Анализ одномерных данных

В качестве данных использовался Российский мониторинг экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE) (29 волна, репрезентативная выборка по индивидам)

Базовая модель

В качестве базовой модели была взята мультиклассовая логистическая регрессия. В качестве данных в модель подаются 5 признаков, среди которых 3 дискретных, 2 категориальных. Искомая переменная «Удовлетворенность» является дискретной и содержит 5 классов.

Основные характеристики модели:

- **Тип модели**: Мультиклассовая классификация (5 классов)
- **Обучение модели**: Градиентный спуск с оптимизатором Adam
- **Функция потерь**: cross entropy
- **Количество признаков**: 5 признаков
- **Число классов**: 5 классов
- **Способ классификации**: Применение softmax к логистической регрессии для мультиклассовой классификации

Метрики качества модели:

- **Accuracy**
- **F1**
- **ROC–AUC**

Определение выбросов и пропусков

Для начала из данных были удалены очевидные ошибки ввода для каждого признака по отдельности (*Таблица 1*). Далее были определены выбросы с помощью метода межквартильного размаха (IQR). С помощью этого метода были найдены верхние и нижние границы возможных значений признаков (*Таблица 2*). Затем были созданы 2 новых датасета, не содержащих пропусков в целевой переменной и очевидных ошибок ввода, в котором все пропуски каждого признака были заменены на его среднее значение или медиану

Таблица 1: Количество ошибок ввода для каждого признака

Признак	Заработная плата	Премия	Отпуск	Рабочая неделя	Образование
Количество ошибок ввода	346	4401	19	368	1

Таблица 2: Верхняя и нижняя границы возможных значений признаков

Признак	Заработная плата	Премия	Отпуск	Рабочая неделя
min	0	0	7.5	28
max	68000	20500	51.5	60

Сэмплирование данных

Заработная плата и премия

На рисунках 1 и 2 видно, что плотности распределения напоминают логнормальное, характеризующееся положительными значениями (>0) и длинным правым хвостом, более наглядно без удаления выбросов это показано на рисунках 3 и 4 в Приложении. Логнормальное распределение часто используется для моделирования доходов населения, так как отражает неравномерность распределения и наличие высоких значений, что соответствует нашим переменным — зарплате и премии

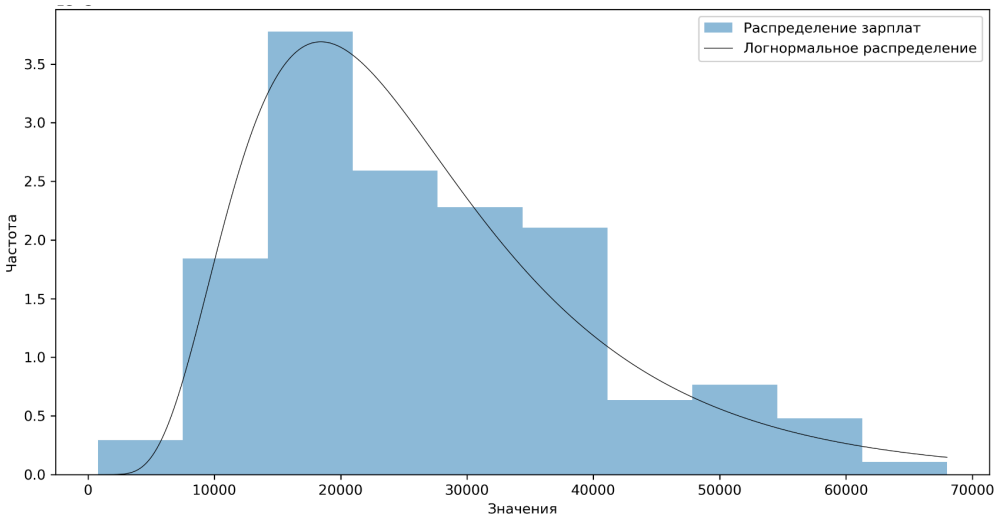

Рис. 1: Логнормальное распределение на плотности распределения заработных плат

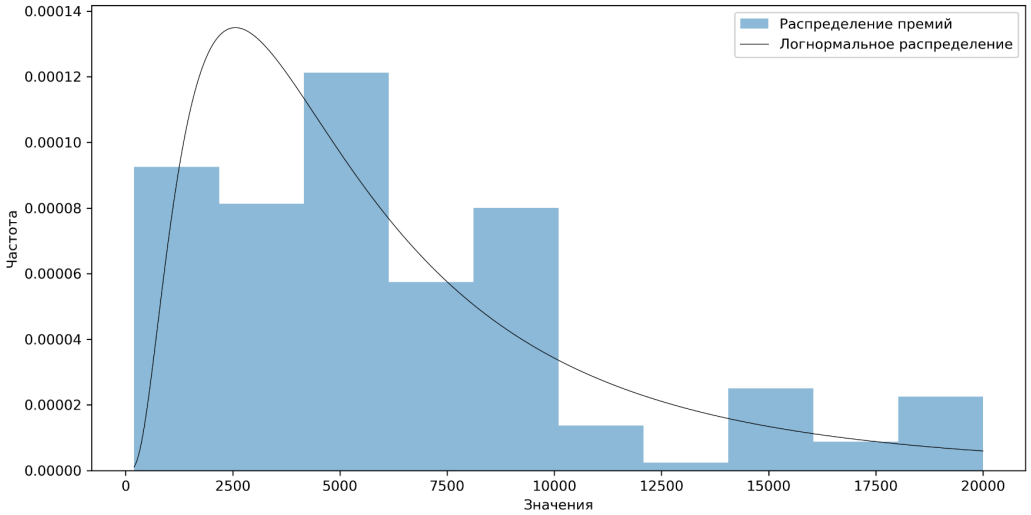

Рис. 2: Логнормальное распределение на плотности распределения премий

Оценка логнормального распределения строилась методом максимального правдоподобия, формула логнормального распределения и ее оцененных параметров представлены ниже:

$$f(x, \mu, \sigma) = \frac{1}{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)\right)^2$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln n_i}{n}$$

$$\hat{\sigma} = \left(\frac{\sum_{i=1}^{n} (\ln n_i - \hat{\mu})^2}{n}\right)^{1/2}$$

Ниже представлены изменения плотности распределения зарплат (*Рисунок 5*) и премий (*Рисунок 6*) после сэмплирования на основе логнормального распределения. Для текущих и последующих переменных на сэмплирование были наложены ограничения, обеспечивающие попадание данных в ранее определенные интервалы, очищенные от выбросов

