Университет науки и технологий МИСИС
Направление «09.04.01 ИНФОРМАТИКА И ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА»
Профиль «Интеллектуальные программные решения для бизнеса»

Отчет о самостоятельной работе по дисциплине «Программная инженерия (Python)»

Бригада № 2:
Лазаренко Д. М., 1 курс, группа МИВТ-22-5
Маковецкий И. А., 1 курс, группа МИВТ-22-5

Москва 2022

Оглавление

1 Общая постановка задачи

1.1 Описание прикладной области и данных

Выбранная прикладная область — «Уровень предлагаемых зарплат технических специалистов в России». Задачей данного исследования является анализ и прогнозирование уровня заработной платы сотрудников по ряду признаков.

Информация, используемая в данном исследовании, была собрана из открытых источников (см. **??**).

В таблице 1 представлено описание фактов, учтенных в анализе.

Таблица 1: Описание фактов, учтенных в анализе

№	Характеристика	Название переменной	Шкала объяснения	Роль
1	Город	city	Номинальная	Объясняющая
2	Долгота	longtitude	Интервальная	Объясняющая
3	Заработная плата	salary	Относительная	Целевая
4	Ключевые навыки	skills	Номинальная	Объясняющая
5	Название	name	Номинальная	Объясняющая
6	Оклад до вычета налогов	gross	Относительная	Объясняющая
7	Опыт	experience	Качественная	Объясняющая
8	Отклик	response_letter_required	Номинальная	Объясняющая
9	Премиум	premium	Номинальная	Объясняющая
10	Тестовое задание	has_test	Номинальная	Объясняющая
11	Широта	latitude	Интервальная	Объясняющая

В анализе присутствуют 6 номинальных переменных, 2 относительных, 2 интервальных и 1 качественная. Зависимая переменная — «Заработная плата».

Приведем описание переменных.

1. Город (city) — название городского округа в России, в котором находится работодатель.
2. Долгота (longitude) — двугранный угол λ между плоскостью меридиана, проходящего через данную точку, и плоскостью начального нулевого меридиана, от которого ведётся отсчёт долготы. Координаты представлены в градусах в виде десятичной дроби.
3. Заработная плата (salary) — средняя величина зарплатной «вилки». При использовании в вакансиях валют, отличных от российского рубля, совершается конвертация в российские рубли (RUB).
4. Ключевые навыки (skills) — информация о наборе ключевых навыков на естественном языке, перечисление представлено в виде строки с разделителем «нижнее подчеркивание».
5. Название (name) — заголовок вакансии на естественном языке.
6. Оклад до вычета налогов (gross) — переменная-признак: оклад указан до вычета налогов.
7. Опыт (experience) — требуемый опыт работы для вакансии.
8. Отклик (response_letter_required) — переменная-признак: обязательно ли заполнять сообщение при отклике на вакансию.
9. Премиум (premium) — переменная-признак: является л/.и данная вакансия премиум-вакансией.

10. Тестовое задание (has_test) — переменная-признак: информации о наличии прикрепленного тестового задания к вакансии.
11. Широта (latitude) — угол ϕ между местным направлением зенита и плоскостью экватора, отсчитываемый от 0° до 90° в обе стороны от экватора. Координаты представлены в градусах в виде десятичной дроби.

1.2 Основные гипотезы, которые планируется провертить в рамках исследования

Для дальнейшего анализа были сформулированы 3 гипотезы о статистической взаимосвязи целевой переменной и объясняющих:

1. Средняя зарплата на вакансиях с требованием знания английского выше, чем без такого требования.
2. При росте опыта работы зарплата разработчиков с навыками JavaScript растет быстрее, чем для разработчиков с навыками 1С.
3. При росте опыта зарплата операторов станка растет быстрее, чем зарплата фрезировщиков.

2 Предварительный анализ собранных данных

2.1 Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы

В ходе выполнения работы была произведена очистка и фильтрация данных.

Переменная «заработная плата» была приведена к единой рублевой шкале. Записи с пропущенными значениями заработной платы и координат были исключены из набора данных.

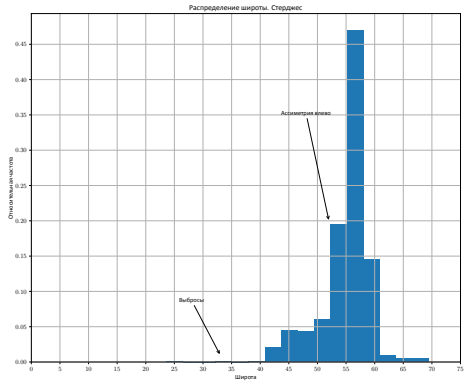
Потенциальная ошибка заключается в неравномерной репрезентации данных — при выделении значений населенных пунктов наблюдается высокое количество уникальных записей (в моногородах и населенных пунктах с малым количеством населения количество вакансий мало). Было принято решение определить по координатам принадлежность вакансии к одному из двенадцати экономических макрорегионов России.[1]

Наоборот, большое количество номинальных признаков, описывающих ключевые навыки вакансии вызывает трудности при их обработке. Было принято решение выделить ключевые навыки в отдельные фиктивные (англ. dummy) признаки.

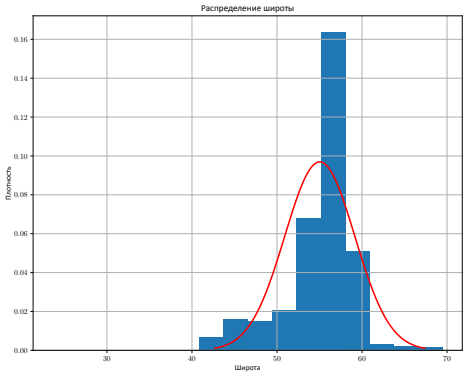
2.1.1 Анализ количественных переменных

К количественным признакам относятся долгота, заработная плата и широта. Построим описательные статистики количественных переменных в таблице 2.

[1] Распоряжение Правительства Российской Федерации от 13 февраля 2019 г. No 207-р.



