



С о п р о в о д и т е л ь н а я д о к у м е н т а ц и я

Веб-сервис: комплексное аналитическое решение

для контроля эффективности сотрудников,

включая визуализацию, машинное обучение

и взаимодействие с базой данных

Исполнитель: «Timebook»

Оглавление

1. Оглавление	2
2. Структура данных таблиц СУБД	3
3. Описание структуры входных и выходных данных	5
4. Описание модулей и основных функций	7
5. Инструменты проекта	9



1. Структура данных таблиц СУБД

База данных **lct** – схема **public**

Таблица **dataset** основная структура, поступающая на вход пайплайна

dataset
abc id
abc фιο
abc date
123 period
123 месячный_доход
123 получаемых_сообщений_за_период
123 отправленных_сообщений_за_период
123 адресатов_в_отправляемых_сообщени
123 сообщений_с_адресами_в_скрытой_коп
123 сообщений_с_адресатами_в_поле_копи
123 сообщений_прочитанных_после_4_часо
123 дней_между_получено_прочитано
123 ответов_на_сообщения
123 символов_в_исходящих_сообщениях
123 сообщений_отправленных_вне_рамоч_
123 соотношение_полученных_отправлен
123 соотношение_объема_полученных_отп
123 входящих_сообщений_с_?_без_ответа
123 target

Id: уникальный идентификатор записи;

фио: полное имя сотрудника;

date: дата получения данных

period: номер пятидневной рабочей недели с момента ведения наблюдений;

месячный доход: доход сотрудника в указанную дату;

полученных сообщений за период: количество сообщений, полученных за указанную дату;

отправленных сообщений за период: количество отправленных сообщений;

адресатов в отправляемых сообщениях: количество уникальных адресатов в отправленных сообщениях за указанную дату;

сообщений с адресатами в скрытой копии: количество сообщений, в которых адресаты указаны в скрытой копии;

сообщений с адресатами в поле копии: количество сообщений с адресатами в поле копии;

сообщений прочитанных после 4 часов: количество сообщений, прочитанных более чем через 4 часа после получения;

дней между получено и прочитано: количество дней между получением и прочтением сообщения;

ответов на сообщения: количество ответов на отправленные сообщения;

символов в исходящих сообщениях: общее количество символов в отправленных сообщениях;

сообщений отправленных вне рамок: количество сообщений, отправленных вне установленных рамок рабочего времени;

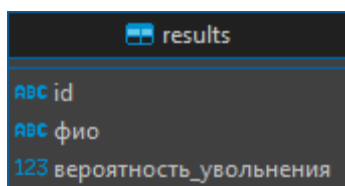
соотношение полученных/отправленных: соотношение между количеством полученных и отправленных сообщений;

соотношение объема полученных/отправленных: соотношение между объемом полученных и отправленных сообщений;

входящих сообщений с ? без ответа: количество входящих сообщений с вопросительным знаком, на которые не было ответа;

target: целевая переменная, 1 – сотрудник уволился, 0 – сотрудник продолжает работать.

Таблица **results** формируется, как результат работы модели машинного обучения. Содержит данные о сотруднике и результат прогнозирования вероятности увольнения за последний период.



results		
abc	id	
abc	фio	
123	вероятность_увольнения	

Id: уникальный идентификатор записи;

фio: полное имя сотрудника;

вероятность_увольнения: спрогнозированная за выбранный период вероятность увольнения.



2. Описание структуры входных и выходных данных

Принципы формирования датасета:

- Датасет сформирован на период с 01.12.2022 по 22.11.2023;
- Гранулярность исходного датасета – 1 день, значения всех признаков указываются суммарно за один рабочий день;
- Период составляет 5 рабочих дней, в модели указывается период, с которого начинается контрольная выборка (заканчивается последней датой данных в БД);
- Количество структурных подразделений – 5;
- Начальное количество сотрудников в датасете – 1000;
- Количество уволившихся сотрудников за указанный период – 250;
- Эффективность работы уволившегося сотрудника и уровень его удовлетворенности условиями труда начинали снижаться в среднем за 50 дней до факта увольнения, это снижение отражается в динамике признаков;
- Признаки эффективности и удовлетворенности рассчитывались случайным образом с заданными параметрами меры среднего, дисперсии, уровнем корреляции между собой.