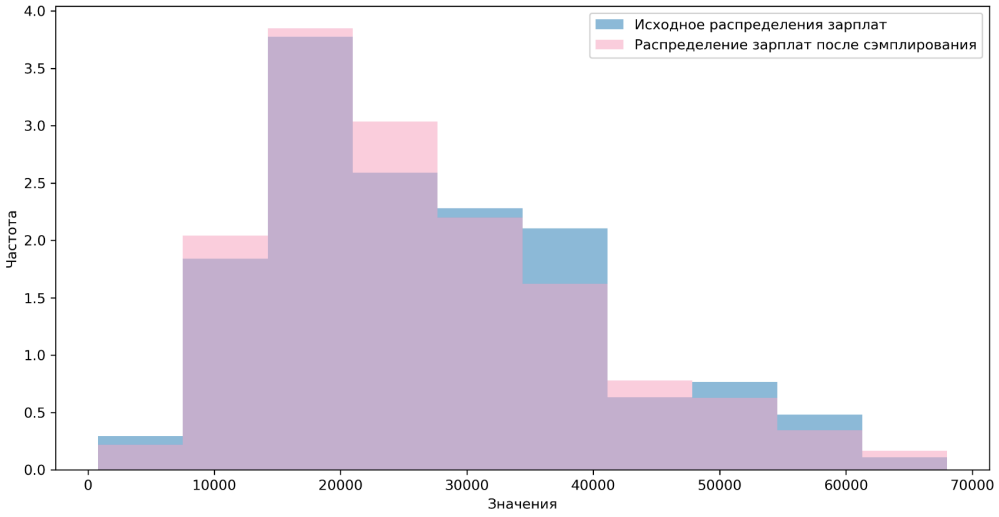

Рис. 5: Распределение плотности зарплат до и после сэмплирования

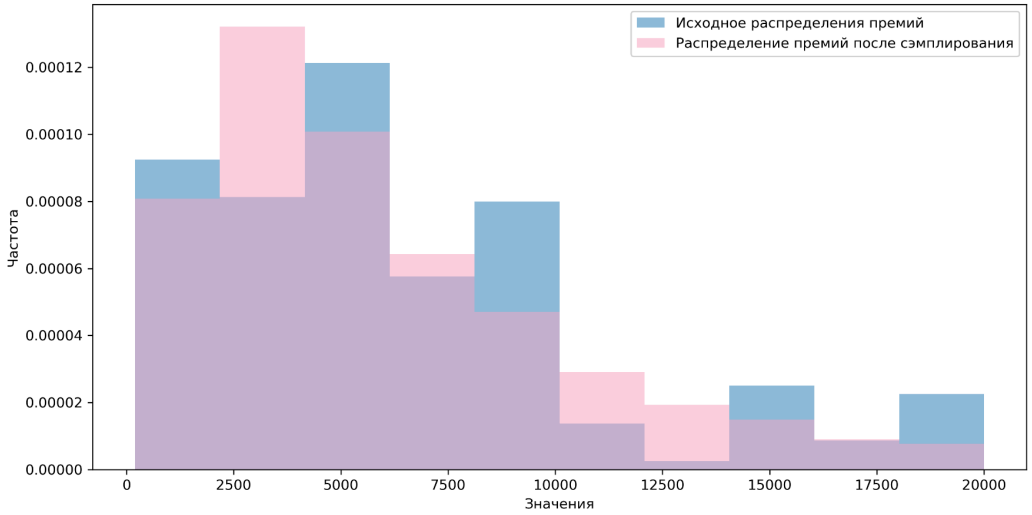

Рис. 6: Распределение плотности премий до и после сэмплирования

Отпуск и рабочая неделя

Плотности распределения дней отпуска и рабочих часов в неделю (*Рисунки 7 и 8 соответственно*) на первый взгляд не соответствуют никакому известному распределению. Однако, принимая во внимание, что отпуск и рабочие часы имеют место у каждого работающего человека, а также учитывая законодательные нормы (в большинстве случаев 28 дней отпуска и 40 часов рабочей недели), которые формируют средние значения у распределений, было решено использовать для сэмплирования нормальное распределение

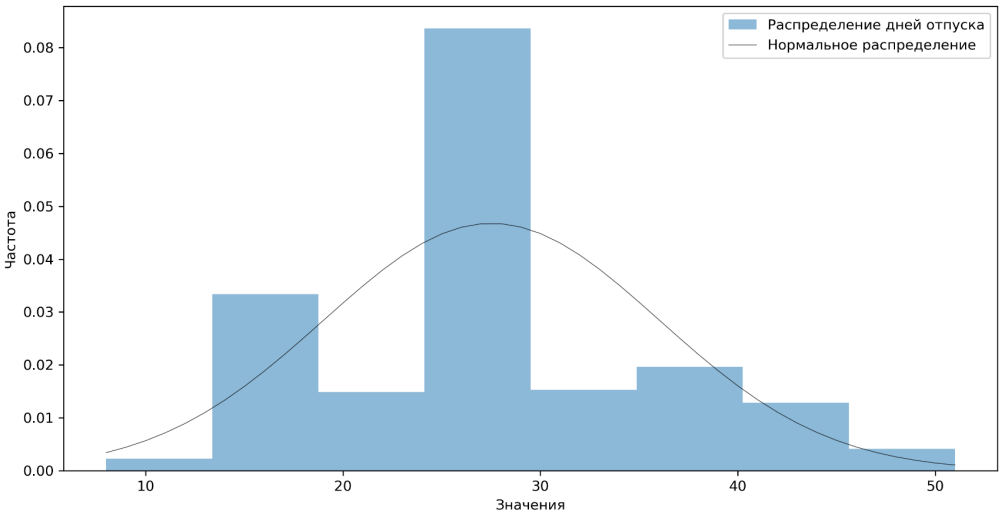

Рис. 7: Нормальное распределение на плотности распределения дней отпуска

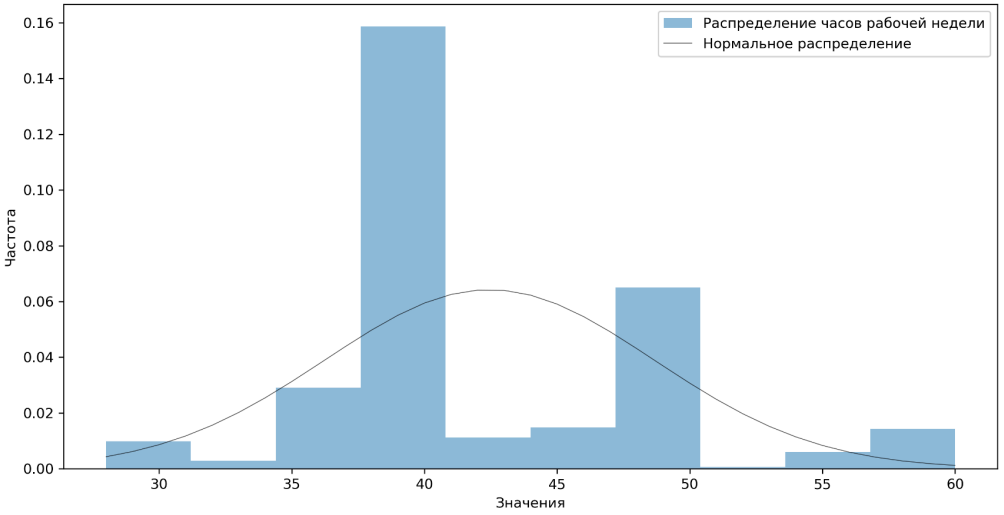

Рис. 8: Нормальное распределение на плотности распределения часов рабочей недели

Как видно на рисунке 8, распределение имеет два выраженных пика, что изначально предполагало использование смеси распределений. Однако этот подход не дал ожидаемых результатов, поскольку второй компонент гауссовой смеси практически не описывал данные, что продемонстрировано на Рисунке 9 в Приложении