(a) Распределение по формуле Стерджесса



(b) Распределение широты

Рис. 1: Широта

Таблица 2: Описательные статистики количественных переменных

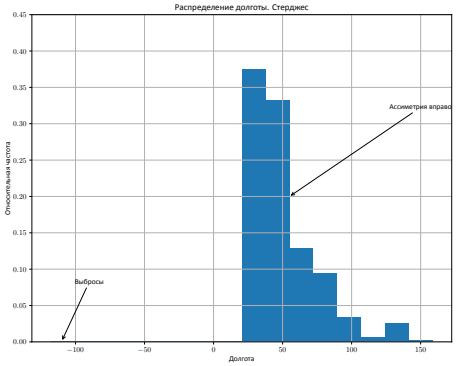
Статистика	Долгота	Заработная плата	Широта
Среднее	50.673 507	68 776.57	55.031 173
Медиана	39.807 944	55 000.00	55.749 451
Минимум	−117.780 920	8700.00	23.608 705
Максимум	158.679 625	522 000.00	69.496 790
Ст. отклонение	23.116 306	44 979.83	4.115 035
Асимметрия	1.800 073	2.41	−0.999 113
Эксцесс	3.544 598	8.885 286	2.253 621
5%	30.307 736	26 100.00	45.049 750
95%	92.981 642	156 600.00	59.971 695
Интерквартильный размах	19.953 478	42 500.00	2.640 537
Пропущенные наблюдения	0	0	0

Построим и проанализируем гистограммы и таблицу 2.

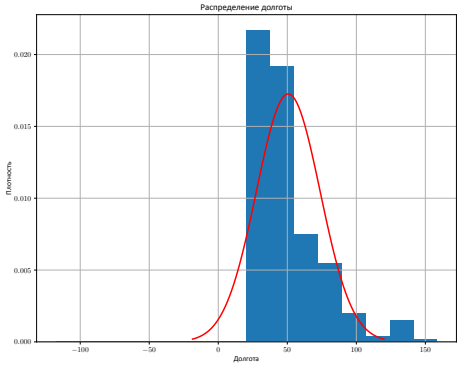
Распределение широты (рис. 1, табл. 2) почти симметрично с ассиметрией влево. Выделим три кластера: градусы широты с 40 по 55, 55 по 60, с 60 по 70. Второй кластер географически расположен между Москвой и Санкт-Петербургом, на этих широтах расположена большая часть российских городов. Кластер «40–50» представлен на юге европейской части России, кластер «60-70» представлен мало из-за климатических условий и малозаселенности. Выбросы в 33 градусе широты связаны с большим количеством вакансий на русском языке в Республике Кипр.

Распределение долготы (рис. 2, табл. 2) асимметрично вправо (среднее значение больше медианы). Большая часть предложений сосредоточено с 30 по 50 градус долготы, что также соответствует распределению населения в европейской части. Два последующих столбца соотносятся с распределением 50–100 градусов — «Урал» и «Сибирь». Малый «всплеск» значений на 130-м градусе соотносится с Дальним Востоком. Выбросы в -100 градусе долготы свидетельствуют о наличии вакансий с удаленным местом работы в США.

Распределение заработной платы (рис. 3, табл. 2) асимметрично вправо (среднее значение больше медианы).

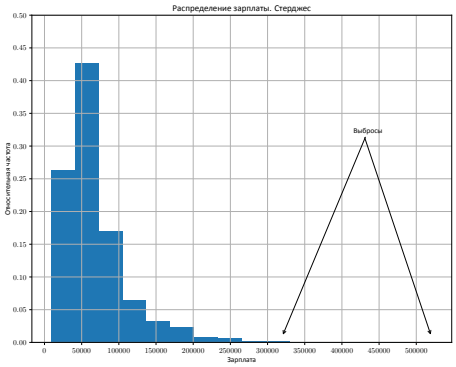


(a) Распределение по формуле Стерджесса

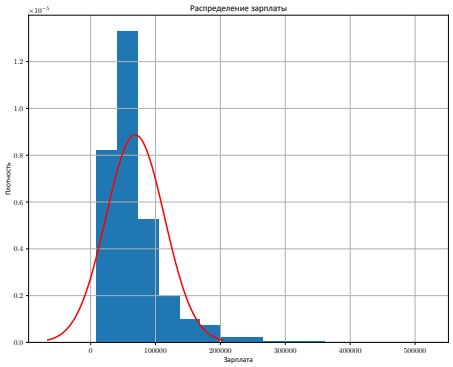


(b) Распределение долготы

Рис. 2: Долгота



(a) Распределение по формуле Стерджесса



(b) Распределение заработной платы

Рис. 3: Заработная плата

Выделим 4 кластера: кластер вакансий для работников с малым доходом (от 8700 до 45 000 рублей), кластер вакансий работников среднего звена (от 45 000 до 68777 рублей), кластер вакансий работников высшего звена (от 68777 до 156 000 рублей) и кластер топ-менеджмента (от 156 000 рублей). Присутствие выбросов в значениях 300 000 и 500 000 связано с наличием заграничных вакансий.

2.1.2 Анализ качественных переменных

В большом количестве (10 тысяч) вакансий город не указан, такие значения заполняются категорией «Неизвестно». Города, представленные в наборе данных меньше трех раз, объединяются в категорию «Другие», в таком случае размер этой категории соотносится с набором данных.

Северный регион был объединен с Уральским по схожему характеру производства, наличия моногородов и малого ($< 5\%$) количества вакансий.

Выборка не является сбалансированной из-за характера размещения большей части

населения в Центральном макрорегионе, остальные регионы представлены равномерно.

График распределения городов по макрорегионам представлен на рисунке 4.

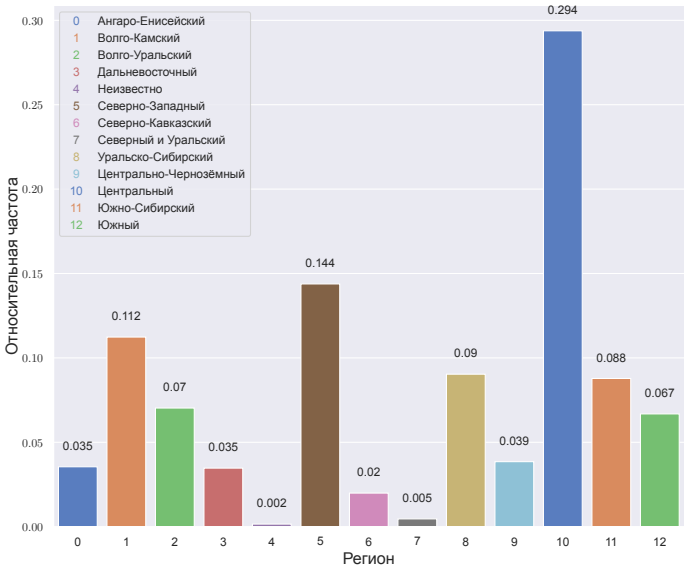


Рис. 4: Распределение городов по макроэкономическим регионам

Рассмотрим две качественных переменных — «Необходимость сопроводительного письма» и «Необходимость прохождения тестового задания».

