ФГБОУ ВО «СЕВЕРО-ОСЕТИНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ К.Л. ХЕТАГУРОВА»

Факультет математики и компьютерных наук
Кафедра прикладной математики и информатики

ОТЧЕТ ПО ПРАКТИКЕ

Наименование практики: Производственная (технологическая (проектно-технологическая)) практика

Выполнил студент Дзуцева Лана Сергеевна
(фамилия, имя, отчество)

направление подготовки 01.03.02 Прикладная математика и информатика,

профиль «Программирование, анализ данных и математическое моделирование»

Подпись студента: _____ Дата сдачи отчета: «30» мая 2025 г.

Отчет принят: _____ Тотрова М.Х., старший преподаватель каф. ПМиИ
подпись *Ф.И.О. ответственного лица, должность*

«30» мая 2025 г.

Оценка _____ _____/ Тотрова М.Х._____
подпись / Ф.И.О. преподавателя-экзаменатора

«___» _____ 2025 г.

Содержание

Введение

Целью Производственной (технологической (проектно-технологической)) практики (далее – Практика) является профессионально-компетентностная подготовка обучающихся к самостоятельной профессиональной деятельности посредством формирования навыков и иных компетенций, опыта самостоятельной профессиональной деятельности в реальных условиях. А также:

- закрепление, углубление и систематизация знаний, полученных при изучении дисциплин профессионального цикла;
- развитие имеющихся и приобретение новых профессиональных умений и навыков; развитие сформированных и формирование новых компетенций по избранной профессиональной деятельности;
- развитие опыта организационной работы, повышение мотивации к профессиональному самосовершенствованию;
- проверка умения обучающихся работать с информационными технологиями;
- укрепление связи обучения с практической деятельностью.

Задачи Практики: приобретение опыта самостоятельной работы в сфере будущей профессиональной деятельности.

Место прохождения практики: ООО М-Софтер.

Форма проведения Практики: непрерывно.

Способ проведения Практики: стационарно.

Период прохождения практики: с 02.05.2025 по 30.05.2025 включительно (4 недели).

Руководство практикой осуществлялось старшим преподавателем каф. ПМиИ Тотровой М.Х.

Прохождение практики осуществлялось в соответствии с рабочим графиком (см. приложение 1) и индивидуальным заданием (см. приложение 2).

Краткое содержание и цель индивидуального задания:

1. Сбор данных и подготовка датасета
2. Выбор оптимального алгоритма и фреймворка для работы.
3. Обучение модели машинного обучения

Перечень выполненных работ и заданий. В соответствии с индивидуальным заданием за период прохождения практики выполнена следующая работа:

1. Собран датасет, состоящий из изображений оконных проёмов и подготовлен для обучения нейронной сети
2. Выбрана оптимальная архитектура нейронной сети и фреймворк для работы
3. Обучены и протестированы 2 модели машинного обучения