Оценка нормального распределения строилась методом максимального правдоподобия, формула нормального распределения и ее оцененных параметров представлены ниже:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\widehat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Ниже представлены изменения плотности распределения дней отпуска (*Рисунок 10*) и часов рабочей недели (*Рисунок 11*) после сэмплирования на основе нормального распределения. С учетом того, что текущие переменные не могут принимать дробные значения сэмплированные данные были дополнительно округлены до ближайшего целого числа

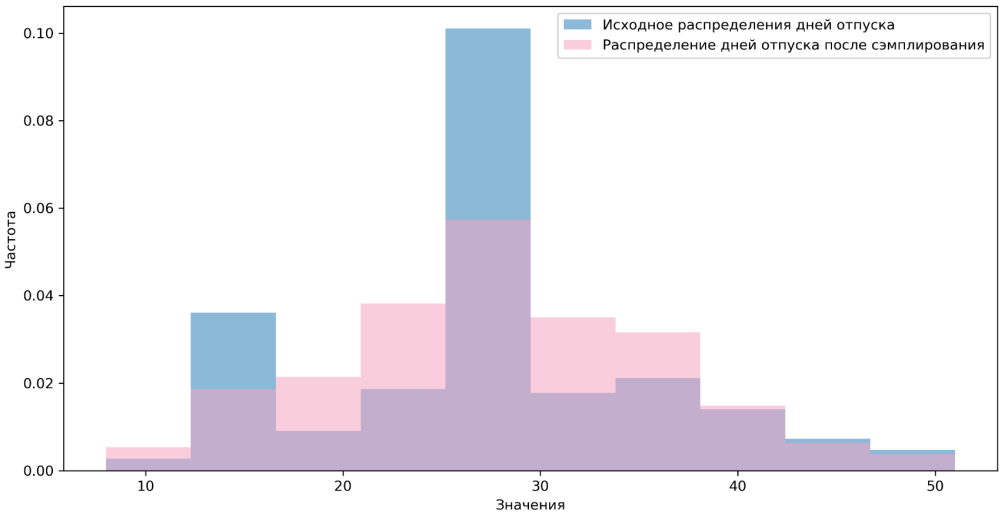

Рис. 10: Распределение плотности дней отпуска до и после сэмплирования

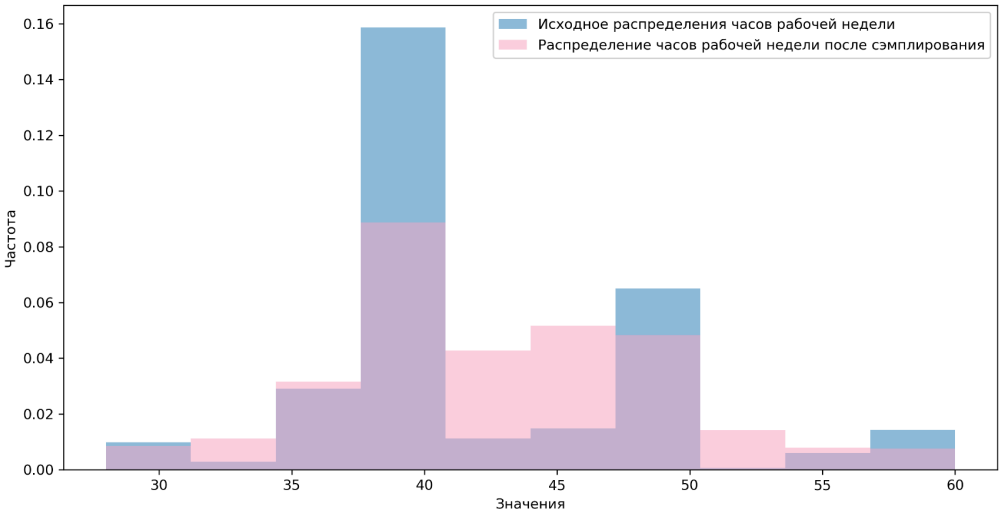

Рис. 11: Распределение часов рабочей недели до и после сэмплирования

Уровень образования и удовлетворенность работой

Учитывая дискретный характер значений уровня образования и удовлетворенности работой, а также низкую вероятность некоторых ответов в этих переменных, было принято решение в качестве оценки распределения использовать распределение Пуассона. Поскольку переменная удовлетворенность работой является целевой и содержит относительно небольшое количество пропусков, сэмплирование для нее не проводилось

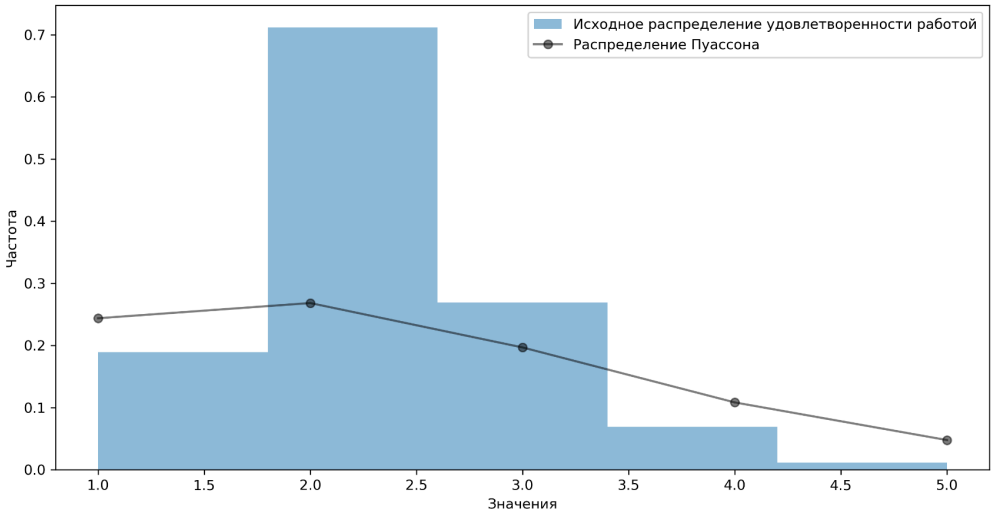

Рис. 12: Распределение Пуассона на плотности распределения удовлетворенности работой

