Финальный проект курса “Специализация Data Science” (SkillFactory)

Определение территории дистрибьютора, на которой находится торговая точка. Создание системы для подсказки оператору

Сергей Пигида  
(группа DST-5, DST-0)

2020

# 

# Основные цели и задачи проекта

## Бизнес-кейс

В отделе продаж компании имеется периодический бизнес процесс выплат дистрибьюторам, которые осуществляют продажу продукции в определенной географической области. Для подсчета размера выплат используются данные по продажам торговых точек, которые должны быть отнесены к тому или иному дистрибьютору. Сложность заключается в том, что в одну торговую точку могут продавать несколько дистрибьюторов, при этом необходимо точку однозначно отнести точку к конкретному дистрибьютору.

Возможность различать торговые точки по географическому принципу затруднена: неявные границы территорий внутри больших городов, пограничные точки на границах областей, дистрибьюторы работают в одной области, но в разных форматах торговых точек и т.п.

## Постановка задачи

Для обеспечения бизнес процесса в каждом регионе имеется оператор, который вручную назначает торговым точкам правильный филиал нужного дистрибьютора. Чтобы снизить человеческий фактор и ускорить процесс формирования выплат необходимо внедрить систему для формирования подсказки оператору о том, к какому филиалу дистрибьютора принадлежит новая точка.

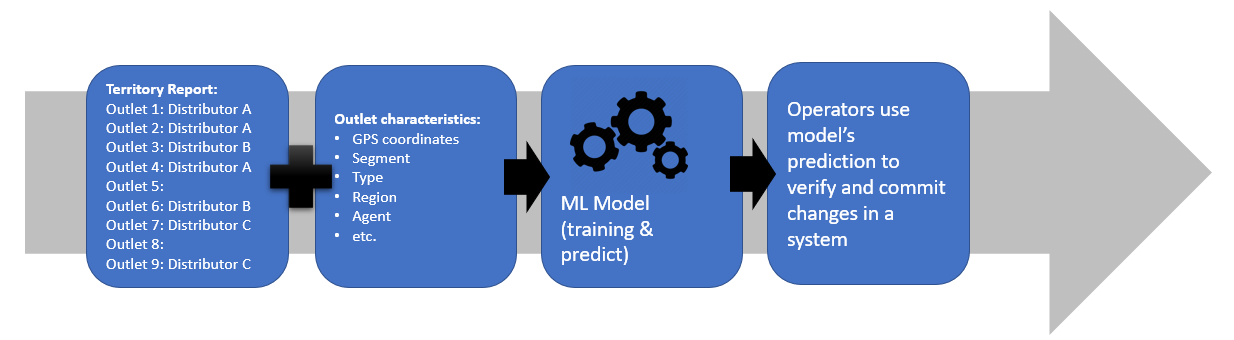
Требуется в имеющийся отчет Territory Management Report, с которым работает оператор, добавить новые колонки в которых только для новых торговых точек показывать филиалы дистрибьюторов, к которым относится данная торговая точка. Предсказание выполнять на основе доступной информации о торговой точке: регион, сеть, формат магазина, координаты и т.п.

## Описание решения

Поскольку бизнес-процесс имеет особый легальный статус, необходимо минимальное вмешательство. Для решения задачи будет разработана программа, которая в качестве входной информации будет использовать два Excel-файла: исходный отчет Territory Management Report и файл с координатами торговых точек. На выходе будет формироваться новый файл Territory Management Report Updated с дополнительными колонками информации.

В процессе выполнения программы поступившая информация о точках будет очищена и подготовлена для обучения в модели машинного обучения. Обучение необходимо выполнять при каждом запуске программы, из-за периодических изменений территорий дистрибьюторов. На выходе модели будет оцениваться качество работы модели, а затем запускаться предсказание для новых торговых точек.

Дополнительная информация в обновленном отчете будет использоваться операторами для проверки и утверждения. По завершению работы оператор будет вручную удалять дополнительную информацию, поскольку отчет в неизменном виде участвует в дальнейшем процессе, в том числе используется для загрузки в систему для последующего выполнения расчета.



# Работа над проектом

## Этапы работ

Данный проект был заявлен бизнесу как быстрое решение существующей проблемы, а также для демонстрации возможности применения методов машинного обучения для решения различных задач отдела продаж. От возникновения идеи до запуска прототипа в работу прошло пять недель. Можно выделить следующие этапы работ:

1. Возникновение идеи
2. Получение данных и их анализ
3. Быстрое обучение модели для подтверждения возможности решения задачи и демонстрация результатов бизнесу
4. Формирование и согласование требований с бизнесом
5. Работа над решением с демонстрацией промежуточных результатов и приемка финального результата
6. Обучение пользователей
7. Запуск прототипа системы в эксплуатацию
8. Получение обратной связи

## Информация о данных

Входная информация поступает из двух источников-файлов, которые формируются в существующей системой отчетности. В процессе обработки файлы могут быть объединены по полю “Код торговой точки”.

Файл **Coordinates** содержит координаты торговых точек. Не все торговые точки имеют координаты. В процессе подготовки данных для обучения модели, точкам без координат будут назначены усредненные координаты соседних точек, которые находятся в том-же населенном пункте, что и точка без координат.

Файл **Report Territory Management** является специальным отчетом, который требуется для работы оператора процесса, а также является входным файлом для последующего расчета. Файл содержит следующую информацию по точкам:

1. Иерархия дистрибьюторов (Регион - Дистрибьютор - Филиал)
2. Иерархия сотрудников
3. Адресный классификатор (пять уровней)
4. Характеристики точки: Сеть, Тир, Сегмент, Формат и т.п.
5. Информация о посещении точки
6. Код филиала на территории которого находится торговая точка (этот код необходимо предсказать, если отсутствует)

**Подробный анализ данных, включая сделанные выводы, а также очистка и подготовка выполнены в рабочем Jupyter ноутбуке *Data analysis and pre-processing.ipynb*.**

## Реализация и продукционирование

Для создания программы был выбран язык Python, поскольку имеет библиотеку для работы с Excel-файлами (Openpyxl), а также библиотеку для работы с машинным обучением (Scikit-Learn). Для продукционирования и передачи программы конечному пользователю Python скрипт был скомпилирован в выполняемый файл (библиотека Pyinstaller), для запуска на корпоративных компьютерах с операционной системой Windows 10. Решение продиктовано необходимостью быстрого внедрения прототипа, а также небольшим количеством пользователей (до десяти человек).

### Git-репозиторий

Финальный скрипт сохранен в папке source/ репозитория <https://github.com/pigidser/Territory-Prediction>.

## Обратная связь и дальнейшее развитие проекта

Запуск системы и предоставление операторам электрнного помощника, для определения к какой территории дистрибьютора относится точка, было положительно воспринято. Операторы заявили, что стало удобнее работать. Кроме того, очень удобно, что предлагается три варианта кодов доставки, а также цветовая дифференциация правильности ответа: если вероятность класса больше 0.75, то заливка ответа зеленая, если больше 0.5, но меньше 0.75, то желтая. Красной заливкой обозначены все предсказания с определенной вероятностью ниже 0.5

Также работа с измененным отчетом помогла обнаружить человеческие ошибки, которые были допущены ранее.

Из моментов к улучшению обозначено, что скрипт работает 10 минут, а хотелось бы быстрее. Также программа чрезвычайно требовательна к памяти и на компьютере должно быть минимум 8 гигабайт RAM. При этом 90% времени занимают операции загрузки из Excel-файлов, а также обновление финального отчета в Excel-файле.