Тестовое задание на позицию Data Scientist

Введение

Данное тестовое задание предназначено для оценки Ваших навыков в методах машинного обучения и анализа данных. Для завершения данного тестового задания Вам потребуются как теоретические, так и практические навыки. Критерия выполнения/не выполнения данного тестового задания нет, нам важно проследить ваш подход к решению и ваши знания.

Выбор инструментов для решения задачи остается на Ваше усмотрение, предпочтительно для нас : Python, Java, R и т.д.

Данные

В качестве входных данных Вам предоставлены тексты 100 тысяч отзывов клиентов о продуктах в китайском онлайн-магазине JD.com (*jd_reviews.csv*). Дополнительно, у Вас имеется дата создания отзыва, product_id и название самого продукта, а также количество звезд, которое имеет данный отзыв (от 1 до 5). Количество звезд -- это общая оценка, которую поставил автор отзыва соответствующему продукту.

Задача

Необходимо построить модель, которая по входным данным предсказывает количество звезд в отзыве. Итоговый результат должен содержать набор скриптов и обученных моделей с подробной инструкцией по их запуску. На вход должен подаваться файл, аналогичный *jd_reviews.csv*, а на выходе файл формата: review_id; количество_предсказанных_звезд. Качество Вашей модели будет тестироваться на отложенной выборке, которая Вам недоступна.

Пожелания к оформлению отчета

Результат выполнения Вашего тестового задания будет интересен не только специалистам в области машинного обучения, поэтому, кроме модели, мы просим Вас также подготовить аналитический отчет. В отчет необходимо включить предварительный анализ исходных данных, шаги/идеи (не обязательно удачные), которые привели к финальному решению, а также описание полученной модели и мотивации ее выбора.