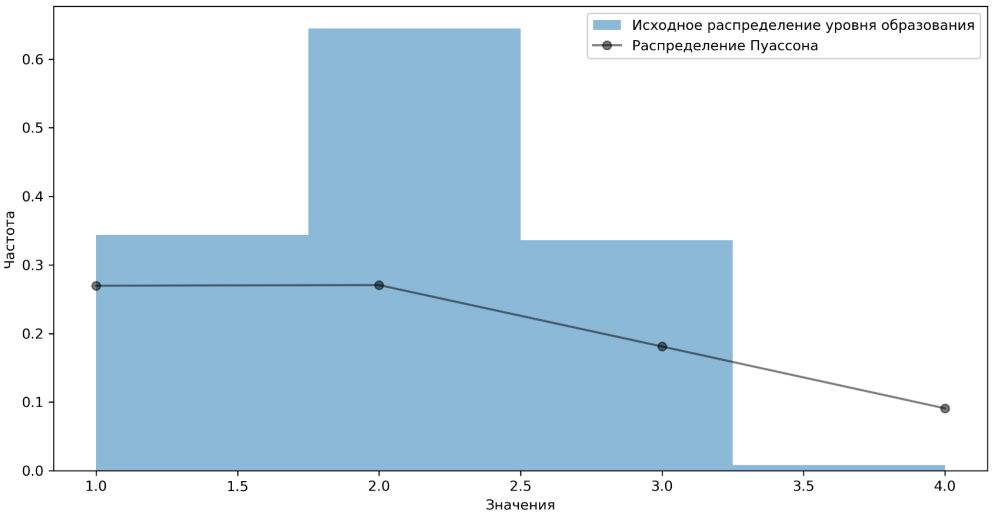

Рис. 13: Распределение Пуассона на плотности распределения уровня образования

Оценка распределения Пуассона строилась методом максимального правдоподобия, формула распределения Пуассона ее оцененных параметров представлены ниже:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Ниже представлено изменение плотности распределения уровня образования (*Рисунок 14*) после сэмплирования на основе распределения Пуассона

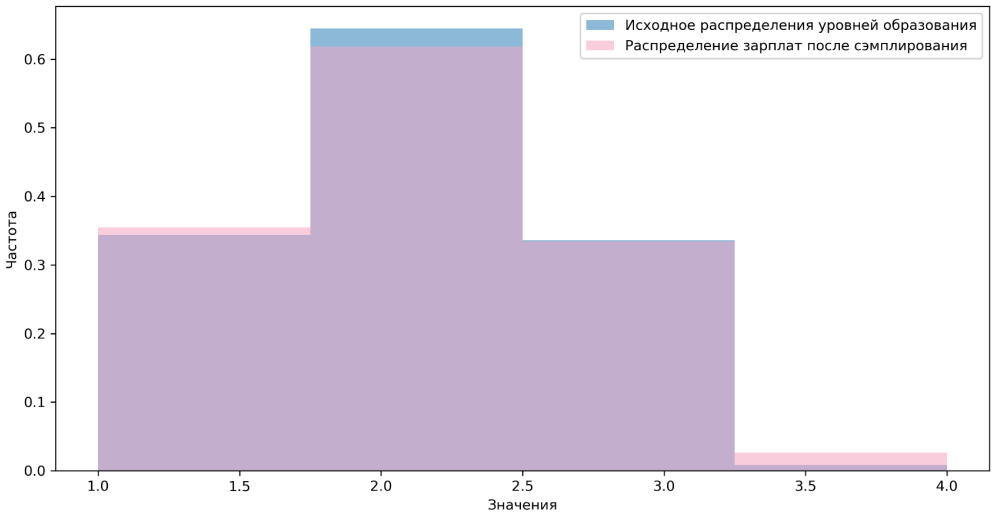

Рис. 14: Распределение уровней образования до и после сэмплирования

Описательная статистика

Описательная статистика проводилась на 4-х вариациях данных:

1. Исходный датасет без пропусков и ошибок ввода. Эти данные использовались для первоначального обучения логистической регрессии
2. Данные, в которых пропуски и ошибки ввода были заменены на средние значения, а выбросы удалены
3. Данные, в которых пропуски и ошибки ввода были заменены на медиану, а выбросы удалены
4. Исходные данные без выбросов + сэмплированные на их основе значения

Далее приведены сравнения 4-х видов данных для каждого параметра и их статистическое описание

Заработная плата

Данная часть датасета содержит ответы на вопрос “За последние 12 месяцев какова была Ваша среднемесячная зарплата на этом предприятии после вычета налогов - независимо от того, платят Вам ее вовремя или нет?”

Таблица 3. Стат.характеристики заработной платы

Стат.характеристика	Датасет 1	Датасет 2	Датасет 3	Датасет 4
Среднее	35397,89	26836,83	26660,4	26791,52
Медиана(Q50%)	30000	25000	25000	25000
Станд.откл	20177,38	11887,7	11898,19	12525,19
Размах	192000	66919	66919	66919
Q25%	22000	18000	18000	17017,7
Q75%	45000	33000	33000	35000
Асимметрия	2,59	0,79	0,83	0,77
Эксцесс	13,94	0,37	0,40	0,11

В датасетах 2, 3, 4 можно заметить, что среднее и медиана очень близки друг к другу, что может говорить о симметричности распределения

Самые минимальные стандартные отклонения у 2-ого и 3-его датасета, так как были удалены выбросы и шумы, что привело к стабильности данных, но при сэмплировании стандартное отклонение увеличилось, что говорит о том, что выбранная нами модель распределения неидеально описывает значения

Самое минимальное значение асимметрии у 4 датасета, что говорит о том, что распределение находится симметричнее относительно среднего значения, нежели остальные датасеты

В датасете 1 эксцесс равен 13,94, что указывает на то, что присутствует много экстремальных значений, что типично для данных с выбросами или шумами.

Четвертый датасет имеет минимальный эксцесс, т.к. сэмплирование происходило по нормальному распределению