Графики распределения переменных представлены на рисунках .graphics/response.pdf и .graphics/test.pdf.

Наблюдается крайняя несбалансированность этих переменных — доля вакансий с требованиями сопроводительного письма или прохождения тестового задания не превышает 1 процента.

На рисунке 5 представлена диаграмма распределения опыта работы по годам. Присутствует явное преобладание категории «От 1 года до 3 лет», присутствие категории «более 6 лет» незначительно.

На рисунке 6 представлена диаграмма распределения упоминаний навыков в вакансиях. На данном множестве невозможно создать отношение порядка (кроме частоты упоминания в вакансии), поэтому данные переменные являются фиктивными (dummy).

Из диаграммы видно, что наиболее представленным уровнем является навык владения «1С», что полностью согласуется со статистикой в России. Вторым навыком по количеству вакансий является нетехнический навык «Грамотная речь».

Будем считать навык «мягким» (soft), если в его назывании не представлено имя технологии, в противном случае будем считать его «твердым» (hard, e. g. Linux). В таком случае hard-навыков в распределении 37 процентов.

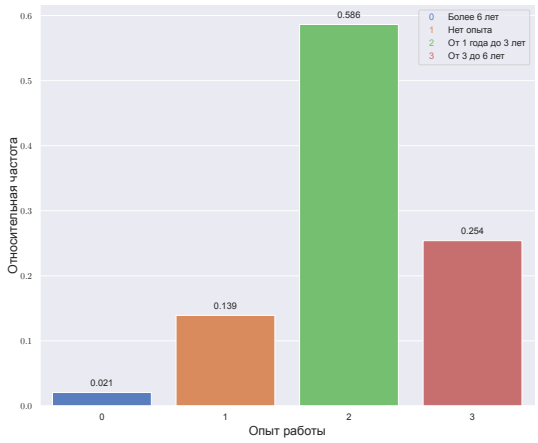


Рис. 5: Распределение опыта работы

2.2 Анализ статистической связи

2.2.1 Графический анализ пары «целевая переменная — качественная объясняющая переменная»

На рисунке 7 представлена категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Наличие тестового задания». Из построенной диаграммы видно, что у вакансий с требованием к наличию тестового задания медиана заработной платы выше. При этом нижнее значение диапазона заработной платы совпадает, а высшее значение также больше у вакансии с требованием к тестовому заданию.

На рисунке 8 представлена категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Наличие письма с откликом». Из построенной диаграммы видно, что у вакансий с требованием к наличию письма с откликом и без него медианы совпадают. Из-за несбалансированности переменных вакансий с требованием письма с откликом меньше.

Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Регион» представлена на рисунке 9.

Из построенной диаграммы видно, что и больший уровень заработной платы, и большее количество вакансий относятся к Центральному макрорегиону. При этом для остальных макрорегионов схожи и медианы, и среднее значение диапазона заработной платы.

Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Опыт» представлена на рисунке 10.

Из построенной диаграммы видно градуальное повышение уровня целевой переменной с повышением опыта, при этом количество вакансий с требованием опыта более 6 лет значительно меньше, чем в остальных категориях.

Категорированная диаграмма для целевой переменной и переменной «Навыки» представлена в файле `../graphics/skills-whiskers.pdf`.

В таблице 3 представлено выполнение непараметрического дисперсионного анализа для пар независимая качественная — зависимая количественная переменная (критерий

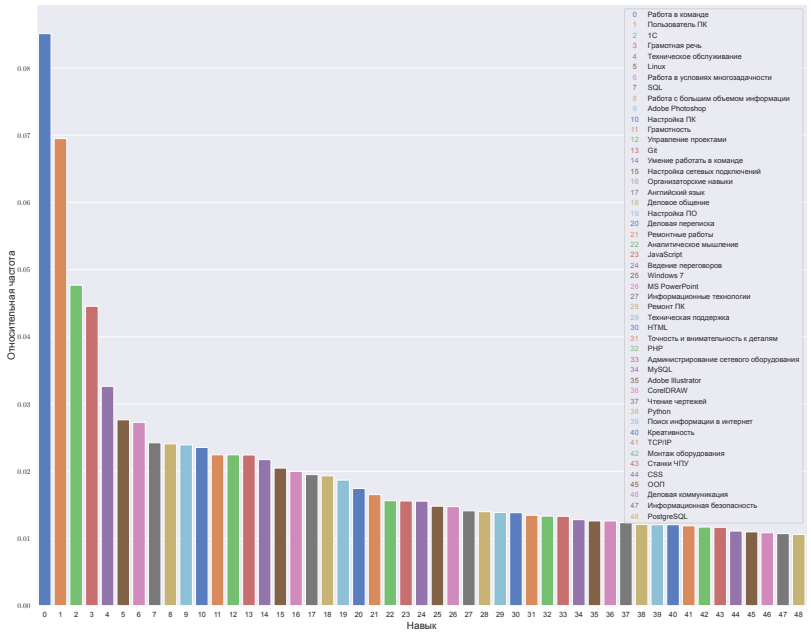


Рис. 6: Частотные упоминания навыков в вакансиях

Крускала-Уоллиса).

Таблица 3: Непараметрический дисперсионный анализ

Независимая переменная	Значимость
Тестовое задание	3.49e-09
Письмо с откликом	0.0009
Регион	0.0
Опыт	0.0
Навыки	0.0

Оценим наличие связи формально с помощью критерия Краскела-Уоллиса: полученное p-value лежат в диапазонах, меньших 5%. Это говорит о том, что все гипотезы об отсутствии связи могут быть отвергнуты на уровне значимости 5%.

2.2.2 Графический анализ пары «числовая зависимая переменная — числовая независимая переменная»

Зависимость диапазона заработной платы от долготы представлена в файле ../graphics-/longtitude-scatter.pdf.

На основании визуального анализа приведенной диаграммы рассеивания, можно сказать, что диапазон заработной платы принимает значения от 0 до 300000 рублей, диапазон долготы принимает значения примерно от 30 до 150 градусов.

