

ПОЧТА БАНК И ЦЕНТР-ИНВЕСТ

Кейс второго отборочного этапа на Молодежную программу FINODAYS

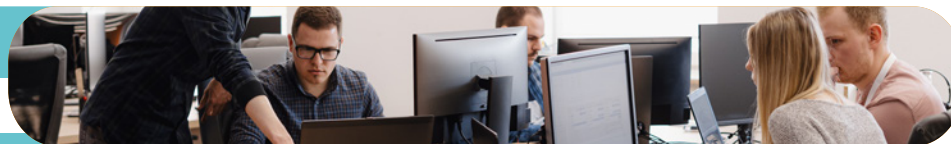
Иван и Михаил дружат с университетской скамьи. Даже профессии выбрали похожие — работают специалистами по машинному обучению: один в Почта Банке, другой — в банке «Центр-инвест». Сейчас перед ними стоят задачи по созданию прототипов сервисов для анализа тональности сообщений и автоматизации оценки имущества.

На втором этапе отбора вам вместе с героями кейса предстоит погрузиться в мир машинного обучения и предложить наиболее подходящее решение для одной из двух задач кейса на ваш выбор. Успехов!

СОДЕРЖАНИЕ

Приветствие

3



Введение

4



Задание 1. Определение эмоционального окраса сообщений в чате

6



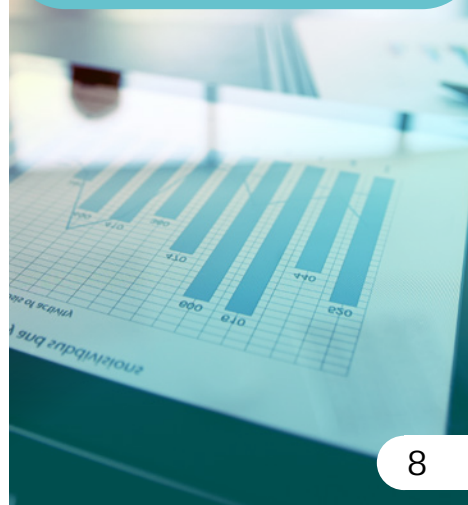
Задание 2. Автоматизированная оценка залогового имущества

7



Приложения

8



Этот кейс подготовлен исключительно в образовательных целях. Авторы не намерены иллюстрировать как эффективное, так и неэффективное решение управленческой проблемы. Кейс не содержит исчерпывающую информацию, необходимую для решения. Для построения модели и решения вы можете использовать любые источники и свои допущения. Некоторые имена в кейсе, а также другая идентификационная информация могли быть изменены с целью соблюдения конфиденциальности.

ПРИВЕТСТВИЕ

Уважаемые участники, добро пожаловать на второй отборочный этап в рамках молодежной программы FINODAYS!

Вам предстоит выбрать одну из двух задач и в течение восьми дней, работая в команде, предложить ее реализацию в виде прототипа сервиса.

Кейс содержит актуальные бизнес-задачи, с которыми сталкиваются IT-департаменты банков.

При использовании конкретной модели машинного обучения, лежащей в основе работы прототипа сервиса, аргументируйте свой выбор.

В качестве результата вам нужно прислать презентацию объемом не более 8 слайдов + приложения, а также все необходимые файлы одним zip-архивом.

Желаем удачи!

Жюри оценит решения пяти критериям:



Подход к разработке, включая описание принципа работы системы (прототипа сервиса) и ее технической архитектуры, а также используемых технологий.



Глубина проработки предложенного решения, анализ альтернатив.



Уровень качества MVP¹.



Соответствие решения поставленным требованиям в задании.



Презентация и продажа решения.

¹ MVP (Minimal Viable Product) — тестовая версия товара, услуги или сервиса с минимальным набором функций (иногда даже одной), которая несет ценность для конечного потребителя.

ВВЕДЕНИЕ

Каждый год в конце августа в Москве проходит конференция по машинному обучению Machine Learning Horizons². В этот раз она посвящена трендам и внедрению методов машинного обучения в сфере финансовых технологий.

Иван, ведущий специалист по машинному обучению IT-департамента Почта Банка, изучал программу конференции в холле, как вдруг услышал, что его окликают.