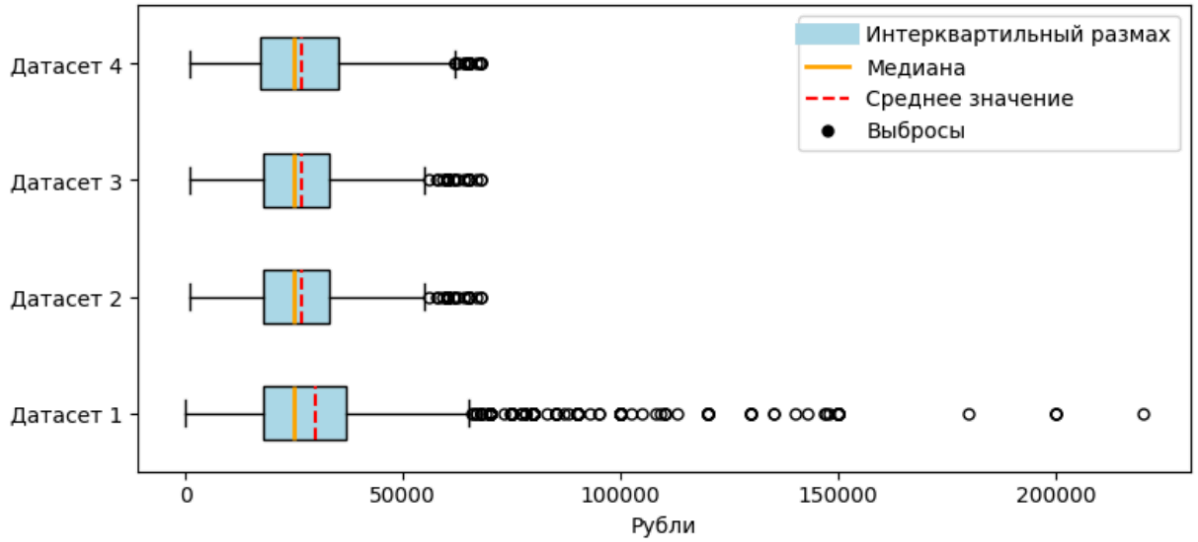

Рис. 15: Ящик с усами для зарплат

Можно заметить, что датасет 1 имеет большое количество выбросов, что свидетельствует о том, что данный датасет нужно чистить. Также среднее и медиана не совпадают, что говорит о том, что данное распределение несимметрично. Медиана находится левее среднего, что говорит о хвосте справа. Асимметрия, равная 2,6 подтверждает этот вывод. Датасет 4 имеет наименьшее количество выбросов

Рабочая неделя

Данная часть датасета содержит ответы на вопрос “Сколько часов в среднем продолжается Ваша обычная рабочая неделя?”

Таблица 4. Стат.характеристики рабочей недели

Стат.характеристика	Датасет 1	Датасет 2	Датасет 3	Датасет 4
Среднее	42,14	42,47	42,47	42,47
Медиана(Q50%)	40	40	40	40
Станд.откл	7,47	6,14	6,14	6,14
Размах	68	32	32	32
Q25%	40	40	40	40
Q75%	45	48	48	48
Асимметрия	2,02	0,92	0,92	0,92
Эксцесс	8,65	1,29	1,29	1,29

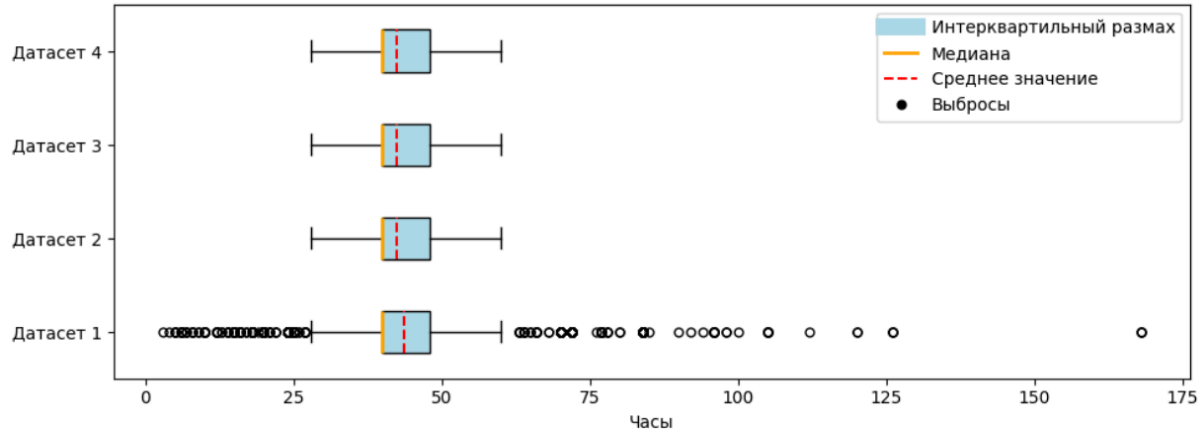

Рис. 16: Ящик с усами для рабочей недели

Во всех 4х датасетах среднее и медиана не совпадают, то говорит о несимметричности распределений, но в 1ой датасете оно максимально. Асимметрия везде положительна, что говорит о хвосте справа

Выбросы отсутствуют во 2, 3, 4 датасетах, ибо был сильно урезан диапазон возможных значений

Образование

Изначальные значения текущей переменной являлись отражением ответа на вопрос “Какой у Вас самый высокий уровень образования, по которому Вы получили аттестат, свидетельство, диплом?”, где минимальным значением являлась 1 и максимальным 17.

Ввиду того, что часть значений расположены не последовательно на условной шкале (*значение 15*), а также наличие пропущенных значений в последней (*значения 7-9*), было принято решение объединить категории по следующему принципу:

- 1 - ‘Школьное образование’ (значения 1–2)
- 2 - ‘Профессиональные училища’ (значения 3–6 и 15)
- 3 - ‘Высшее образование’ (значения 10–12)
- 4 - ‘Последипломное образование’ (значения 16 и 17)

Таблица 5. Стат.характеристики образования

Уровень образования	Датасет 1	Датасет 2	Датасет 3	Датасет 4
1	27	434	434	434
2	150	1863	1863	1863
3	154	1092	1092	1092
4	1	13	13	13

Можно заметить отличие датасета 1 от остальных, т.к. данный датасет крайне ограничен (всего 332 значений) и может не отражать реальную картину

Премия

Данная часть датасета содержит ответы на вопрос “Если Вы получали премию по основному месту работы в течение последних 30 дней, то сколько рублей Вы получили?”

Таблица 6. Стат.характеристики премий

Стат.характеристика	Датасет 1	Датасет 2	Датасет 3	Датасет 4
Среднее	8507,38	6483,54	5135,96	5971,8
Медиана(Q50%)	5000	6487,01	5000	4915,36
Станд.откл	8587,84	1448,83	1509,19	4232,91
Размах	59800	19800	19800	19800
Q25%	3000	6487,01	5000	2801,02
Q75%	10000	6487,01	5000	8018,02
Асимметрия	2,46	3,92	5,99	1,19
Эксцесс	7,81	40,09	50,47	0,99

Можно заметить высокий эксцесс у 2-ого и 3-его датасетов, что может говорить о несбалансированности распределений, т.е. каких-то значений, резко отличающихся от других, очень много

