

**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

<b>1</b>	<b>Общая постановка задачи</b>	<b>3</b>
1.1	Описание прикладной области и данных . . . . .	3
1.2	Основные гипотезы, которые планируется проверить в рамках исследования	3
<b>2</b>	<b>Предварительный анализ собранных данных</b>	<b>4</b>
2.1	Анализ особенностей данных: потенциальные ошибки и пропущенные значения, группы и выбросы . . . . .	4
2.1.1	Анализ количественных переменных . . . . .	4
2.1.2	Анализ качественных переменных . . . . .	4
2.2	Анализ статистической связи . . . . .	5
2.2.1	Графический анализ пары «целевая переменная — качественная объясняющая переменная» . . . . .	5
2.2.2	Графический анализ пары «числовая зависимая переменная — числовая независимая переменная» . . . . .	5
2.2.3	Анализ статистической взаимосвязи между независимыми переменными . . . . .	5
2.2.4	Предварительная проверка гипотез . . . . .	5
<b>3</b>	<b>Проверка гипотез с помощью моделирования</b>	<b>5</b>
3.1	Построение базовой модели . . . . .	6
3.2	Проверка гипотез с помощью моделирования . . . . .	6
3.3	Оптимизация итоговой модели, сравнение качества моделей . . . . .	6
3.4	Проверка прогностических способностей модели . . . . .	7
3.5	Диагностика регрессионной модели . . . . .	7
<b>4</b>	<b>Заключение</b>	<b>8</b>
	<b>Приложение А</b>	<b>9</b>

# 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. Приложение А).

В таблице 1 представлено описание фактов, учтенных в анализе.

Таблица 1: Описание фактов, учтенных в анализе

№	Характеристика объекта/явления	Название переменной	Шкала объяснения	Роль: целевая/объясняющая
1	Заработная плата	salary_from, salary_to	Порядковая	Целевая
2	Адрес места работы	coordinates	Номинальная	Объясняющая
3	Сопроводительное письмо	response_letter	Отношений	Объясняющая
4	Город	city	Номинальная	Объясняющая
5	Широта	longitude	Номинальная	Объясняющая
6	Долгота	latitude	Номинальная	Объясняющая
7	Необработанный адрес	raw	Номинальная	Объясняющая
8	Опыт	experience	Порядковая	Объясняющая
9	Время работы	schedule, employment	Номинальная	Объясняющая
10	Ключевые навыки	skills	Номинальная	Объясняющая
11	Проверенный работодатель	has_test	Отношений	Объясняющая
12	Заработная плата до вычета	gross	Порядковая	Объясняющая
13	Валюта	currency	Номинальная	Объясняющая
14	Премиум-аккаунт	premium	Номинальная	Объясняющая

В анализе присутствуют 9 номинальных переменных, 3 порядковых и 1 отношений. Зависимая переменная — «Заработная плата».

## 1.2 Основные гипотезы, которые планируется проверить 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 механиков растет быстрее, чем токарей.

## 2 Предварительный анализ собранных данных

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

#### 2.1.1 Анализ количественных переменных

Здесь необходимо построить и проанализировать гистограммы для всех количественных (интервальных и относительных) переменных в анализе. Необходимо охарактеризовать вид распределения по отношению к нормальному распределению — асимметрию, эксцесс, полимодальность. Для этого следует привести график гистограммы совместно с графиком плотности нормального распределения, а также таблицу основных статистик.

Таблица 2: Описание фактов, учтенных в анализе

Статистика	Значение
Среднее	
Медиана	
Стандартное отклонение	
Межквартильный размах	
Верхняя квартиль	
Нижняя квартиль	
Коэффициент асимметрии	
Коэффициент эксцесса	
Количество наблюдений	
Количество пропущенных значений	

Необходимо дать интерпретацию статистических свойств количественных переменных в контексте предметной области. Например, на основании гистограммы и числовых характеристик распределения можно сделать вывод о наличии небольшого количества субъектов федерации с очень большой долей бедного населения. Также, для целевой переменной следует проанализировать наличие выбросов на основании правила «трех-сигм». Следует отметить в базе все выбросы и на основании сравнения соответствующих значений объясняющих переменных с их средними или медианными значениями объяснить, почему эти наблюдения могут интерпретироваться как выбросы.

### **2.1.2 Анализ качественных переменных**

Здесь следует привести столбчатые диаграммы, которые отражают количество измерений с разными уровнями для данной переменной.

Необходимо проанализировать степень представленности всех уровней и при необходимости (наличии уровней с долей менее 5 %) произвести укрупнение уровней.

Результат привести на новых диаграммах. Принцип укрупнения пояснить.

## **2.2 Анализ статистической связи**

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

Здесь для каждой пары (количественная зависимая переменная — качественная независимая переменная) необходимо построить категоризованную диаграмму Бокса- Уискера (Box-Whisker).

На основании анализа диаграммы следует охарактеризовать связь среднего значения и разброса количественной зависимой переменной с уровнями качественной независимой переменной. Интерпретацию дать в контексте предметной области.

Для формальной проверки гипотезы о наличии статистической связи следует выполнить непараметрический дисперсионный анализ (критерий Крускала-Уоллиса)

### **2.2.2 Графический анализ пары «числовая зависимая переменная – числовая независимая переменная»**

Здесь для каждой пары (количественная зависимая переменная – количественная независимая переменная) необходимо построить диаграммы рассеивания (Scatter plot).

На основании визуального анализа диаграммы следует сделать предположение о наличии и характере статистической взаимосвязи. Интерпретацию результатов дать в контексте предметной области.

Для формальной проверки гипотезы о наличии связи следует подсчитать коэффициенты корреляции Пирсона и Спирмена, а также тау Кендала и привести результаты проверки их значимости.

