HR_Сборный проект — 2

Описание проекта

Этапы работы – декомпозировать задачи – разделить их на более мелкие.

Задача. HR-аналитики компании «Работа с заботой» помогают бизнесу оптимизировать управление персоналом: бизнес предоставляет данные, а аналитики предлагают, как избежать финансовых потерь и оттока сотрудников. HR-аналитики используют машинное обучение, с помощью которого можно быстрее и точнее отвечать на вопросы бизнеса.

Компания «Работа с заботой» предоставила данные с характеристиками сотрудников компании. Среди них — уровень удовлетворённости сотрудника работой в компании. Информацию получили из форм обратной связи: сотрудники заполняют тест-опросник, и по его результатам рассчитывается доля их удовлетворённости от 0 до 1, где 0 — совершенно неудовлетворен, 1 — полностью удовлетворён.

Первая задача — построить модель, которая сможет предсказать уровень удовлетворённости сотрудника на основе данных заказчика.

Бизнесу это важно: удовлетворённость работой напрямую влияет на отток сотрудников. А предсказание оттока — одна из важнейших задач HR-аналитиков. Внезапные увольнения несут в себе риски для компании, особенно если уходит важный сотрудник.

Вторая задача — построить модель, которая сможет на основе данных заказчика предсказать то, что сотрудник уволится из компании.

Подробнее о задачах:

Задача 1: предсказание уровня удовлетворённости сотрудника

Заказчик предоставил данные с признаками:

- `id` — уникальный идентификатор сотрудника;
- `dept` — отдел, в котором работает сотрудник;
- `level` — уровень занимаемой должности;
- `workload` — уровень загруженности сотрудника;
- `employment_years` — длительность работы в компании (в годах);
- `last_year_promo` — показывает, было ли повышение за последний год;
- `last_year_violations` — показывает, нарушал ли сотрудник трудовой договор за последний год;
- `supervisor_evaluation` — оценка качества работы сотрудника, которую дал руководитель;

- `salary` — ежемесячная зарплата сотрудника;
- `job_satisfaction_rate` — уровень удовлетворённости сотрудника работой в компании, целевой признак.

Шаг 1. Загрузка данных

Загружены файлы с данными:

Тренировочная выборка:

train_job_satisfaction_rate.csv

Входные признаки тестовой выборки:

test_features.csv

Целевой признак тестовой выборки:

test_target_job_satisfaction_rate.csv

Шаг 2. Предобработка данных

Изучены данные и сделаны выводы, выполнена предобработка, пропуски, заполнены в пайплайне.

Шаг 3. Исследовательский анализ данных

Исследованы все признаки и сделаны выводы о том, как их нужно подготовить.

Шаг 4. Подготовка данных

Подготовка признаков выполнена в пайплайне, пайплайн дополнен шагом предобработки. При кодировании учтены особенности признаков и моделей и использованы два кодировщика.

Шаг 5. Обучение моделей

Обучены две модели. Взята одна линейная модель, а в качестве второй — дерево решений. Подобраны гиперпараметры для одной модели с помощью одного из известных инструментов. param_grid Настройка GridSearchCV.

Выбрана лучшая модель и проверено её качество. Выбор сделан на основе новой метрики — SMAPE (англ. symmetric mean absolute percentage error, «симметричное среднее абсолютное процентное отклонение»).

Метрика SMAPE вычислена по формуле:

SMAPE=100n∑i=1n|yi−yi^|(|yi|+|y^i|)/2,SMAPE=$_{n}$100∑$_{i=1}$$_{n}$(|$y_i$|+|$y$^$_i$|)/2|$y_i$−$y_i$^|,

где:

- yiyi — фактическое значение целевого признака для объекта с порядковым номером ii в выборке;
- yi^yi^ — предсказанное значение целевого признака для объекта с порядковым номером ii в выборке;
- nn — количество объектов в выборке;
- ∑i=1n∑i=1n — сумма значений, полученная в результате операций, которые следуют за этим знаком, для всех объектов с порядковым номером от ii до nn в выборке.

Напиана функция, которая принимает на вход массивы NumPy или объекты `Series` в pandas и возвращает значение метрики SMAPE. Использована эта метрика при подборе гиперпараметров и оценке качества моделей.

Критерий успеха: SMAPE ≤15≤15 на тестовой выборке.

В решении сохранена работа со всеми моделями, которые опробованы. Сделаны выводы.

Шаг 6. Оформление выводов

Сделаны промежуточные выводы о том, какая модель справилась лучше и почему.

Задача 2: предсказание увольнения сотрудника из компании

Для этой задачи использованы те же входные признаки, что и в предыдущей задаче. Однако целевой признак: `quit` — увольнение сотрудника из компании.

Шаг 1. Загружены данные из файлов:

Тренировочная выборка:

train_quit.csv

Входные признаки тестовой выборки те же, что и в первой задаче:

test_features.csv

Целевой признак тестовой выборки:

test_target_quit.csv

Шаг 2. Предобработка данных

Изучены данные и выполнена предобработка, пропуски заполнены в пайплайне.

Шаг 3. Исследовательский анализ данных

3.1. Проведен исследовательский анализ данных.

3.2. Составлен портрет «уволившегося сотрудника». Определено, в каком отделе с большей вероятностью работает уволившийся сотрудник и какой у него уровень загруженности. Проведено сравнение средней зарплаты ушедших сотрудников с теми, кто остался в компании.

3.3. Уровень удовлетворённости сотрудника работой в компании влияет на то, уволится ли сотрудник: визуализированы и сравнены распределения признака `job_satisfaction_rate` для ушедших и оставшихся сотрудников. Использованы данные с обоими целевыми признаками тестовой выборки.

Шаг 4. Добавлен новый входного признак

Проверено, что `job_satisfaction_rate` и `quit` действительно связаны и получено необходимое значение метрики в первой задаче. Добавлен `job_satisfaction_rate`, предсказанный лучшей моделью первой задачи, к входным признакам второй задачи.

Шаг 5. Подготовка данных

Подготовлены признаки, как в первой задаче: выполнена подготовка в пайплайне, дополнен пайплайн предобработки. При кодировании учтены особенности признаков и моделей и использованы два кодировщика.

Шаг 6. Обучение модели

Обучены три модели. Для двух из них подобраны гиперпараметры. Проверено качество лучшей модели.

Метрика оценки качества — ROC-AUC. Критерий успеха: ROC-AUC $\geq 0.91 \geq 0.91$ на тестовой выборке. Проведен отбор признаков, что помогло улучшить метрику.

Шаг 7. Выводы

Сделаны промежуточные выводы о том, какая модель справилась лучше и почему.

Общий вывод

Сформулирован общий вывод:

- описана задача;
- описаны все этапы работы;
- добавлены выводы и дополнительные предложения для бизнеса.

Оформление

Выполнено задание в Jupyter Notebook. Заполнен программный код в ячейках типа `code`, текстовые пояснения — в ячейках типа `markdown`. Использовано форматирование и заголовки.