Независимые признаки (входные данные), используемые в модели:

- **Id** - уникальный идентификатор записи;
- **ФИО** - полное имя сотрудника;
- **Date** - дата получения данных;
- **Period** - номер пятидневной рабочей недели с момента ведения наблюдений;
- **Количество входящих сообщений за период** – не зависит от уровня удовлетворенности;
- **Количество исходящих сообщений за период** – имеет положительную корреляцию с эффективностью (чем ниже эффективность, тем меньше отправленных сообщений);
- **Количество адресатов в исходящих сообщениях** – зависит от количества отправленных сообщений и имеет положительную корреляцию с эффективностью;

- **Количество исходящих сообщений с адресами в скрытой копии** - зависит от количества отправленных сообщений и имеет отрицательную корреляцию с удовлетворённостью;
- **Количество исходящих сообщений с адресатами в поле копия** - зависит от количества отправленных сообщений;
- **Количество входящих сообщений, прочитанных после 4 часов** - зависит от количества отправленных сообщений и имеет положительную корреляцию с эффективностью;
- **Количество дней между получением и прочтением входящих писем** - зависит от количества отправленных сообщений и имеет положительную корреляцию с эффективностью;
- **Количество ответов на входящие сообщения** - зависит от количества полученных сообщений и имеет положительную корреляцию с эффективностью;
- **Количество символов в исходящих сообщениях** - зависит от количества исходящих сообщений, имеет слабую положительную корреляцию с эффективностью;
- **Количество сообщений, отправленных вне рамок рабочего дня** – зависит от количества исходящих сообщений и имеет положительную корреляцию с удовлетворенностью;
- **Соотношение количества исходящих сообщений к входящим** – рассчитывается как отношение исходящих к отправленным;
- **Соотношение объема исходящих сообщений к входящим в байтах** – рассчитывается как отношение исходящих к отправленным, имеет слабую положительную корреляцию с удовлетворенностью;
- **Количество входящих сообщений с вопросом без ответа** – имеет положительную корреляцию с удовлетворенностью;
- **Месячный доход** – изменяется от месяца к месяцу у каждого сотрудника, интерпретируется как доход (зарплата, надбавки, премии и др.), имеет положительную корреляцию с удовлетворенностью;

Целевая переменная для обучающей выборки:

- **Target** - 1 – сотрудник уволился, 0 – сотрудник продолжает работать.

Целевая переменная(выходные данные):

- **Target** – вероятность увольнения сотрудника за выбранный период.



3. Описание модулей и основных функций

Модуль 1: Подключение к Базе Данных

Задание параметров подключения к базе данных PostgreSQL.

Использование библиотеки `psycopg2` для установления соединения.

Выполнение SQL-запроса для извлечения данных о сотрудниках.

Модуль 2: Обработка и Подготовка Данных

Преобразование дат в удобный формат и создание календарных признаков.

Преобразование данных и подготовка `DataFrame` с использованием библиотеки `pandas`.

Модуль 3: Визуализация и Анализ Данных

Вывод и исследование структуры данных БД.

Создание дашборда с вероятностью увольнения для каждого сотрудника за выбранный период.

Использование библиотек `Plotly` и `Matplotlib` для создания графиков.

Модуль 4: Машинное Обучение и Прогнозирование

Формирование обучающей и контрольной выборок для обучения модели машинного обучения.

Использование библиотеки `CatBoostClassifier`, работающей с задачами бинарной классификации, для обучения модели вероятности увольнения сотрудника.

Предусмотрена возможность работы с категориальными признаками, в том числе добавление новых фичей.

Установка порога для классификации сотрудников по вероятности увольнения.

Модуль 5: Вывод Результатов и Экспорт Данных

Визуализация результатов в виде интерактивных графиков и таблиц с цветовой кодировкой.

Возможность скачивания данных в формате `CSV` и `Excel` для последующего анализа.

Предоставление доступа к дашборду в `Grafana` для более детального изучения данных.

Модуль 6: Отправка Результатов и Рассылка

Организация отправки электронных писем с результатами анализа начальникам отделов.

Временная заглушка о успешной рассылке, пока не будут предоставлены реальные адреса электронной почты.

Модуль 7: Завершение Работы и Закрытие Соединения

Закрытие соединения с базой данных PostgreSQL после завершения всех операций.

Завершение работы веб-приложения с использованием Streamlit.

Данное приложение является инструментом для анализа и прогнозирования, позволяющим принимать обоснованные управленческие решения.



4. Инструменты проекта

Язык программирования:

Python

Библиотеки для обработки данных:

Pandas

NumPy

Визуализация данных:

Matplotlib

Plotly

Веб-приложение:

Streamlit

Взаимодействие с базой данных:

Psycopg2 (для работы с PostgreSQL)

SQLAlchemy

Машинное обучение:

CatBoostClassifier

Работа с электронной почтой:

MIMEText, MIMEMultipart

smtplib

Генерация случайных данных:

Random (для управления случайным зерном)

Веб-документация:

Markdown

Дополнительные инструменты для визуализации:

IPython.display

HTML, Javascript

Интерактивные графики и дашборды:

Grafana (для визуализации внешних дашбордов)

Системные операции с файлами и директориями:

io

os

Управление временными задержками:

time