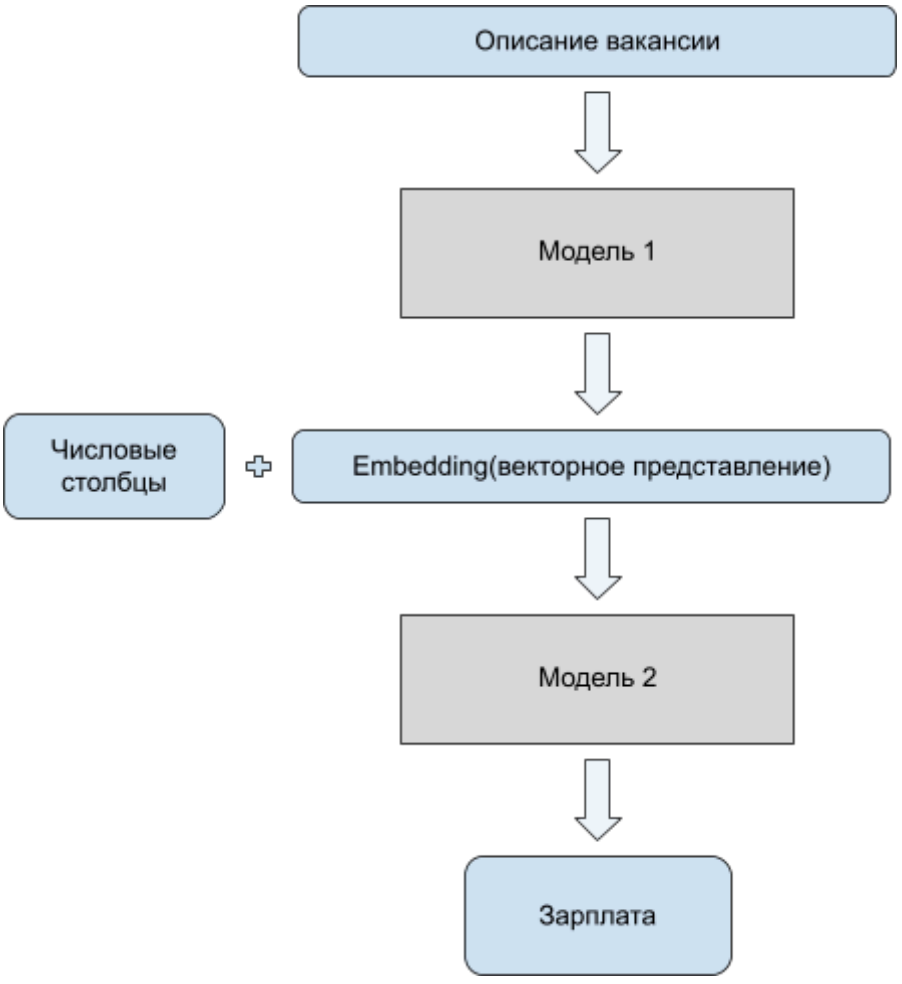
Все виды деятельности в период прохождения практики отражены в Дневнике практики (см. приложение 3).

Прогнозирование зарплатной вилки на основе текстового описания вакансии с использованием методов машинного обучения

Во время практики решается задача создания нейронной сети которая будет предсказывать зарплатную вилку на основе текстового описания вакансии. Для разработки данного решения необходимо решить ряд задач, связанных с обработкой естественного языка и машинным обучением. Из-за разнообразия форматов описаний вакансий и необходимости обеспечения высокой точности предсказаний это представляет собой сложную техническую задачу.

Общая схема работы всего алгоритма видна на следующем рисунке.

рис.1



В рамках практики была проделана следующая работа:

1. Обзор и сравнение различных архитектур нейронных сетей.
2. Сбор и разметка датасета для обучения и тестирования модели.
3. Выбор наиболее подходящей архитектуры нейросети и фреймворка для работы на мобильных устройствах

Сбор и подготовка данных

Данные для обучения были собраны следующим образом. Используя ЯП python и открытое апи hh.ru было собрано около 17000 вакансий по 8 основным направлениям:

1. C#-разработчик
2. Flutter-разработчик
3. Python-разработчик
4. Системный администратор
5. разработчик 1C
6. ML инженер
7. Веб-разработчик
8. Golang-разработчик

Для хранения данных была выбрана база DuckDB, которая является высокопроизводительной аналитической СУБД. DuckDB запускается на Linux, macOS, Windows, Android, iOS и всех других популярных платформах. Она бесплатна для использования и имеет апи для множества языков программирования.

Из модели вакансии были выделены следующие поля, которые представляют важность для аналитики и обучения модели:

```
id_vacancy TEXT NOT NULL,
name TEXT NOT NULL,
premium BOOLEAN NOT NULL,
billing_type TEXT,
area_id TEXT,
```

```
  area_name TEXT,
  salary_from INTEGER,
  salary_to INTEGER,
  salary_currency TEXT,
  salary_gross BOOLEAN,
  type TEXT,
  city TEXT,
  street TEXT,
  lat DOUBLE,
  lng DOUBLE,
  experience TEXT,
  schedule TEXT,
  employment TEXT,
  employer_id TEXT,
  employer_name TEXT,
  employer_url TEXT,
  published_at TIMESTAMP,
  description TEXT,
  key_skills TEXT[],
  group_tag TEXT NOT NULL
```