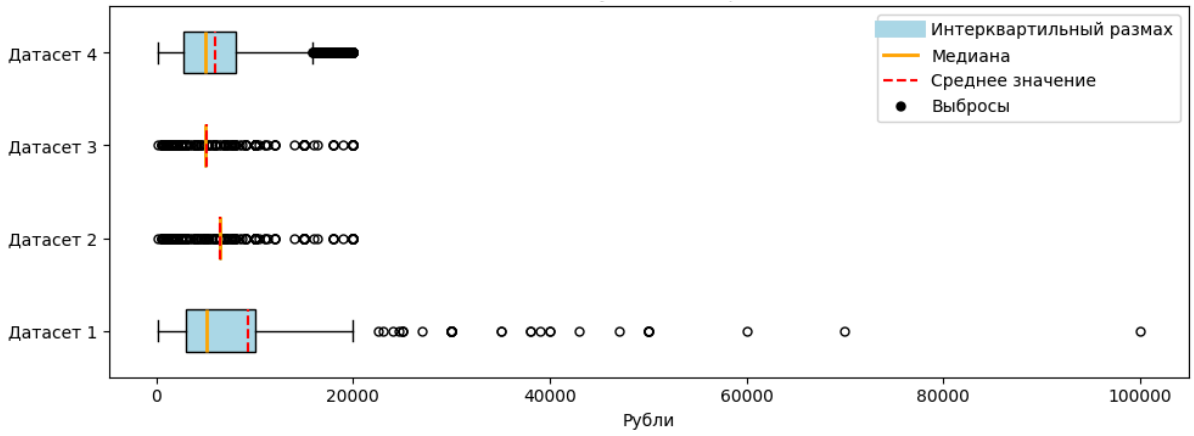

Рис.17: Ящик с усами для премии

Датасет 1 и датасет 4 похожи, т.к. датасет 4 построен на датасете 1. Датасеты 2 и 3 имеют такой вид, т.к. большое количество данных сконцентрировано на среднем и медиане из-за способа избавления от пропусков (заменой на медиану или среднее)

Отпуск

Данная часть датасета содержит ответы на вопрос “Сколько всего календарных дней продолжался или продолжается Ваш отпуск за последние 12 месяцев?”

Таблица 7. Стат.характеристики отпусков

Стат.характеристика	Датасет 1	Датасет 2	Датасет 3	Датасет 4
Среднее	30,42	27,42	27,59	27,8
Медиана(Q50%)	28	27,49	28	28
Станд.откл	11,31	6,72	6,72	8,27
Размах	53	43	43	43
Q25%	28	27,49	28	24
Q75%	36	28	28	32
Асимметрия	0,59	0,17	0,09	0,1
Эксцесс	0,24	1,75	1,72	-0,06

Выводы по данным характеристикам аналогичны выводам по заработной платы

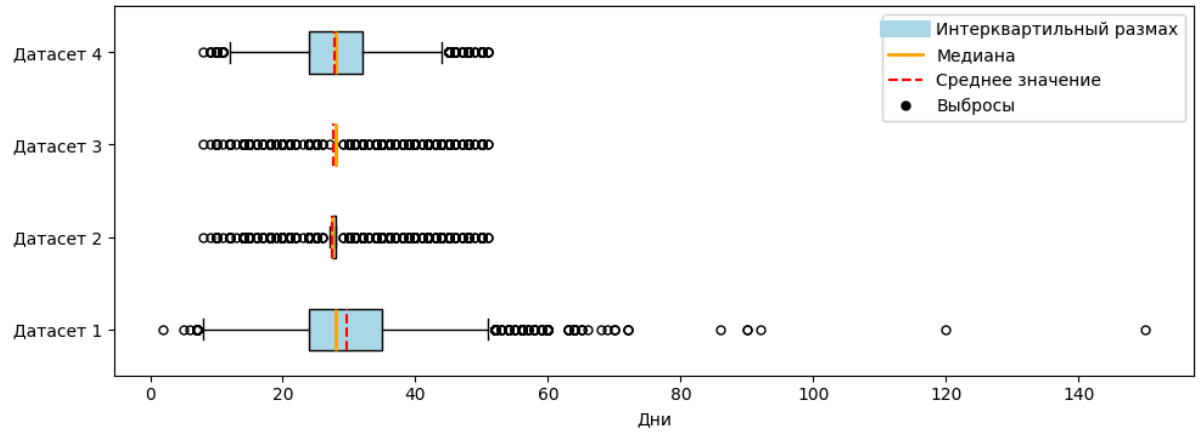

Рис. 18: Ящики с усами для отпусков

Датасет 1 и датасет 4 похожи, т.к. датасет 4 построен на датасете 1. Датасеты 2 и 3 имеют такой вид, т.к. большое количество данных сконцентрировано на среднем и медиане из-за способа избавления от пропусков (заменой на медиану или среднее)

Результаты обучения модели

Необработанный датасет

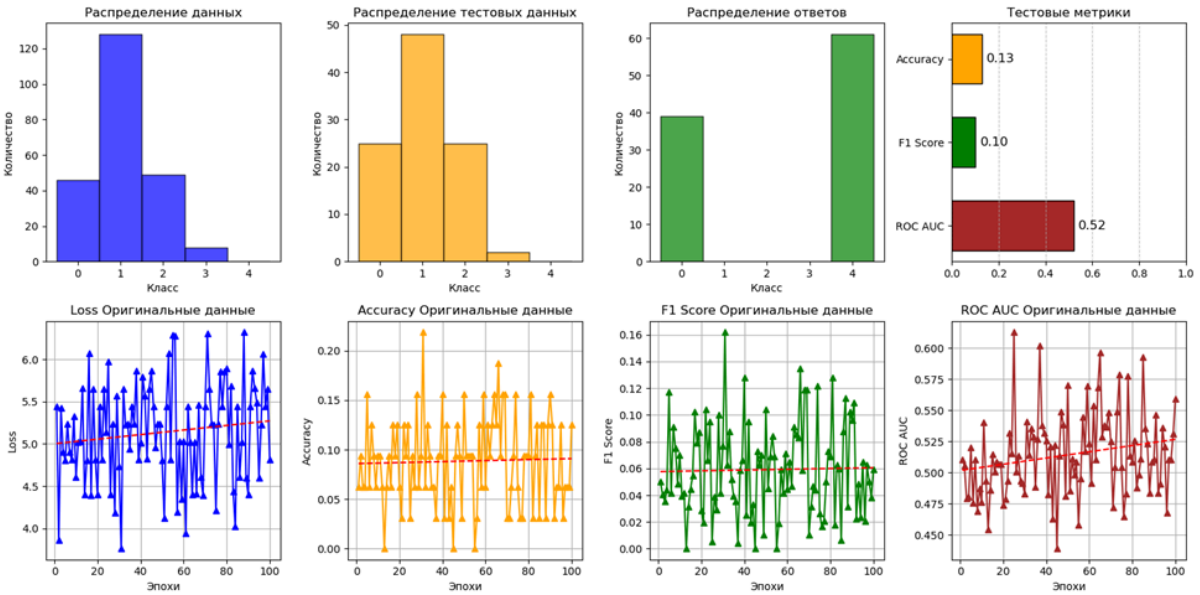

Рис. 19: Графики метрик модели на необработанном датасете

Линейная модель в целом демонстрируют далеко не самые лучше результаты (*Рисунок 19*). На необработанных данных заметно сильное

колебание точности и f1 на уровне значений 0.1. Вероятнее всего это связано с небольшой выборкой (300 строк для тренировочных данных) и значительным количеством шумов и выбросов. Значения ROC–AUC подтверждают, что модель обладает самой лучшей предсказательной способностью