### **2.2.3 Анализ статистической взаимосвязи между независимыми переменными**

Следует проанализировать силу связи между независимыми переменными, используя инструменты пп. 3.2.1 и 3.2.2. Для анализа силы связи между качественными переменными следует использовать анализ таблиц: необходимо привести таблицу кросс-табуляции, значения статистики хи-квадрат и V-Крамера.

### **2.2.4 Предварительная проверка гипотез**

Здесь необходимо рассказать о результатах проверки гипотез из п.1.3 на основании предварительного анализа данных.

### 3 Проверка гипотез с помощью моделирования

Данный раздел предполагает проверку прогностических способностей построенной модели. В связи с этим исходную выборку следует случайным образом разделить на обучающую и тестовую в пропорции 80:20. На обучающей выборке будет осуществляться построение моделей, тестовая выборка будет использоваться для проверки прогностических способностей.

#### 3.1 Построение базовой модели

Базовая модель служит для анализа изменения качества моделирования при учете сформулированных гипотез. В качестве базовой модели следует использовать модель линейной регрессии целевой переменной на все объясняющие. Для базовой модели следует проверить значимость всех объясняющих переменных, а также уровень мультиколлинеарности (показатель VIF) и наличие гетероскедастичности (критерий Уайта). Исходная базовая модель и результаты ее анализа включаются в отчет.

Далее необходимо оптимизировать структуру модели для повышения ее качества и возможного снижения уровня мультиколлинеарности. Для этого следует пошагово удалять незначимые переменные, переоценивая модель после каждого удаления. Необходимо также пошагово удалять переменные, которые демонстрируют высокую взаимосвязь с другими переменными ( $VIF > 3$ ). Оценку мультиколлинеарности и гетероскедастичности следует выполнять на каждом шаге оптимизации. В отчете следует привести один промежуточный и итоговый вариант, который не содержит незначимых объясняющих переменных и имеет удовлетворительный уровень мультиколлинеарности. Следует привести оценку мультиколлинеарности вошедших в модель переменных и оценку наличия гетероскедастичности. Необходимо также привести оценку качества полученной модели (критерий Akaike, R-sq и adjusted R-sq).

В ходе оптимизации следует оставить в модели объясняющие переменные, которые необходимы для проверки гипотез даже, если они незначимы или имеют высокое значение показателя VIF. Это следует отметить в отчете.

#### 3.2 Проверка гипотез с помощью моделирования

Для проверки выдвинутых в п. 1.2. сложных гипотез выполняется модификация оптимизированной базовой модели поэтапно для каждого сочетания сформулированных гипотез. Сначала модифицируют базовую модель для каждой сложной гипотезы отдельно, далее для всевозможных пар и т.д. Для простых гипотез модификация не требуется. Модифицированные модели оцениваются и выполняется проверка как сложных, так и простых гипотез. Методология проверки каждой гипотезы должна быть описана в отчете в виде ограничений на коэффициенты и пары статистических гипотез. Результаты использования каждой модифицированной модели включаются в отчет.

Модель, которая учитывает все сформулированные гипотезы объявляется итоговой.

### 3.3 Оптимизация итоговой модели, сравнение качества моделей

Итоговая модель подвергается оптимизации за счет пошагового удаления незначимых переменных. На каждом шаге модель переоценивается. Для финального варианта оценивается качество модели с использованием критерия Akaike и adjusted R-sq. Оптимизированная итоговая модель и результаты ее анализа включаются в отчет.

По результатам работы формируется таблица с перечнем моделей включенных в отчет и оценками их качества — значениями критерия Akaike, R-sq и adjusted R-sq

Таблица 3: Сравнение качества построенных моделей

Номер или критерий	$R^2$	$Adj \setminus R^2$	Akaike
1			
2			
3			

### 3.4 Проверка прогностических способностей модели

Проверка прогностических способностей осуществляется для всех включенных в отчет моделей. Необходимо подсчитать значения прогнозов для элементов тестовой выборки и построить для них центральные доверительные интервалы на основе нормального распределения для доверительной вероятности 95%. Для результатов следует рассчитать среднеквадратическую погрешность прогнозирования и максимальную абсолютную погрешность прогнозирования, а также эмпирическую оценку доверительной вероятности. Результаты следует представить в виде таблицы

Таблица 4: Сравнение прогностических способностей моделей

Номер или критерий	Среднеквадратическая погрешность	Абсолютная погрешность	Доверительная вероятность
1		0-10	Целевая
2			
3			

Таблицу следует прокомментировать, в частности, оценку доверительной вероятности. Результаты, представленные в таблице, следует сопоставить с оценками качества данных моделей.

### 3.5 Диагностика регрессионной модели

Для оптимизированной базовой модели и для оптимизированной итоговой модели необходимо выполнить поиск:

- точек разбалансировки с помощью hat-value;
- выбросов, с помощью студентизированных остаточных разностей;
- измерений сильно влияющих на оценки коэффициентов, с помощью расстояния Кука.

Необходимо сравнить полученные множества для двух моделей и выделить измерения, которые входят в указанные множества для обеих моделей.

Также, необходимо проанализировать несколько точек (две — три) с аномальными значениями расстояния Кука для оптимизированной итоговой модели. Следует установить, входят ли они в множества точек разбалансировки и выбросов, а также проанализировать, чем это объясняется. Для этого следует сравнить значения объясняющих и целевой переменных с средними значениями по всей выборке.

## 4 Заключение

В данном разделе следует перечислить результаты проверки сформулированных гипотез в различных сочетаниях, проверки прогностических способностей моделей и их диагностики.



## Приложение А

В РАЗРАБОТКЕ