Всего было собрано порядка 17 000 строк данных, которые требуется подготовить для дальнейшего обучения модели. Для этого был написана программа на Python, которая создает новый столбец “input_text” очищенный от html тегов и дополненный текстовыми метками для лучшего обучения моделей, применяет one-hot encoding к категориальному признаку “key_skills” и преобразует строковое поле “experience” с сохранением важности.

```python
import duckdb
import pandas as pd
import numpy as np
from bs4 import BeautifulSoup

def html_to_text(html):
    if pd.isna(html):
        return ""
    soup = BeautifulSoup(html, "html.parser")
    return soup.get_text(separator="\n", strip=True)

def to_lower(text):
  return text.lower()
```

```
def clean_html_column(df, column="description",
new_column="description_clean"):
    """
    Преобразует HTML-текст в колонке DataFrame в обычный текст и сохраняет
в новую колонку.
    """
    df[new_column] = df[column].apply(html_to_text).apply(to_lower)
    return df

def build_input(row):
    temp = f"[TITLE] {row['name']}"
    skills = ', '.join(row['key_skills']) if isinstance(row['key_skills'],
list) else ''
    if skills:
      remp += f" [SKILLS] {skills}"
    temp += f" [DESC] {row['description_clean']}"
    return temp
```

```
df = df[~(df["salary_from"].isnull() & df["salary_to"].isnull())].copy()
df = df[~df["description"].isna() & (df["description"].str.strip() != "")]
df = clean_html_column(df)
```

Функция для создания целевого столбца:

```
def calculate_target_salary(row):
    salary_from = row['salary_from']
    salary_to = row['salary_to']

    # Случай 1: Есть и salary_from, и salary_to
    if pd.notna(salary_from) and pd.notna(salary_to):
        return (salary_from + salary_to) / 2
    # Случай 2: Есть только salary_from
    elif pd.notna(salary_from):
        return salary_from + 30000
    # Случай 3: Есть только salary_to
    elif pd.notna(salary_to):
        return salary_to - 30000
    # Случай 4: Оба значения отсутствуют
    else:
        return np.nan
```

```
df["input_text"] = df.apply(build_input, axis=1)
df["target_salary"] = df.apply(calculate_target_salary, axis=1)
```

Выбор модели для векторизации текста

Перед обучением регрессионной модели необходимо преобразовать текстовые данные в числовой формат, так как большинство алгоритмов машинного обучения не работают напрямую с текстовой информацией.

Для векторизации используются следующие методы:

Метод	Год	Сложность	Захватывает контекст	Пример использования
Мешок слов (BoW)	1950-е	Низкая	Нет	Классификация спама
TF-IDF	1972	Низкая	Частично	Поиск релевантных документов
Word2Vec	2013	Средняя	Да, локальный	Семантический поиск
GloVe	2014	Средняя	Да, глобальный	Анализ тональности
FastText	2016	Средняя	Да, с n-граммами	Классификация текстов на редких языках
Doc2Vec	2014	Высокая	Да, для документов	Кластеризация новостей
BERT	2018	Высокая	Да, двунаправленный	Вопросно-ответные системы

В своей работе я буду использовать самый продвинутый варинт - модель BERT. BERT использует трансформер — механизм “внимания”, который изучает контекстуальные отношения между словами (или подсловами) в тексте. В своей оригинальной форме трансформер включает в себя два отдельных механизма — кодировщик, который считывает введенный текст, и декодер, который выдает прогноз для задачи. Поскольку целью BERT является создание языковой модели, то ей необходим только кодировщик.

Готовая функция для преобразования в эмбеддинг:

```
      import pandas as pd
      import numpy as np
      import re

      # Функция для получения эмбеддингов RuBERT
      def get_rubert_embeddings(texts, tokenizer, model, max_length=512,
batch_size=8, device='cpu'):
          embeddings = []
          model = model.to(device)
          for i in range(0, len(texts), batch_size):
```

```
            batch_texts = texts[i:i + batch_size]
            inputs = tokenizer(batch_texts, return_tensors="pt",
max_length=max_length,
                               truncation=True, padding=True).to(device)
            with torch.no_grad():
                outputs = model(**inputs)
            batch_embeddings = outputs.last_hidden_state[:, 0,
:].cpu().numpy()
            embeddings.append(batch_embeddings)
        return np.vstack(embeddings)
```

Обучение модели

Для обучения воспользуюсь такими моделями как RandomForest и CatBoost.

