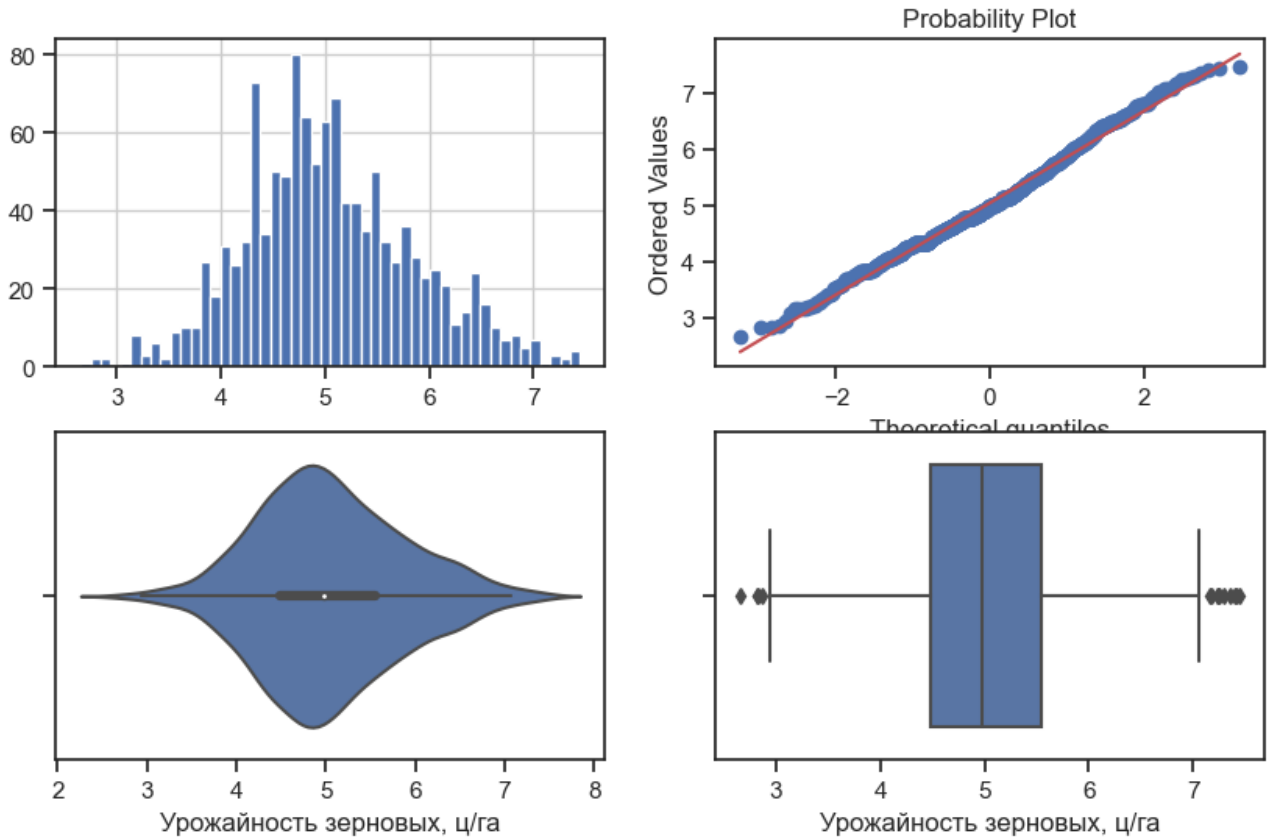
Поле-Общая площадь земли, га, метод-OutlierBoundaryType.SIGMA, строк-1181



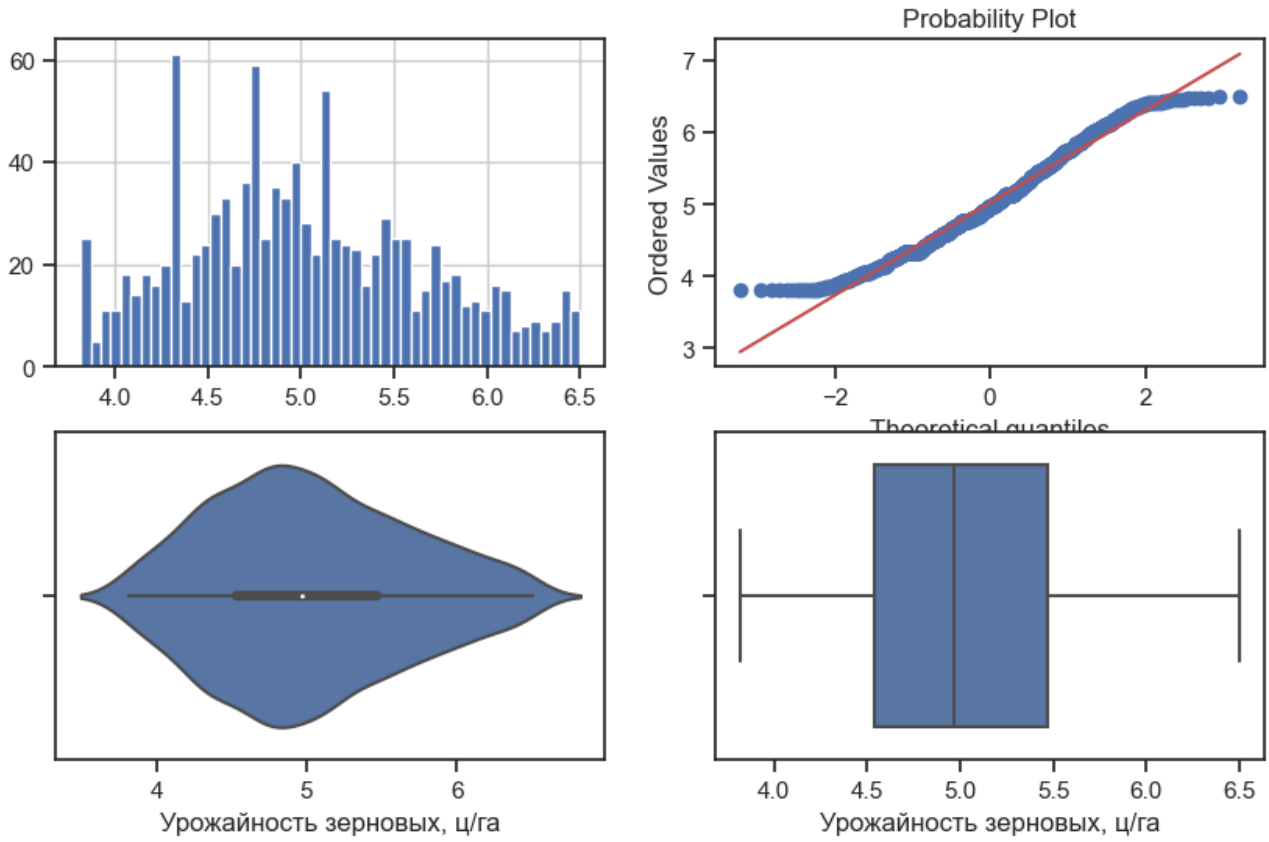
Поле-Общая площадь земли, га, метод-OutlierBoundaryType.QUANTILE, строк-1181