Присутствуют незначительные выбросы.

В таблице 4 представлены коэффициенты корелляции для пары «заработная плата — долгота».

Наблюдается отрицательная корелляция средней силы на уровне значимости 0. С ростом долготы незначительно уменьшается количество вакансий, в представленной пред-

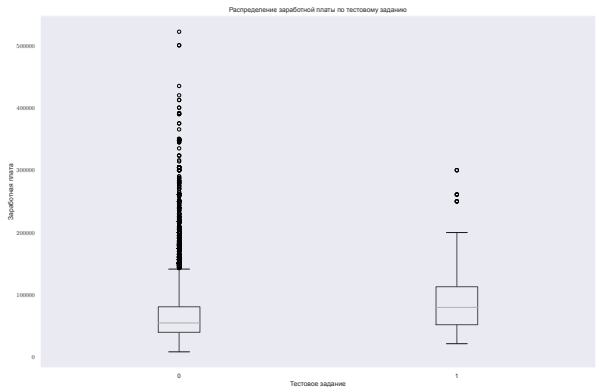


Рис. 7: Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Тестовое задание»

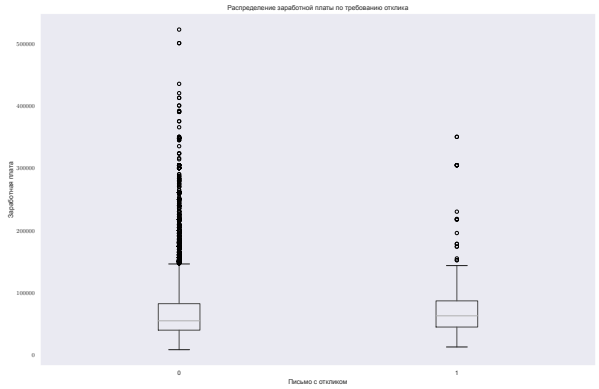


Рис. 8: Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Письмо»

метной области значение долготы не является ключевым для формирования заработной платы.

Таблица 4: Коэффициенты корелляции для пары «заработная плата — долгота»

	Пирсон	Спирмен	Кендалл
Полученное значение	-0.16	-0.29	-0.19
Значимость	0	0	0

Зависимость диапазона заработной платы от широты представлена в файле ../graphics-/latitude-scatter.pdf.

На основании визуального анализа приведенной диаграммы рассеивания, можно сказать, что диапазон заработной платы принимает значения от 0 до 300000 рублей, диапазон долготы принимает значения примерно от 40 до 70 градусов.

Присутствуют незначительные выбросы.

В таблице 5 представлены коэффициенты корелляции для пары «заработная плата — широта».

Наблюдается положительна корелляция малой силы на уровне значимости 0. С ростом широты незначительно увеличивается количество вакансий, в представленной предметной области значение широты не является ключевым для формирования заработной

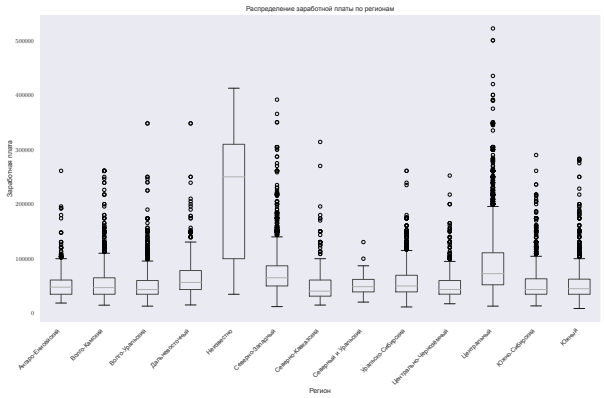


Рис. 9: Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Регион»

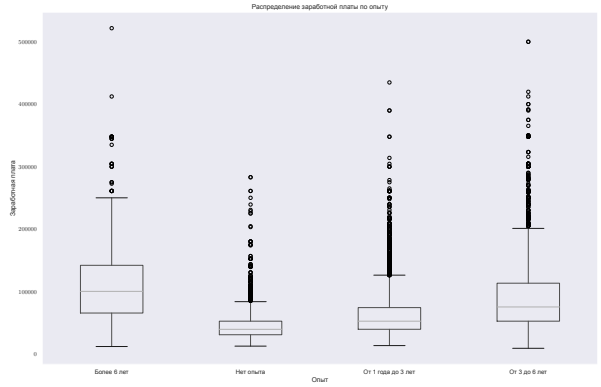


Рис. 10: Категорированная диаграмма Бокса-Уискера для целевой переменной и переменной «Опыт»

платы.

Таблица 5: Коэффициенты корелляции для пары «заработная плата — широта»

	Пирсон	Спирмен	Кендалл
Полученное значение	0.08	0.16	0.1
Значимость	0	0	0

2.2.3 Анализ статистической взаимосвязи между независимыми переменными

Таблица 6: Таблица кросс-табуляции между парой «регион» — «опыт»

experience	> 6	Нет	1–3	3–6	All
Ангаро-Енисейский	6	134	430	145	715
Волго-Камский	31	355	1394	483	2263
Волго-Уральский	19	217	881	299	1416
Дальневосточный	10	133	409	147	699
Неизвестно	3	0	4	24	31
Северно-Западный	81	334	1630	852	2897
Северно-Кавказский	2	79	243	77	401
Северный и Уральский	0	13	57	23	93
Уральско-Сибирский	37	237	1123	423	1820
Центрально-Чернозёмный	9	129	479	159	776
Центральный	175	666	3273	1804	5918
Южно-Сибирский	24	263	1062	420	1769
Южный	18	243	825	260	1346

Критерий HI^2 для переменных «experience»и «Region» = 406.8, df=52, p-value = 0.0.

Статистика Cramer V для переменных «experience» и «Region» — 0.07.

Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

Таблица 7: Таблица кросс-табуляции между парой «наличие теста» — «опыт»

has_test	False	True	All
> 6	408	7	415
Нет	2762	41	2803
1–3	11730	80	11810
3–6	5083	33	5116

Критерий HI^2 для переменных «experience» и «has_test» =23.43, df=8, p-value=0.003. Статистика Cramer V для переменных «experience» и «has_test» = 0.02.

Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

Таблица 8: Таблица кросс-табуляции между парой «наличие письма» — «опыт»

response_letter_required	False	True	All
> 6	408	7	415
Нет	2767	36	2803
1–3	11666	144	11810
3–6	5073	43	5116

Критерий HI^2 для переменных experience и response_lettter_required — 6.34, df=8, pvalue=0.6. Статистика Cramer V для переменных ’experience’ и ’response_letter_required’ — 0.0125. Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

Таблица 9: Таблица кросс-табуляции между парой «регион» — «письмо»

response_letter_required	False	True	All
Ангаро-Енисейский	706	9	715
Волго-Камский	2252	11	2263
Волго-Уральский	1401	15	1416
Дальневосточный	685	14	699
Неизвестно	28	3	31
Северно-Западный	2853	44	2897
Северно-Кавказский	397	4	401
Северный и Уральский	93	0	93
Уральско-Сибирский	1812	8	1820
Центрально-Черноземный	770	6	776
Центральный	5838	80	5918
Южно-Сибирский	1758	11	1769
Южный	1321	25	1346

Критерий HI^2 для переменных Region и response_letter_required — 58.9, df=26, p-value = 0.0002. Статистика Cramer V для переменных Region и response_letter_required — 0.03. Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

Таблица 10: Таблица кросс-табуляции между парой «регион» — «тест»

has_test	False	True	All
Ангаро-Енисейский	715	0	715
Волго-Камский	2250	13	2263
Волго-Уральский	1414	2	1416
Дальневосточный	689	10	699
Неизвестно	31	0	31
Северно-Западный	2871	26	2897
Северно-Кавказский	397	4	401
Северный и Уральский	93	0	93
Уральско-Сибирский	1802	18	1820
Центрально-Черноземный	775	1	776
Центральный	5846	72	5918
Южно-Сибирский	1762	7	1769

Южный	1338	8	1346

Критерий HI^2 для переменных Region и has_test = 40.75, df=26, pvalue=0.03. Статистика Cramer V для переменных Region и has_test — 0.03. Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

Таблица 11: Таблица кросс-табуляции между парой «письмо» — «тест»

has_test	False	True	All
False	19761	153	19914
True	222	8	230

Критерий HI^2 для переменных response_letter_required и 'has_test = 21, df=4, pvalue=0. Статистика Cramer V для переменных response_letter_required и has_test — 0.0228. Из сравнения значений с уровнем 0.05 получаем наличие статистической связи.

2.2.4 Предварительная проверка гипотез

1. Средняя зарплата на вакансиях с требованием знания английского выше, чем без такого требования.
   Для проверки этой гипотезы построим диаграмму Бокса-Вискера для категории английский язык (../../graphics/hyp-1.pdf). И медиана, и 75 квартиль для вакансий с требованием английского языка выше вакансий без такого требования. Проверим равенство медиан с помощью критерия Манна-Уитни. p-value= 1.61e-57 — отвергаем гипотезу о равенстве медиан.
2. При росте опыта работы зарплата разработчиков с навыками JavaScript растет быстрее чем для разработчиков с навыками 1С.
   По графику процентного изменения цены от опыта работы (../../graphics/hyp2.pdf) мы видим, что разработчики Javascript работающие от трех лет имеют большую прибавку к зарплате, чем разработчики 1С.
3. При росте опыта работы зарплата операторов станка растет быстрее, чем зарплата фрезеровщиков.
   По графику процентного изменения цены от опыта работы мы видим, что операторы станка имеют большую прибавку к зарплате, чем фрезеровщики (../../graphics/hyp-3.pdf).

3 Проверка гипотез с помощью моделирования

В качестве базовой модели используется регрессионная модель с линейным включением всех переменных. Выборка разделяется случайным образом на обучающую (80% от общего объема) для построения модели и тестовую (20%) для проверки прогностических свойств.

Уравнение регрессии записывается следующим образом: $salary = a_0 + a_1 * response - letter - required + a_2 * latitude + a_3 * longtitude + a_4 * experience + a_5 * has\_test + a_6 *$

$rabota\_v\_komande + + a_7 * polzovatel\_pk + a_8 * gramotnaya\_rech + a_9 * tekhnicheskoe-obsluzhivanie + a_{10} * linux + a_{11} * rabota - v - usloviyah - mnogozadachnosti + a_{12} * sql + a_{13} * rabota - s - bolshim - obemom - informacii + a_{14} * adobe - photoshop + a_{15} * nastrojka - pk + a_{16} * gramotnost + a_{17} * upravlenie - proektami + a_{18} * git + a_{19} * umenie - rabotat - v - komande + a_{20} * nastrojka - setevyh - podklyuchenij + a_{21} * organizatorskie - navyki + a_{22} * anglijskij - yazyk + a_{23} * delovoe - obshchenie + a_{24} * nastrojka - po + a_{25} * delovaya - perepiska + a_{26} * remontnye - raboty + a_{27} * analiticheskoe - myshlenie + a_{28} * javascript + a_{29} * vedenie - peregovorov + a_{30} * windows - 7 + a_{31} * ms - powerpoint + a_{32} * informacionnye - tekhnologii + a_{33} * remont - pk + a_{34} * tekhnicheskaya - podderzhka + a_{35} * html + a_{36} * tochnost - i - vnimatelnost - k - detalyam + a_{37} * php + a_{38} * administrirovanie - setevogo - oborudovaniya + a_{39} * mysql + a_{40} * adobe - illustrator + a_{41} * coreldraw + a_{42} * chtenie - chertezhej + a_{43} * python + a_{44} * kreativnost + a_{45} * poisk - informacii - v - internet + a_{46} * tcp/ip + a_{47} * montazhoborudovaniya + a_{48} * stankichpu + a_{49} * css + a_{50} * oop + a_{51} * delovayakommunikaciya + a_{52} * informacionnaya - bezopasnost + a_{53} * postgresql + a_{54} * 1c + a_{55} * volgo - kamskij + a_{56} * volgo - uralskij + a_{57} * dalnevostchnyj + a_{58} * neizvestno + a_{59} * severno - zapadnyj + a_{60} * severno - kavkazskij + a_{61} * severnyj - i - uralskij + a_{62} * ural + a_{63} * centralno - chernozyomnyj + a_{64} * centralny + a_{65} * yuzhno - sibirskij + a_{66} * yuzhnyj.$

Созданная нами базовая модель имеет следующую форму:

