**Описание проекта**

Позади много уроков и заданий, и закрепить свои знания, как обычно, вы сможете в самостоятельном проекте. Это непростой проект, в котором от вас потребуется больше самостоятельности. Этапы работы описаны кратко, и вам понадобится декомпозировать задачи, то есть разделить их на более мелкие. Опирайтесь на знания об этапах анализа данных и машинного обучения из прошлых курсов.

Перейдём к задаче. HR-аналитики компании «Работа с заботой» помогают бизнесу оптимизировать управление персоналом: бизнес предоставляет данные, а аналитики предлагают, как избежать финансовых потерь и оттока сотрудников.

В этом HR-аналитикам пригодится машинное обучение, с помощью которого получится быстрее и точнее отвечать на вопросы бизнеса.

Компания предоставила данные с характеристиками сотрудников компании. Среди них — уровень удовлетворённости сотрудника работой в компании. Эту 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 неудовлетворён, 1 — полностью удовлетворён.

Собирать данные такими опросниками не так легко: компания большая, и всех сотрудников надо сначала оповестить об опросе, а затем проследить, что все его прошли.

У вас будет несколько задач.

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

Почему 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.

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

Теперь расскажем подробнее о задачах.

**Задача 1: предсказание уровня удовлетворённости сотрудника**

Для этой задачи 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**

Загрузите файлы с данными:

Тренировочная выборка:

[train\_job\_satisfaction\_rate.csv](https://code.s3.yandex.net/datasets/train_job_satisfaction_rate.csv)

Входные признаки тестовой выборки:

[test\_features.csv](https://code.s3.yandex.net/datasets/test_features.csv)

Целевой признак тестовой выборки:

[test\_target\_job\_satisfaction\_rate.csv](https://code.s3.yandex.net/datasets/test_target_job_satisfaction_rate.csv)

**Шаг 2. Предобработка данных**

Изучите данные и сделайте выводы. При необходимости выполните предобработку. Если есть пропуски, заполните их в пайплайне.

**Шаг 3. Исследовательский анализ данных**

Исследуйте все признаки и сделайте выводы о том, как их нужно подготовить.

**Шаг 4. Подготовка данных**

Подготовку признаков выполните в пайплайне, дополнив пайплайн шага предобработки. При кодировании учитывайте особенности признаков и моделей и используйте как минимум два кодировщика.

**Шаг 5. Обучение моделей**

Обучите как минимум две модели. Возьмите одну линейную модель, а в качестве второй — дерево решений. Подберите гиперпараметры как минимум для одной модели с помощью одного из известных вам инструментов.

Выберите лучшую модель и проверьте её качество. Выбор делайте на основе новой метрики — SMAPE (англ. symmetric mean absolute percentage error, «симметричное среднее абсолютное процентное отклонение»).

Метрика SMAPE вычисляется так:

SMAPE=100𝑛∑𝑖=1𝑛∣𝑦𝑖−𝑦𝑖^∣(∣𝑦𝑖∣+∣𝑦^𝑖∣)/2,SMAPE=*n*100​∑*i*=1*n*​(∣*yi*​∣+∣*y*^​*i*​∣)/2∣*yi*​−*yi*​^​∣​,

где:

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

Напишите функцию, которая принимает на вход массивы NumPy или объекты Series в pandas и возвращает значение метрики SMAPE. Используйте эту метрику при подборе гиперпараметров и оценке качества моделей.

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

В решении сохраните работу со всеми моделями, которые вы пробовали. Сделайте выводы.

**Шаг 6. Оформление выводов**

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

**Задача 2: предсказание увольнения сотрудника из компании**

Для этой задачи вы можете использовать те же входные признаки, что и в предыдущей задаче. Однако целевой признак отличается: это quit — увольнение сотрудника из компании.

**Шаг 1. Загрузка данных**

Загрузите данные из файлов:

Тренировочная выборка:

[train\_quit.csv](https://code.s3.yandex.net/datasets/train_quit.csv)

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

[test\_features.csv](https://code.s3.yandex.net/datasets/test_features.csv)

Целевой признак тестовой выборки:

[test\_target\_quit.csv](https://code.s3.yandex.net/datasets/test_target_quit.csv)

**Шаг 2. Предобработка данных**

Изучите данные и при необходимости выполните предобработку. Если есть пропуски, заполните их в пайплайне.

**Шаг 3. Исследовательский анализ данных**

3.1. Проведите исследовательский анализ данных.

3.2. Составьте портрет «уволившегося сотрудника». Например, можно узнать, в каком отделе с большей вероятностью работает уволившийся сотрудник и какой у него уровень загруженности. Также можно сравнить среднее значение зарплаты ушедших сотрудников с теми, кто остался в компании.

3.3. Аналитики утверждают, что уровень удовлетворённости сотрудника работой в компании влияет на то, уволится ли сотрудник. Проверьте это утверждение: визуализируйте и сравните распределения признака job\_satisfaction\_rate для ушедших и оставшихся сотрудников. Используйте данные с обоими целевыми признаками тестовой выборки.

**Шаг 4. Добавление нового входного признака**

Допустим, job\_satisfaction\_rate и quit действительно связаны и вы получили необходимое значение метрики в первой задаче. Тогда добавьте job\_satisfaction\_rate, предсказанный лучшей моделью первой задачи, к входным признакам второй задачи.

**Шаг 5. Подготовка данных**

Подготовьте признаки так же, как и в первой задаче: выполните подготовку в пайплайне, дополнив пайплайн предобработки. При кодировании учитывайте особенности признаков и моделей и используйте как минимум два кодировщика.

**Шаг 6. Обучение модели**

Обучите как минимум три модели. Как минимум для двух из них подберите гиперпараметры. Проверьте качество лучшей модели.

Метрика оценки качества в этой задаче — ROC-AUC. Критерий успеха: ROC-AUC ≥0.91≥0.91 на тестовой выборке. Напомним: отбор признаков часто помогает улучшить метрику.

**Шаг 7. Выводы**

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

**Общий вывод**

Сформулируйте общий вывод:

* опишите задачу;
* опишите все этапы работы;
* добавьте выводы и дополнительные предложения для бизнеса.

**Оформление**

Выполните задание в Jupyter Notebook. Заполните программный код в ячейках типа code*,* текстовые пояснения — в ячейках типа markdown. Используйте форматирование и заголовки.

**Как будут проверять мой проект**

Ваш проект будут оценивать по критериям. Прежде чем решать кейс, внимательно изучите их.

На что обращают внимание при проверке проектов:

* как вы описываете найденные в данных проблемы;
* какие методы построения графиков применяете для исследовательского анализа;
* какие методы предобработки и подготовки данных применяете и как обосновываете принятое решение;
* используете ли пайплайны и автоматизируете ли повторяющиеся шаги;
* какие ML-решения вы реализуете;
* как выбираете и объясняете лучшее ML-решение;
* как вы интерпретируете результаты работы лучшей модели;
* соблюдаете ли структуру проекта и поддерживаете ли аккуратность кода;
* какие выводы делаете для бизнеса;
* оставляете ли комментарии к шагам.

Успехов!

Перейти к заданию