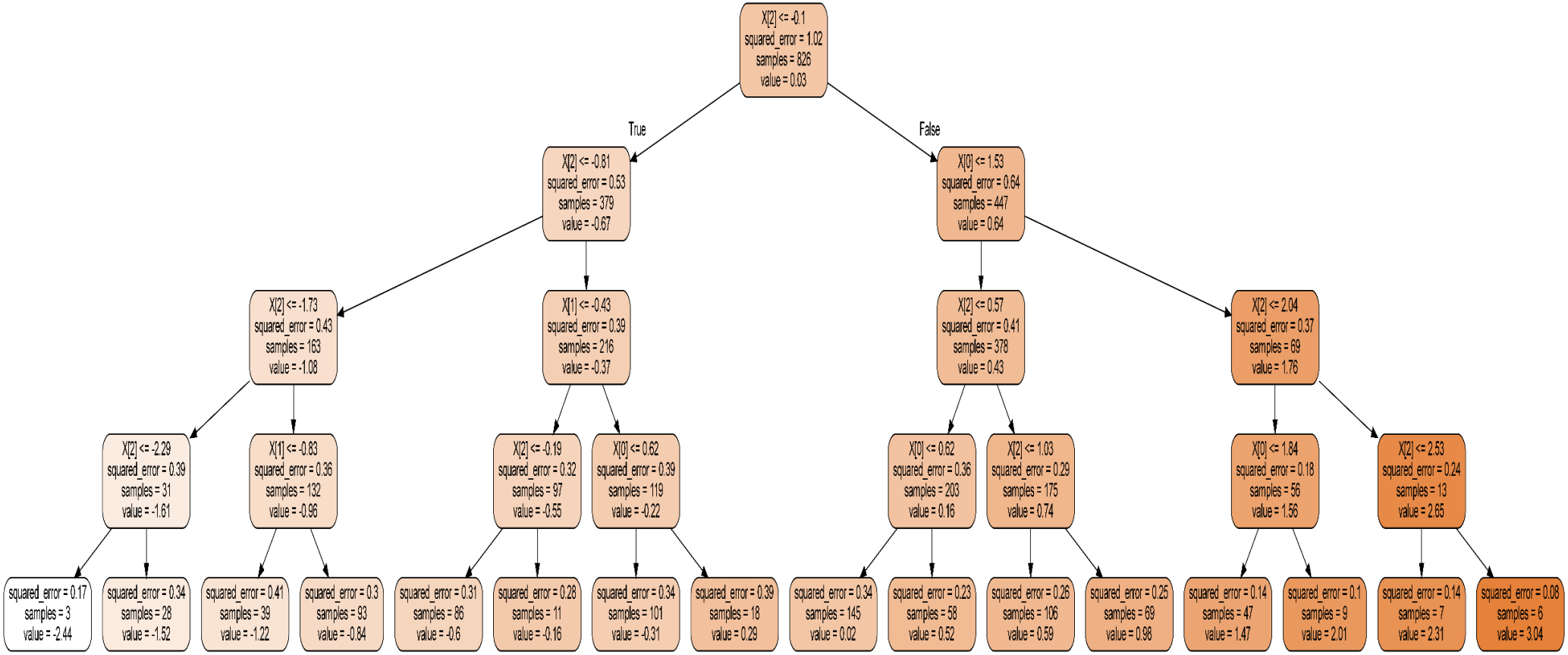
Поле-Урожайность зерновых, ц/га, метод-OutlierBoundaryType.SIGMA, строк-1181



Поле-Урожайность зерновых, ц/га, метод-OutlierBoundaryType.QUANTILE, строк-1181



Приложение В



Приложение Г