Случайный лес (Random Forest) — это ансамблевый алгоритм машинного обучения, который объединяет множество решающих деревьев для повышения точности и устойчивости предсказаний. Каждое дерево обучается на случайной подвыборке данных и случайном наборе признаков, что снижает корреляцию между деревьями и уменьшает риск переобучения. В задачах классификации итоговое решение принимается большинством голосов деревьев, а в задачах регрессии — усреднением их предсказаний. Благодаря своей универсальности, устойчивости к шуму и способности обрабатывать данные с пропущенными значениями, случайный лес широко применяется в различных областях, включая медицину, финансы и маркетинг.

CatBoost — это алгоритм градиентного бустинга, разработанный компанией Яндекс, который эффективно обрабатывает как числовые, так и категориальные признаки без необходимости их предварительного кодирования. Он использует уникальные методы, такие как упорядоченное кодирование и симметричные деревья решений, что повышает точность предсказаний и снижает риск переобучения. CatBoost также автоматически обрабатывает пропущенные значения и поддерживает ускоренное обучение с использованием GPU, что делает его особенно полезным для задач классификации и регрессии на структурированных данных.

Для обученных моделей RandomForest и CatBooest ошибка MSE составила соответство

3182697626.489668

3162352253.801595

Как видно различия несущественны.

Пример работы RandomForest на случайном элементе датасета:

```
Вакансия: Разработчик C# / .NET Core
Предсказанная зарплата: 311558.11
А на самом деле зарплата от 220000 и до 350000
```

Заработная плата попала в изначальный диапазон, что говорит о корректной работе модели.

Заключение

В ходе прохождения практики я получила практические навыки работы с текстовыми данными и познакомилась с различными способами их векторизации. Особое внимание было уделено использованию модели BERT, как одного из современных и эффективных инструментов для преобразования текста в числовое представление. Это позволило получить более качественные признаки для последующего анализа.

На основе подготовленных данных были обучены модели машинного обучения — Random Forest и CatBoost. В процессе я освоила методы обучения и оценки моделей, научилась выбирать подходящие параметры, а также проводить интерпретацию результатов.

Практика помогла мне лучше понять этапы построения модели — от предварительной обработки данных до анализа полученных предсказаний. Полученные знания и навыки являются важной частью моей профессиональной подготовки и могут быть применены при решении реальных задач анализа данных в будущем.

Список литературы

1. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter / Wes McKinney / 2022
2. Как обучить модель случайного леса (Random Forest) в scikit-learn – Текст : электронный –URL: https://labex.io/ru/tutorials/python-how-to-train-random-forest-in-scikit-learn-425422?utm_source=chatgpt.com "Как обучить модель случайного леса (Random Forest) в scikit-learn | LabEx"
3. Как работает алгоритм CatBoost – Текст : электронный – URL: https://pro.arcgis.com/ru/pro-app/latest/tool-reference/geoai/how-catboost-works.htm?utm_source=chatgpt.com "Как работает алгоритм CatBoost—ArcGIS Pro | Документация"
4. UltraLytics CatBoost – Текст : электронный – URL: https://www.ultralytics.com/ru/glossary/catboost?utm_source=chatgpt.com "CatBoost"
5. Transformers for Natural Language Processing. Build innovative deep / Denis Rothman / 2024
6. Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners / Scott Hartshorn / 2016
7. BERT в двух словах: Инновационная языковая модель для NLP. – Текст : электронный – URL: https://habr.com/ru/companies/otus/articles/702838/ (дата обращения 29.05.2025)

Приложение 1

ФГБОУ ВО «Северо-Осетинский государственный университет
имени Коста Левановича Хетагурова»

Кафедра прикладной математики и информатики

УТВЕРЖДАЮ
заведующий кафедрой
_____ / Е.К. Басаева /
«02» мая 2025 г.