$salary = 2.381e04 + 63.77 * response - letter - required + 354.17 * latitude + -76.6 * longtitude + 1.888e04 * experience + 1.682e04 * has\_test + a_6 * -2769.8 + + - 7472.07 * polzovatel\_pk + -1948.8254 * gramotnaya\_rech + -1823.3335 * tekhnicheskoe-obsluzhivanie + 3147.7461 * linux + -4556.1398 * rabota - v - usloviyah - mnogozadachnosti + 1.313e04 * sql + -9384.0084 * rabota - s - bolshim - obemom - informacii + -1.001e4 * adobe - photoshop + -5426.2571 * nastrojka - pk + -5604 * gramotnost + 2.299e04 * upravlenie - proektami + 2.617e04 * git + -5332 * umenie - rabotat - v - komande + -4045 * nastrojka - setevyh - podklyuchenij + -825 * organizatorskie - navyki + -1488e04 * anglijskij - yazyk + 37 * delovoe - obshchenie + -83 * nastrojka - po + -8058 * delovaya - perepiska + -7057 * remontnye - raboty + -505 * analiticheskoe - myshlenie + 1.636e04 * javascript + 1.499e04 * vedenie - peregovorov + -8384 * windows - 7 + 3472 * ms - powerpoint + -3802 * informacionnye - tekhnologii + -668 * remont - pk + -8204 * tekhnicheskaya - podderzhka + -7297 * html + -9059 * tochnost - i - vnimatelnost - k - detalyam + 443 * php + -6258 * administrirovanie - setevogo - oborudovaniya + -2936 * mysql + 1676 * adobe - illustrator + -9928 * coreldraw + -2814 * chtenie - chertezhej + 1.92e04 * python + -4704 * kreativnost + -3979 * poisk - informacii - v - internet + -4093 * tcp/ip + -4815 * montazhoborudovaniya + -3046 * stankichpu + 1788 * css + 8632 * oop + 2504 * delovayakommunikaciya + 5143 * informacionnaya - bezopasnost + 4.332e04 * postgresql + 2.572e04 * 1c + -5882 * volgo - kamskij + -9001 * volgo - uralskij + 1.163e04 * dalnevostchnyj + 1.379e05 * neizvestno + 6562 * severno - zapadnyj + -3513 * severno - kavkazskij + = 1.193e04 * severnyj - i - uralskij + -4473 * ural + -5010 * centralno - chernozyomnyj + 2.24e04 * centralny + -3088 * yuzhno - sibirskij + -3513 * yuzhnyj.$

Таблица 12: Характеристика базовой модели

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.381e+04	1.11e+04	2.143	0.032	2032.824	4.56e+04
response_letter_required	63.7748	2700.765	0.024	0.981	-5230.027	5357.577
latitude	354.1698	167.400	2.116	0.034	26.048	682.292
longtitude	-76.6166	71.949	-1.065	0.287	-217.644	64.411
experience	1.884e+04	433.505	43.453	0.000	1.8e+04	1.97e+04
has_test	1.682e+04	3130.090	5.374	0.000	1.07e+04	2.3e+04
Работа в команде	-2769.8075	899.017	-3.081	0.002	-4531.981	-1007.634
Пользователь ПК	-7472.0762	976.885	-7.649	0.000	-9386.880	-5557.272
Грамотная речь	-1948.8254	1267.236	-1.538	0.124	-4432.751	535.100
Техническое обслуживание	-1823.3335	1483.991	-1.229	0.219	-4732.121	1085.454
Linux	3147.7461	1519.750	2.071	0.038	168.866	6126.627
Работа в условиях многозадачности	-4556.1398	1389.469	-3.279	0.001	-7279.654	-1832.626
SQL	1.313e+04	1581.741	8.301	0.000	1e+04	1.62e+04
Работа с большим объемом информации	-9384.0084	1553.360	-6.041	0.000	-1.24e+04	-6339.249
Adobe Photoshop	-1.001e+04	2263.039	-4.421	0.000	-1.44e+04	-5569.226
Настройка ПК	-5426.2571	2362.844	-2.296	0.022	-1.01e+04	-794.818
Грамотность	-5604.8384	1655.487	-3.386	0.001	-8849.779	-2359.898
Управление проектами	2.299e+04	1670.759	13.758	0.000	1.97e+04	2.63e+04
Git	2.617e+04	1803.585	14.513	0.000	2.26e+04	2.97e+04
Умение работать в команде	-5332.4983	1567.917	-3.401	0.001	-8405.791	-2259.206
Настройка сетевых подключений	-4045.4878	2276.349	-1.777	0.076	-8507.387	416.411
Организаторские навыки	-825.8288	1691.633	-0.488	0.625	-4141.618	2489.960
Английский язык	1.488e+04	1631.616	9.122	0.000	1.17e+04	1.81e+04
Деловое общение	37.2880	1836.516	0.020	0.984	-3562.489	3637.065
Настройка ПО	-83.0943	2494.392	-0.033	0.973	-4972.382	4806.194
Деловая переписка	-8058.0872	1883.334	-4.279	0.000	-1.17e+04	-4366.541
Ремонтные работы	-7057.2475	1985.965	-3.554	0.000	-1.09e+04	-3164.534
Аналитическое мышление	-505.8391	1862.126	-0.272	0.786	-4155.813	3144.135
JavaScript	1.636e+04	2356.936	6.939	0.000	1.17e+04	2.1e+04
Ведение переговоров	1.499e+04	1946.045	7.704	0.000	1.12e+04	1.88e+04
Windows 7	-8384.4998	1995.272	-4.202	0.000	-1.23e+04	-4473.544
MS PowerPoint	3472.5188	1940.879	1.789	0.074	-331.821	7276.858
Информационные технологии	-3802.5038	2002.862	-1.899	0.058	-7728.336	123.329
Ремонт ПК	-668.6550	2286.822	-0.292	0.770	-5151.082	3813.772
Техническая поддержка	-8204.7166	1988.039	-4.127	0.000	-1.21e+04	-4307.937
HTML	-7297.5838	2633.187	-2.771	0.006	-1.25e+04	-2136.244
Точность и внимательность к деталям	-9059.2966	1987.923	-4.557	0.000	-1.3e+04	-5162.744
PHP	443.7024	2744.997	0.162	0.872	-4936.799	5824.204
Администрирование сетевого оборудования	-6258.6186	2190.250	-2.857	0.004	-1.06e+04	-1965.483
MySQL	-2936.2935	2701.090	-1.087	0.277	-8230.733	2358.146
Adobe Illustrator	1676.3743	2793.291	0.600	0.548	-3798.789	7151.537
CorelDRAW	-9928.6038	2505.197	-3.963	0.000	-1.48e+04	-5018.138
Чтение чертежей	-2814.9596	2086.178	-1.349	0.177	-6904.102	1274.183
Python	1.92e+04	2195.237	8.744	0.000	1.49e+04	2.35e+04
Креативность	-4704.5474	2218.276	-2.121	0.034	-9052.616	-356.479
Поиск информации в интернет	-3979.9981	2127.929	-1.870	0.061	-8150.978	190.981
TCP/IP	-4093.8090	2233.319	-1.833	0.067	-8471.363	283.745
Монтаж оборудования	-4815.1528	2176.876	-2.212	0.027	-9082.073	-548.232
Станки ЧПУ	-3046.3091	2165.566	-1.407	0.160	-7291.061	1198.443
CSS	1788.0801	3118.025	0.573	0.566	-4323.597	7899.757
ООП	8632.4211	2362.950	3.653	0.000	4000.775	1.33e+04
Деловая коммуникация	2504.5506	2296.775	1.090	0.276	-1997.384	7006.485
Информационная безопасность	5143.1524	2299.105	2.237	0.025	636.649	9649.656
PostgreSQL	4.342e+04	2369.081	18.329	0.000	3.88e+04	4.81e+04
1С	2.572e+04	1076.384	23.894	0.000	2.36e+04	2.78e+04
Region_Волго-Камский	-5882.6667	3784.355	-1.554	0.120	-1.33e+04	1535.092
Region_Волго-Уральский	-9001.1175	3726.532	-2.415	0.016	-1.63e+04	-1696.698
Region_Дальневосточный	1.163e+04	3545.818	3.281	0.001	4683.197	1.86e+04
Region_Неизвестно	1.379e+05	9283.300	14.850	0.000	1.2e+05	1.56e+05
Region_СеверноЗападный	6562.2847	5147.038	1.275	0.202	-3526.485	1.67e+04
Region_СеверноКавказский	-3513.6970	4957.794	-0.709	0.479	-1.32e+04	6204.134
Region_Северный и Уральский	-1.193e+04	6136.383	-1.944	0.052	-2.4e+04	101.558
Region_Уральско-Сибирский	-4473.9594	3026.909	-1.478	0.139	-1.04e+04	1459.121
Region_Центрально-Чернозёмный	-5010.3261	4634.894	-1.081	0.280	-1.41e+04	4074.585
Region_Центральный	2.24e+04	4498.973	4.978	0.000	1.36e+04	3.12e+04
Region_Южно-Сибирский	-3088.4704	2016.626	-1.532	0.126	-7041.283	864.342
Region_Южный	-3513.7833	4534.097	-0.775	0.438	-1.24e+04	5373.554