Сэмплированный датасет

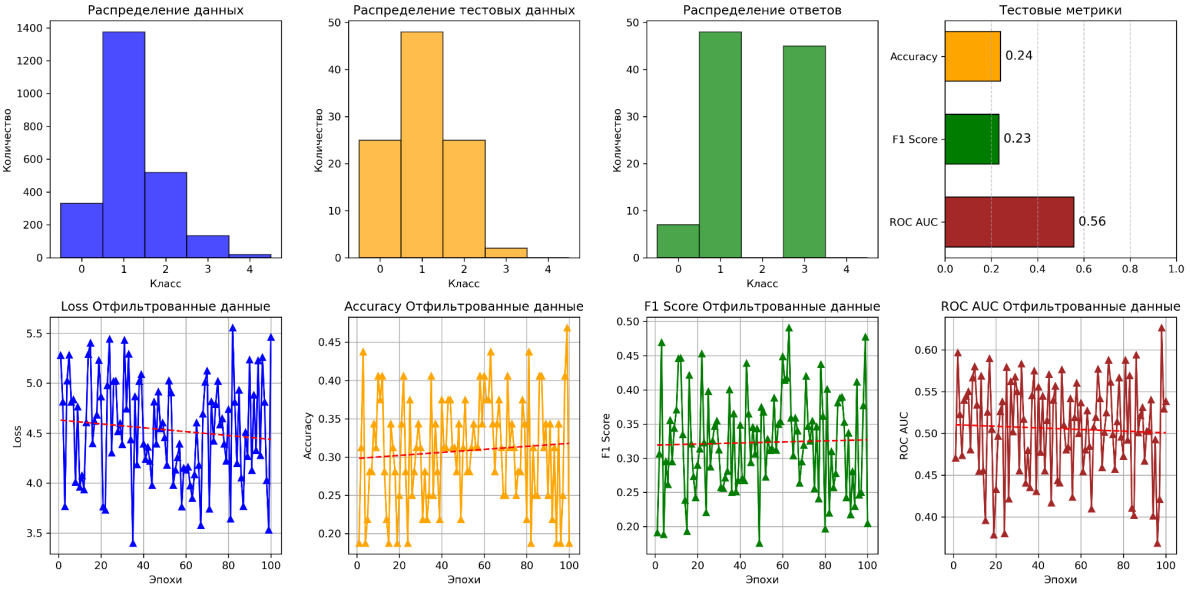

Рис. 20: Графики метрик модели на семплированном датасете

Сэмплирование данных показало явное, но не очень большое улучшение метрик модели (*Рисунок 20*). Так Loss стал более стабилен, а значения точности и f1 повысились. Улучшение предсказательных способностей отразил и ROC–AUC. Наиболее любопытные моменты можно отметить на графике распределения ответов модели, которая теперь научилась предсказывать большее количество распространенных классов. Предположительно это связано с увеличением размера датасета (3000 строк) и со стабилизацией разброса данных

Датасет с заменой пропусков на среднее

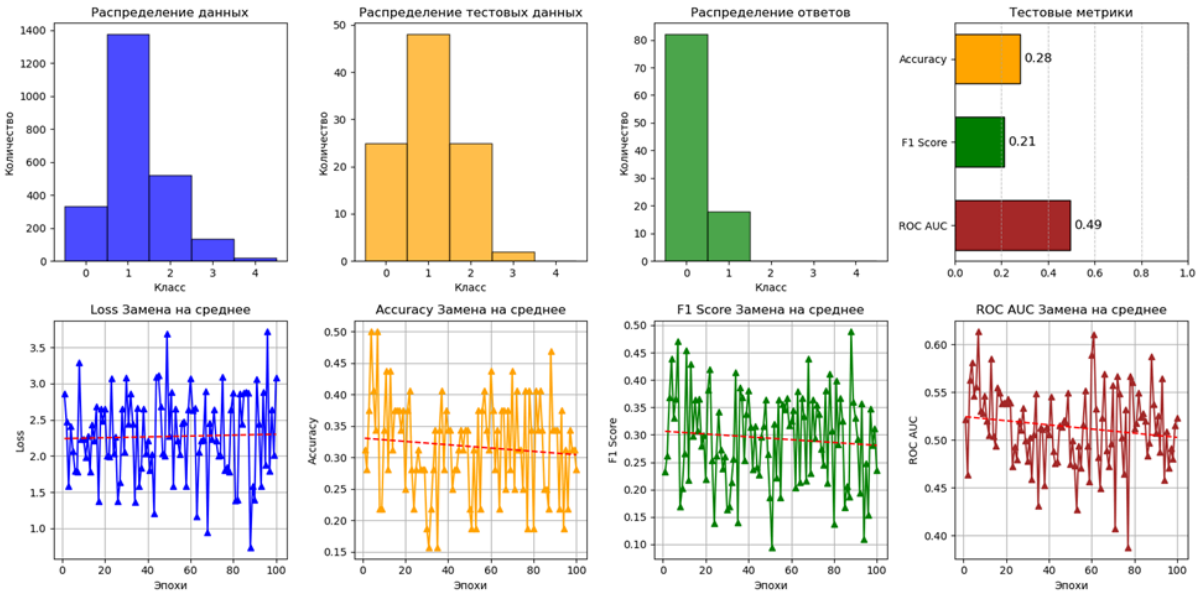

Рис. 21: Графики метрик модели на датасете с заменой на среднее

Замена пропусков на среднее показала схожие метрики с сэмплированными данными (*Рисунок 21*), но предсказательные способности явно пострадали, так как новые добавленные данные не совсем отражают распределение исходных данных, на которых и тестировалась модель