— Ваня, ты? Сколько лет, сколько зим. Тоже выступаешь на конференции?

— Миша, привет! Какая встреча! Да, сегодня после полудня, — Иван улыбнулся, увидев старого университетского приятеля.

— Мое выступление завтра, но послушаю сегодняшних спикеров: у них актуальные темы, — ответил Михаил. — Мы в департаменте информационных технологий «Центр-инвеста» как раз работаем над несколькими ML-моделями по автоматизации оценки стоимости имущества, пока что обучаем их на открытых данных.

— Звучит интересно! Я слышал, что вы уделяете много внимания направлению ESG³? — спросил Иван.

— Да, именно. Оценка залогового имущества — трудоемкая процедура, предполагающая привлечение сторонних специалистов, временные и финансовые затраты. Поэтому автоматизация минимизирует издержки и значительно ускорит этот процесс. Также наша цель — научить модели включать в оценку стоимости имущества не только, к примеру, год постройки, но и степень экологичности жилья, чтобы предоставлять владельцам экодомов более выгодные условия кредитования. Здорово ведь, правда? — улыбнулся Михаил.

— Полностью согласен! Меня еще заинтересовал спикер, который расскажет о задачах по анализу текста. Это очень актуально для Почта Банка. Один из наших проектов направлен на измерение и постоянный анализ клиентской удовлетворенности. Мы хотим знать, насколько Почта Банк нравится клиентам и как они воспринимают качество наших продуктов и сервисов, — Иван задумчиво повертел в руках брошюру с программой конференции.



— Понимаю, мы в «Центр-инвесте» тоже фокусируемся на улучшении взаимодействия с клиентами. Какие модели машинного обучения вы планируете использовать, если не секрет? — спросил Михаил.

— Думаю, что при построении модели важно начинать с более простых и интерпретируемых моделей и заканчивать более сложными. Простая модель, например логистическая регрессия, задает baseline⁴, на который в дальнейшем нужно ориентироваться.

² Все имена и названия вымышленные, любые совпадения случайны. Данные в кейсе могут быть изменены в целях конфиденциальности.

³ ESG — совокупность характеристик управления компанией, при которой достигается вовлечение данной компании в решение экологических, социальных и управленческих проблем.

⁴ Baseline — базовая модель, используемая как ориентир для оценки качества работы модели.

ВВЕДЕНИЕ

После построения и оценки пробной модели можно переходить к улучшению результатов. Для анализа текста, который является типичной задачей квалификации, я бы использовал Random Forest⁵, SVM⁶ или Naïve Bayes⁷. На этом этапе я бы посмотрел сразу на несколько опций и сравнил их качество. Это позволит выбрать подходящий алгоритм для конкретной задачи, — ответил Иван.

— А если все эти модели покажут низкую точность, что собираешься использовать в качестве тяжелой артиллерии? Word2Vec, GloVe или глубокое обучение на основе нейронных сетей? — поинтересовался Михаил.

— Да, можно попробовать все, что ты перечислил. Также значимо на итоговый результат повлияет то, какие данные использовать для обучения моделей. Ты упоминал, что при работе над оценкой имущества вы обучаете модели на открытых дан-

ных. Какими именно данными вы пользуетесь? — спросил Иван.

— Для обучения моделей наша команда использует открытые данные от [Kaggle](#) или [Dataset Search](#), разработанный Google. Для целей анализа тональности я рекомендую [Large Movie Dataset](#) или [Sentiment140](#) от Стэнфордского университета, — перечислил Михаил без запинки.

— Спасибо! И как ты запоминаешь все названия ресурсов наизусть?! — улыбнулся Иван. — Ладно, через десять минут начинаются пленарные сессии — нам стоит поторопиться. Let's go!

Друзья стали подниматься по лестнице на второй этаж, по пути не переставая обсуждать тренды машинного обучения. Впереди их ждал насыщенный двухдневный интенсив конференции, много дискуссий и новые знакомства.



⁵ Random Forest — алгоритм машинного обучения, заключающийся в использовании комитета решающих деревьев.

⁶ SVM (Support Vector Machines) — линейный алгоритм используемый в задачах классификации и регрессии.

⁷ Naive Bayes [Classifier] — простой вероятностный классификатор, основанный на применении теоремы Байеса.