Проведем анализ мультиколлениарности:

Таблица 13: Таблица значений для проверки мультиколлениарности базовой модели

	VIF Factor	features
0	5.975267934359127	const
14	1.8730699375079103	Настройка сетевых подключений
9	1.7893377630263387	Настройка ПК

Продолжение на следующей странице

Таблица 13 — *Продолжение*

	VIF Factor	features
8	1.630735772775094	Adobe Photoshop
25	1.511587122222877	CorelDRAW
18	1.4526987064386938	JavaScript
12	1.3626208357887688	Git
22	1.2932659427692192	HTML
24	1.223018368744047	Администрирование сетевого оборудования
37	1.194345606157767	Region_Центральный
6	1.1627224160281804	SQL
20	1.1565884489171454	Windows 7
27	1.1545140324888308	Креативность
29	1.1520825274943565	ООП
19	1.1518778013433573	Ведение переговоров
36	1.1439322140528216	Region_Северно-Западный
31	1.1376856476316821	PostgreSQL
4	1.1375555765868899	Пользователь ПК
3	1.1309665461165916	Работа в команде
11	1.1282050087401703	Управление проектами
26	1.1246820136986657	Python
21	1.082907347590384	Техническая поддержка
16	1.0800909567520036	Деловая переписка
33	1.0777558566479093	Region_Волго-Уральский
10	1.0615420242469884	Грамотность
1	1.0581677105698448	experience
17	1.0487439508791838	Ремонтные работы
7	1.0464349350705453	Работа с большим объемом информации
34	1.0447090661958827	Region_Дальневосточный
13	1.0431666011584064	Умение работать в команде
28	1.0365801934793484	Монтаж оборудования
32	1.0326565219448112	1С
15	1.0291915485960264	Английский язык
5	1.0186071879930594	Работа в условиях многозадачности
35	1.018313397453688	Region_Неизвестно
30	1.0178685417055608	Информационная безопасность
23	1.0171909797404994	Точность и внимательность к деталям
2	1.0080961297658366	has_test

Все значение меньше 10 — мультиколлинеарности нет. Проведем анализ гетероскадастичности:

```
'Test Statistic': 3798.8382212678603,
'Test Statistic p-value': 1.1589080013261682e-158,
'F-Statistic': 2.5862894216339507,
'F-Test p-value': 7.147080738627432e-195
```

Мы принимаем гипотезу о том, что в регрессионной модели присутствует гетероскедастичность.

В данной модели присутствуют незначимые коэффициенты. Проведем пошаговое исключение переменных из модели, на первом шаге исключая переменную с самым большим p-value. Повторим данную операцию до тех пор, пока в модели не останутся только значимые переменные. Проведем оптимизацию базовой модели.