Датасет с заменой пропусков на медиану

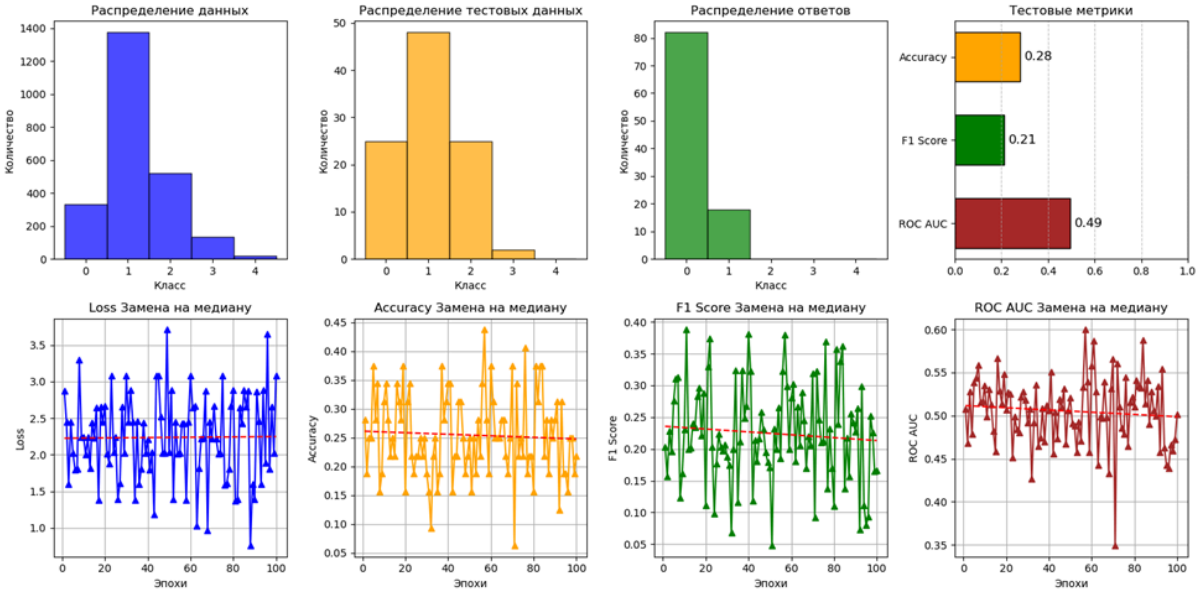

Рис. 22: Графики метрик модели на датасете с заменой на медиану

Результаты обучения модели при использовании замены на медиану показывают схожие результаты с заменой на среднее (*Рисунок 22*)

Заключение

Результаты применения методов одномерного анализа данных показали, что качественная обработка данных данных влияет на показатели метрик модели машинного обучения. Наиболее эффективным подходом оказалось сэмплирование признаков, которое повысило точность и улучшило стабильность обучения

Замена пропущенных значений на среднее и медианное значение показала схожие с семплированием, но более нестабильные результаты обучения модели

Приложение

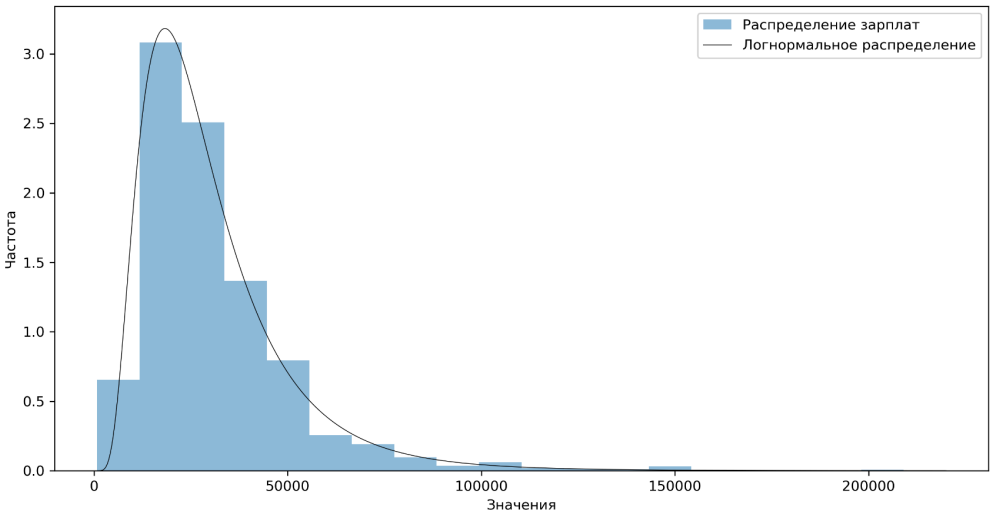

Рис. 3: Логнормальное распределение на исходной плотности распределения заработных плат

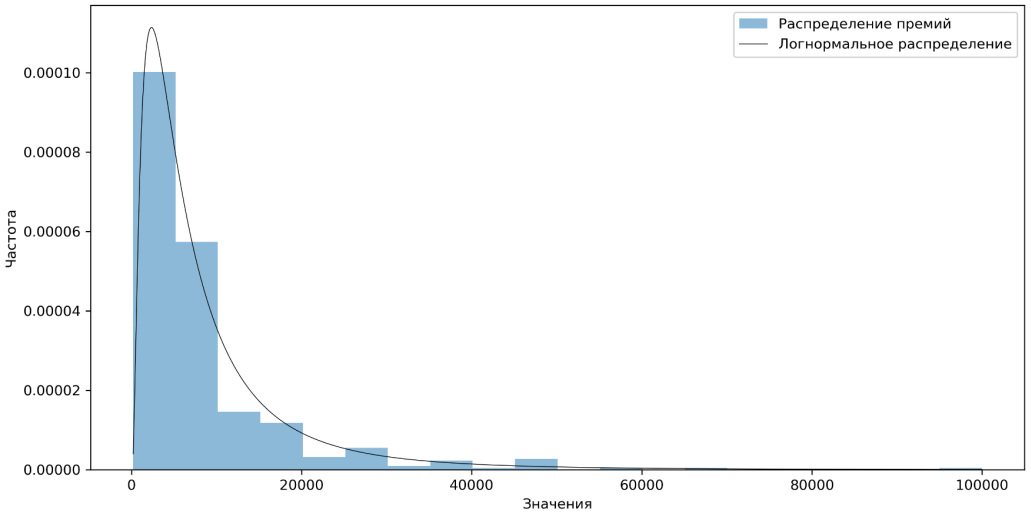

Рис. 4: Логнормальное распределение на исходной плотности распределения премий

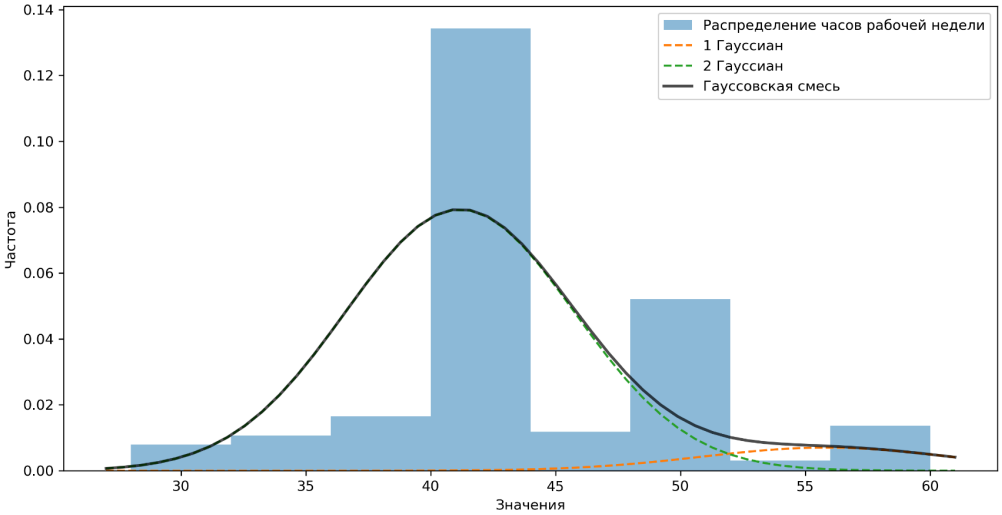

Рис. 9: Смесь гауссианов на плотности распределения часов рабочей недели