ЗАДАНИЕ 1. ОПРЕДЕЛЕНИЕ ЭМОЦИОНАЛЬНОГО ОКРАСА СООБЩЕНИЙ В ЧАТЕ

После конференции полный энтузиазма Иван начал утро понедельника с идеи, над которой думал все выходные. Не теряя времени, он написал письмо команде.

Коллеги, привет!

Каждый день мы получаем большое количество сообщений от клиентов, но нет четкого понимания, как оценивать эмоциональный фон подобных месседжей. Для Почта Банка важно знать потребности клиентов, число которых должно вырасти до 21 млн человек к 2023 году согласно обновленной стратегии.

Поэтому наша задача на этой неделе — **создать прототип сервиса, который с помощью машинного обучения сможет определять эмоциональный окрас входящих сообщений и классифицировать их в одну из трех категорий: негативные, позитивные или нейтральные.** Проект масштабный, так как сервис должен охва-

тывать всех пользователей онлайн-банка и мобильного приложения Почта Банка, а предполагаемый срок внедрения — второй квартал 2022 года.

При реализации сервиса необходимо:

- Протестировать его работоспособность на открытых данных, используя датасет с отзывами о фильмах, Large Movie Dataset или Sentiment140 от Стэнфордского университета, а также на реальных кейсах, собранных Почта Банком (см. приложения). Чем выше точность вычислений на тестовых выборках, тем лучше.
- Научить его выдавать рекомендации для сотрудников, как банк может помочь сделать жизнь клиентов удобнее. Например, если комментарии положительные, то можно предложить клиентам, которые их оставили, выставить оценку приложению. В случае негативных комментариев к работе с такими пользователями нужно подключать операторов, специали-



рующихся на работе с конфликтами. Повысить полезность сервиса за счет учета степени интенсивности той или иной эмоции на основе полученных данных, то есть давать количественную оценку отзывам, например отзыв на 80% положительный, а на 20% — нейтральный.

- Оценить потенциальный экономический эффект от внедрения сервиса с учетом увеличения числа клиентов банка.

Хорошего дня!

ЗАДАНИЕ 2. АВТОМАТИЗИРОВАННАЯ ОЦЕНКА ЗАЛОГОВОГО ИМУЩЕСТВА

Утро понедельника Михаил тоже решил начать с общения с командой. Он собрал коллег в переговорной банка «Центр-инвест». После брейншторма идей по новому проекту автоматизации оценки залогового имущества Михаил зафиксировал план действий и выслал коллегам фоллоу-ап.

Коллеги, благодарю за вовлечение!

Напоминаю, наша задача — **реализовать прототип сервиса, позволяющего оценить стоимость автомобиля и/или объекта недвижимости по их характеристикам с максимально возможной точностью данных**. При оценке стоимости желательно также определять степень экологичности автомашины или жилья, чтобы предлагать владельцам экологичных объектов более выгодные условия по кредитам. Мы планируем внедрить сервис в ближайшие 3–4 месяца. Ограничений по технологиям и способам реализации задачи, а также по детализации дополнительного функционального наполнения нет.

При реализации сервиса нам нужно учитывать следующее:

- Поиск данных - одна из основных задач любого дата-аналитика. Примеры ключевых слов для поиска подходящих датасетов: Cars, Real Estates, Estate price. Этим можно не ограничиваться, а предложить альтернативы, главное - обосновать свой выбор.
- Основные требования к конечному результату: нужна высокая точность определения стоимости объектов (> 80% на тестовых выборках), а также создание MVP для проверки жизнеспособности прототипа сервиса. ПО должно иметь UI (Web) для кредитного сотрудника, который на основании полученных от клиента данных (если машина — марка/модель/пробег/фото; если недвижимость — адрес/площадь и пр.) оценит стоимость предлагаемого залога.



- Важно оценить потенциальный экономический эффект от создания сервиса. Нашей финальной целью является снижение этого показателя на 70%, но на этапе разработки прототипа допускается менее значимое снижение издержек.

Проект важен для «Центр-инвеста», так как случаи, когда клиенты предлагают нам залог для получения лучших условий кредитования, возникают часто.

Спасибо!

ПРИЛОЖЕНИЯ

Диалоги с клиентами из банковского чата



ПОЧТА БАНК

