Исследовательский анализ данных (EDA – Exploratory data analysis) для джуниор-аналитиков в компании «Y Analizer»

Процесс исследования данных – это процесс, для которого требуется набор исторических показателей, например, выгрузка из базы данных, с которой впоследствии можно произвести манипуляции и извлечь определенные зависимости. Эти зависимости главным образом помогут руководителям бизнеса принять наиболее выгодные решения и повысить показатели.

Исследовательский анализ данных может быть произведен в системе электронных таблиц (Excel, Pages), при помощи Системы управления базами данных (SQLite, PostgreSQL), в системах визуализации (Power BI, Tableau), в системе пошаговой интерпретации (Jupyter Notebook, IPython), с использованием языка программирования (Python, Java), а также в системах статистических исследований (R).

В последнее время все большей популярностью пользуется пара Jupyter Notebook и язык программирования Python.

Процесс исследовательского анализа данных состоит из следующих основных этапов:

1. **Загрузка библиотек и данных.**

На данном этапе необходимо загрузить вспомогательные библиотеки:

- Numpy – библиотека для расширенной работы с массивами и матрицами;
- Pandas – библиотека на основе Numpy для работы с панельными данными.

Далее следует загрузить данные для исследований и обработки в различных форматах (csv, xlsx, xls, json). Гибкость языка Python позволяет преобразовывать табличные данные практически любого формата или подключаться и формировать запросы напрямую из базы данных (библиотека Sqlalchemy). Дополнительно данные можно извлекать с помощью парсинга сайтов (библиотека Requests, Re, BeautifulSoup) или запросами через API-интерфейс различных сайтов и сервисов (библиотека Requests, Json).

2. Подготовка данных.

Это очень тонкий и важный процесс, во время которого нужно бережно предобработать данные, выявить аномалии, избавиться от дубликатов и по возможности заполнить или удалить пропуски. Также данные необходимо привести к нужным форматам – это позволит среде обработки правильно их интерпретировать.

Какие данные потребуют такой обработки? Ошибки в логах или на этапах выгрузки, особенности работы предприятия и огромное количество других факторов, которые неизбежны, когда данных очень много.

3. Исследование закономерностей в данных.

На данном этапе необходимо четко сформулировать вопросы, которые аналитик ставит перед собой для решения задач. Здесь пригодятся теоретические знания в математическом анализе и статистике. Они должны помочь аналитику сделать верные выводы, найти закономерности и на их основании сформулировать гипотезы.

4. Проверка статистических гипотез.

На данном этапе аналитик формулирует гипотезы и проверяет их уместность.

Библиотека Scipy дает аналитику набор инструментов, в которых содержатся все известные статистические тесты и математические методы. Все это позволяет сделать правильные выводы и выразить рекомендации для последующих действий бизнеса. Бывает и обратная ситуация, когда у бизнеса есть предположения, которые аналитику требуется проверить и построить выводы.

Проведение таких тестов увеличивает ценность аналитика как важного специалиста в рамках современных data-driven компаний и позволяет предприятию значительно минимизировать издержки. Например, определить оптимальное количество звонков для колцентра.

5. Визуализация данных.

Данный этап может возникать в разных частях анализа данных. Аналитик может визуализировать полученные результаты с помощью соответствующих библиотек: Matplotlib, Seaborn, Plotly, Bokeh. Это нужно главным образом для упрощения восприятия как самим аналитиком, так и командой, которая нуждается в результатах исследования, но не обладает достаточными компетенциями. Также на данном этапе аналитик готовит презентацию, с которой может выступить и рассказать о своих умозаключениях на собрании. Помимо всего прочего можно создать «Дашборд» – интерактивную страницу с дружественным интерфейсом, которая позволит практически любому желающему манипулировать

предобработанными данными и использовать информацию для решения собственных задач.

6. Автоматизация данных.

На этом этапе аналитик, обладая определенными навыками разработки, может построить пайплайн автоматизации, который, например, позволит вести автоматическую отчетность раз в месяц и очень сильно уменьшит количество рутинных задач.

Для автоматизации необходимы навыки программирования скриптов (Python), владение командной строкой (Bash) и сервисом периодизации задач (Ctontab). Эти инструменты также позволяют агрегировать данные в более емкие наборы и таким образом оптимизировать быстродействие рутинных процессов.

7. Выводы и заключения.

На данном этапе аналитик строит финальные выводы, говорит о перспективах, предлагает варианты решения задач, опираясь на полученные результаты.