РАБОЧИЙ ГРАФИК (ПЛАН) ПРОВЕДЕНИЯ ПРАКТИКИ

Общие сведения

ФИО обучающегося	Дзуцева Лана Сергеевна
Курс, группа, форма обучения	*4 курс, ПМ(б)-24-4-ОФО*
Направление подготовки, профиль	*01.03.02 Прикладная математика и информатика, профиль «Программирование, анализ данных и математическое моделирование»*
Наименование структурного подразделения (кафедра)	*кафедра прикладной математики и информатики*
Вид практики	*производственная практика*
Тип практики	*технологическая (проектно-технологическая) практика*
Способ проведения практики	*Стационарно*
Форма проведения практики	*Непрерывно*
Место прохождения практики	*ООО "М-Софтер"*
Период прохождения практики	*с «02» мая 2025 г. по «30» мая 2025 г.*
Реквизиты договора о прохождении практики (при проведении практики в профильной организации)	50-25 от 12.03.25

Планируемые работы

№ п/п	Содержание работы	Срок выполнения	Отметка о выполнении
1.	Оформление документов по прохождению практики	до начала практики	не требуется
2.	Проведение медицинских осмотров (обследований) в случае выполнения обучающимся работ, при выполнении которых проводятся обязательные предварительные и периодические медицинские осмотры (обследования) в соответствии с законодательством РФ	до начала практики	не требуется

3.	Вводный инструктаж по правилам охраны труда, технике безопасности, пожарной безопасности, оформление временных пропусков для прохода в профильную организацию (при необходимости).	в первый день практики	не требуется
4.	Установочный инструктаж по целям, задачам, срокам и требуемой отчетности.	*02.05.2025 14^{30}–16^{00}*	
5.	Оформление индивидуального задания и плана работы.	*02.05.2025 14^{30}–16^{00}*	
6.	Выполнение индивидуального задания практики	*03.05.2025–30.05.2025*	
7.	Консультации руководителя(-ей) практики о ходе выполнения заданий, оформлении и содержании отчета, по производственным вопросам	*05.05.2025 14^{30}–16^{00}* *12.05.2025 14^{30}–16^{00}* *14.05.2025 14^{30}–16^{00}* *16.05.2025 14^{30}–16^{00}* *19.05.2025 14^{30}–16^{00}* *21.05.2025 14^{30}–16^{00}* *23.05.2025 14^{30}–16^{00}* *26.05.2025 14^{30}–16^{00}*	
8.	Подготовка отчета по практике	*27.05.2025–29.05.2025*	
9.	Проверка отчета по практике, оформление характеристики руководителя(-ей) практики (при наличии)	*29.05.2025 14^{30}–16^{00}*	
10.	Промежуточная аттестация по практике	*30.05.2025*	

Рабочий график (план) составил:
руководитель практики от образовательной организации

ст. преп. каф. ПМиИ	_____	Тотрова М.Х.	«02» мая 2025 г.
(уч. степень, уч. звание, должность)	(подпись)	(И.О. Фамилия)	(дата)

Согласовано (при проведении практики в профильной организации):
руководитель практики от профильной организации

ген. директор	_____	Д.Г. Минасян	«02» мая 2025 г.
(уч. степень, уч. звание, должность)	(подпись)	(И.О. Фамилия)	(дата)

С рабочим графиком (планом) ознакомлен:

обучающийся	_____	Л.С. Дзуцева	«02» мая 2025 г.
	(подпись)	(И.О. Фамилия)	(дата)

Приложение 2

ФГБОУ ВО «Северо-Осетинский государственный университет
имени Коста Левановича Хетагурова»

Кафедра прикладной математики и информатики

УТВЕРЖДАЮ
заведующий кафедрой

_____ / Е.К. Басаева /

«02» мая 2025 г.

ИНДИВИДУАЛЬНОЕ ЗАДАНИЕ НА ПРАКТИКУ

Общие сведения

ФИО обучающегося	*Дзуцева Лана Сергеевна*
Курс, группа, форма обучения	*4 курс, ПМ(б)-24-4-ОФО*
Направление подготовки, профиль	*01.03.02 Прикладная математика и информатика, профиль «Программирование, анализ данных и математическое моделирование»*
Наименование структурного подразделения (кафедра / отделение)	*кафедра прикладной математики и информатики*
Вид практики	*производственная практика*
Тип практики	*технологическая (проектно-технологическая) практика*
Способ проведения практики	*стационарно*
Форма проведения практики	*непрерывно*
Место прохождения практики	*ООО "М-Софтер"*
Период прохождения практики	*с «02» мая 2025 г. по «30» мая 2025 г.*
Реквизиты договора о прохождении практики (при проведении практики в профильной организации)	50-25 от 12.03.25