Удаляем колонку Region_Уральско-Сибирский. Обучаем модель на оставшихся колонках:

Промежуточные результаты:

```
R-squared: 0.3855033832154069
Adj. R-squared: 0.38359073189324366
AIC: 383499.91852490103
```

Тест Уайта на гетероскедастичность промежуточной модели:

```
'Test Statistic': 2748.1609680923743,
'Test Statistic p-value': 4.717364833164232e-161,
'F-Statistic': 3.081075901396072,
'F-Test p-value': 1.2380667171275686e-184}
```

Таблица 14: Характеристика значений промежуточной модели

Коэффициент	Значение
const	23585.43
latitude	207.42
experience	18816.88
has$_test$	16843.27
Работа в команде	-2747.73
Пользователь ПК	-7532.77
Грамотная речь	-1875.06
Техническое обслуживание	-1896.02
Linux	3040.38
Работа в условиях многозадачности	-4518.61
SQL	13130.89
Работа с большим объемом информации	-9482.87
Adobe Photoshop	-9190.66
Настройка ПК	-5702.63
Грамотность	-5694.59
Управление проектами	22954.32
Git	25901.32
Умение работать в команде	-5369.27
Настройка сетевых подключений	-4084.21
Английский язык	14921.12
Деловая переписка	-7841.44
Ремонтные работы	-7000.52
JavaScript	16413.18
Ведение переговоров	15066.13
Windows 7	-8302.55
MS PowerPoint	3420.36
Информационные технологии	-3827.91
Техническая поддержка	-8165.31
HTML	-6715.47
Точность и внимательность к деталям	-9033.16
Администрирование сетевого оборудования	-6326.02
CorelDRAW	-9793.12
Чтение чертежей	-2864.61
Python	19218.64
Креативность	-4553.24
Поиск информации в интернет	-3997.09
TCP/IP	-4074.98
Монтаж оборудования	-4857.16
Станки ЧПУ	-3084.61
ООП	8245.44
Информационная безопасность	5238.66
PostgreSQL	42954.13
1С	25719.07
Region_Волго-Камский	-1129.11
Region_Волго-Уральский	-4787.70
Region_Дальневосточный	8976.92
Region_Неизвестно	142555.42
Region_Северно-Западный	13235.73
Region_Северный и Уральский	-5605.85
Region_Центральный	27929.31
Region_Южно-Сибирский	-1121.21

Приведем уравнение итоговой (оптимальной модели): $salary = a_0 + a_1 * experience + a_2 * has\_test + a_3 * rabota\_v\_komande + a_4 * polzovatel\_pk + a_5 * rabota - v - usloviyah - mnogozadachnosti + a_6 * sql + a_7 * rabota - s - bolshim - obemom - informacii + a_8 * adobe - photoshop + a_9 * nastrojka - pk + a_{10} * gramotnost + a_{11} * upravlenie - proektami + a_{12} * git + a_{13} * umenie - rabotat - v - komande + a_{14} * nastrojka - setevyh - podklyuchenij + a_{15} * anglijskij - yazyk + a_{16} * delovaya - perepiska + a_{17} * remontnye - raboty + a_{18} * javascript + a_{19} * vedenie - peregovorov + a_{20} * windows - 7 + a_{21} * tekhnicheskaya - podderzhka + a_{22} * html + a_{23} * tochnost - i - vnimatelnost - k - detalyam + a_{24} *$

$administrirovanie - setevogo - oborudovaniya + a_{25} * coreldraw + a_{26} * kreativnost + a_{27} * montazhoborudovaniya + a_{28} * oop + a_{29} * informacionnaya - bezopasnost + a_{30} * postgresql + a_{31} * 1c + a_{32} * volgo - uralskij + a_{33} * dalnevostchnyj + a_{34} * neizvestno + a_{35} * severno - zapadnyj + a_{36} * centralny.$

Получаем модель:

$salary = 33892.94 + 18911.98 * experience + 16543.28 * has\_test + -3084.26 * rabota\_v\_komande + -8240.52 * polzovatel\_pk + -4425.73 * rabota - v - usloviyah - mnogozadachnosti + 13250.32 * sql - 9381.83 * rabota - s - bolshim - obemom - informacii - 8858.02 * adobe - photoshop - 5839.54 * nastrojka - pk - 6855.32 * gramotnost + 23021.44 * upravlenie - proektami + 26478.97 * git - 5295.17 * umenie - rabotat - v - komande - 4784.70 * nastrojka - setevyh - podklyuchenij + 15107.03 * anglijskij - yazyk - 7746.76 * delovaya - perepiska - 7796.84 * remontnye - raboty + 16403.87 * javascript + 14948.63 * vedenie - peregovorov - 8441.23 * windows - 7 - 9034.00 * tekhnicheskaya - podderzhka - 6724.48 * html - 9529.90 * tochnost - i - vnimatelnost - k - detalyam - 6402.83 * administrirovanie - setevogo - oborudovaniya - 9604.04 * coreldraw + 20104.09 * kreativnost - 4592.28 * montazhoborudovaniya - 5067.31 * oop + 8606.42 * informacionnaya - bezopasnost + 4207.38 * postgresql + 43452.98 * 1c + 25790.67 * volgo - uralskij - 4311.08 * dalnevostchnyj + 8458.31 * neizvestno + 140442.12 * severno - zapadnyj + 15031.77 + 28881.73 * centralny.$

3.1 Проверка гипотез

1-я Гипотеза: Средняя зарплата на вакансиях с требованием знания английского выше, чем без такого требования.

Видим положительный коэффициент у переменной Английский язык. Это означает, что зарплата для вакансий с требованием знания английского языка на 15 тыс. рублей выше, чем без.

2-я Гипотеза: При росте опыта работы зарплата разработчиков с навыками JavaScript растет быстрее чем для разработчиков с навыками 1С.

В таком случае коэффициент a(”опыт работы”) = c1 + b1 * JavaScript + b2 * 1С.

Тогда модель будет иметь следующий вид: ... + c1 * experience + b1 * JavaScript * experience + b2 * experience * 1С ...

Мы видим, что коэффициент перед переменной опыт * Javascript больше чем опыт * 1С. Данные переменные являются статистически значимыми. Это подтверждает нашу гипотезу (см. ноутбук Hypothesis).

3-я Гипотеза: При росте опыта работы зарплата операторов станка растет быстрее, чем зарплата фрезеровщиков

В таком случае коэффициент a(”опыт работы”) = c1 + b1 * Фрезеровщик + b2 * Оператор станка.

Тогда модель будет иметь следующий вид: ... + c1 * experience + b1 * Фрезеровщик * experience + b2 * experience * Оператор станка ...

Мы видим, что коэффициент перед переменной опыт * Фрезеровщик меньше чем опыт * Оператор станка. Данные переменные являются статистически значимыми. Это

подтверждает нашу гипотезу.

4 Заключение

Мы смогли подтвердить с помощью моделирования несколько гипотез, которые не были очевидны на этапе предварительного анализа данных в связи со спецификой датасета. В процессе проверки прогнозируемой способности модели удалось достичь наилучших значений метрик в процессе работы над оптимизацией модели, исключив признаки, наиболее сильно зависящие