Содержание индивидуального задания

1. Сбор данных и подготовка датасета
2. Выбор оптимального алгоритма и фреймворка для работы.
3. Обучение модели машинного обучения

Задание на практику составил:
руководитель практики от образовательной организации

ста. преп. каф. ПМиИ	_____	Тотрова М.Х.	«02» мая 2025 г.
(уч. степень, уч. звание, должность)	(подпись)	(И.О. Фамилия)	(дата)

Согласовано (при проведении практики в профильной организации):
руководитель практики от профильной организации

ген. директор	_____	Д.Г. Минасян	«02» мая 2024 г.
(уч. степень, уч. звание, должность)	(подпись)	(И.О. Фамилия)	(дата)

Задание на практику принял:

обучающийся	_____	Л.С.Дзуцева	«02» мая 2025 г.
	(подпись)	(И.О. Фамилия)	(дата)

Приложение 3

ФГБОУ ВО «Северо-Осетинский государственный университет
имени Коста Левановича Хетагурова»

Кафедра прикладной математики и информатики

ДНЕВНИК ПРАКТИКИ

Общие сведения

ФИО обучающегося	Дзуцева Лана Сергеевна
Курс, группа, форма обучения	*4 курс, ПМ(б)-24-4-ОФО*
Направление подготовки, профиль	*01.03.02 Прикладная математика и информатика, профиль «Программирование, анализ данных и математическое моделирование»*
Наименование структурного подразделения (кафедра / отделение)	*кафедра прикладной математики и информатики*
Вид практики	*производственная практика*
Тип практики	*технологическая (проектно-технологическая) практика*
Способ проведения практики	*стационарно*
Форма проведения практики	*непрерывно*
Место прохождения практики	*ООО "М-Софтер"*
Период прохождения практики	*с «02» мая 2025 г. по «30» мая 2025 г.*
Реквизиты договора о прохождении практики (при проведении практики в профильной организации)	50-25 от 12.03.25

Учет выполняемой работы

№ п/п	Содержание работы	Дата выполнения	Отметка о выполнении
1.	Выдача и оформление индивидуального плана прохождения практики и индивидуального задания на практику	02.05.2025	*выполнено*
2.	Сбор данных и подготовка датасета	03.05.2024–10.05.2025	*выполнено*
3.	Выбор оптимального алгоритма и фреймворка для работы.	12.05.2025–17.05.2025	*выполнено*
4.	Обучение модели машинного обучения	19.05.2024–26.05.2025	*выполнено*
5.	Подготовка отчета по практике. Защита отчета по практике.	27.05.2024–29.05.2025 30.05.2025	*выполнено*

Дневник заполнил:
обучающийся _____ Л.С. Дзуцева 30 мая 2025 г.
(подпись) (И.О. Фамилия) (дата)

Дневник проверил (при проведении практики в профильной организации):
руководитель практики от профильной организации

ген. директор _____ Д.Г. Минасян «30» мая 2025 г.
(уч. степень, уч. звание, должность) (подпись) (И.О. Фамилия) (дата)

Дневник проверил:
руководитель практики от образовательной организации

к.ф.-м.н., зав. кафедрой _____ Е.К. Басаева 30 мая 2025 г.
(уч. степень, уч. звание, должность) (подпись) (И.О. Фамилия) (дата)

Характеристика (отзыв) руководителя практики от профильной организации
(при проведении практики в профильной организации)

Оценка трудовой деятельности и дисциплины: За время прохождения практики Дзуцева Л. С. проявила себя ответственным и дисциплинированным работником, справилась со всеми возложенными на него обязанностями и полностью выполнил программу практики, проявив самостоятельность и исследовательские способности.

Оценка содержания и оформления отчета по практике: Отчет содержит подробное и структурированное описание всех этапов работы, выполненной в ходе практики. Текст отчета написан грамотным и профессиональным языком, что облегчает его понимание.

Оценка по практике: _____.

Руководитель практики от профильной организации

ген. директор	_____	Д.Г. Минасян	«30» мая 2025 г.
(уч. степень, уч. звание, должность)	(подпись)	(И.О. Фамилия)	(дата)