Федеральное государственное образовательное бюджетное

учреждение высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**(Финансовый университет)**

**Институт развития профессиональных**

**компетенций и квалификаций**

**ИТОГОВАЯ РАБОТА**

|  |  |
| --- | --- |
| Группа обучения | **«**AD\_potok\_2\_11.05.23**»** |
| Срок обучения | **«**11.05.23**»** |
| **Козеровский Максим Витальевич** | |
| Номер Кейса | **«10»** |
| Название Датасета | **«Продажа авиамоделей в США за период 2017–2019 гг.»** |

Москва 2023 г.

Ссылка на Excel файл кейса № 10

<https://docs.google.com/spreadsheets/d/1zTyJZXKnN38o431OAIX958pTTSSZR9PD/edit?usp=sharing&ouid=112933034334416794803&rtpof=true&sd=true>

Ссылка на очищенные и подготовленные данные:  
https://drive.google.com/file/d/1Autj\_hYQ8YXhmakiW5\_Bl1do4O5E5aJa/view?usp=sharing

Ссылка на файл KNIME:  
https://drive.google.com/file/d/18wCkYPY822GLXV71J6Ye26lHa2SAohX1/view?usp=sharing  
Ссылка на Colaboratory файл:  
https://colab.research.google.com/drive/1Ey5xhJ2XfJOm9ehGDDDf2fwx8OdcEU2Y?usp=drive\_link  
Ссылка на Tableau:

Кейс

Данные выбранного кейса включают в себя информацию о продажах авиамоделей в US за период с 2017 по 2019 годы. Каждая строчка датасета это уникальный заказ.

Столбцы в нём следующие –

RangeIndex: 377741 entries, 0 to 377740

Data columns (total 26 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 OrderNumber 377741 non-null object

1 OrderDate 377741 non-null object

2 ShipDate 377741 non-null object

3 Quantity 377741 non-null int64

4 UnitPrice 377741 non-null float64

5 DiscountAmount 377741 non-null float64

6 PromotionCode 377741 non-null object

7 ProductName 377741 non-null object

8 ProductCategory 377741 non-null object

9 ItemGroup 377741 non-null object

10 KitType 377741 non-null object

11 Channels 377741 non-null int64

12 Demographic 377741 non-null object

13 RetailPrice 377741 non-null float64

14 StateName 377741 non-null object

15 RegionName 377741 non-null object

16 UsedPromoCode 377741 non-null int64

17 Year 377741 non-null int64

18 SumOrder 377741 non-null float64

19 Profit 377741 non-null float64

20 Marginality 377741 non-null float64

21 ABC\_Class\_2017 377741 non-null object

22 ABC\_Class\_2018 377741 non-null object

23 ABC\_Class\_2019 377741 non-null object

24 XYZ 377741 non-null object

25 RFM 377741 non-null int64

dtypes: float64(6), int64(5), object(15)

По порядку:

Изначально полученные данные. Номер заказа (уникальное значение), дата заказа(год/месяц/день), дата отгрузки товара (год/месяц/день), количество в заказе, закупочная цена, размер скидки, промо-код, название номенклатуры, категория товара, общая группа товара, готовая модель (rtf) или собирать (kit), количество радиоканалов у модели, уровень сложности авиамодели, цена продажи, штат покупки, регион покупки (укрупнённые зоны проживания покупателей).

Данные полученные в процессе обогащения. Столбец параметра показывающий использовал ли покупатель промо-код. Год совершения заказа. Общая сумма заказа. Посчитанная прибыль. Маржинальность. ABC классификация товара по годам. XYZ классификация товара за весь период. RFM классификация штатов где живут покупатели.

**KNIME.**

Для очистки и обработки данных использовалась швейцарская программа KNIME.

При обработке проведенного анализа (ABC, XYZ и RFM) можно сделать вывод что несмотря на то, что общая выручка ритейлера растёт, из года в год сокращается количество номенклатурных позиций приносящие 80%-ю долю выручки, большинство ассортимента имеет устойчивую динамику к понижению объёма продаж. При этом немногочисленные товары категории ‘А’ относятся к группе ‘Z’, они имеют коэффициент вариативности выше 25%, спрос спрогнозировать сложно. Всего два товара со стабильным спросом имеют низкие продажи. По мимо этого покупатели (они разбиты по штатам) зачастую имеют низкую активность и низкий объём покупок.

На основе этих исследований данных можно сделать следующие выводы:

1. Т.к. подавляющее большинство заказов имеют одну позицию, а количество заказов растёт, то отдел маркетинга компании успешно привлекает клиентов.
2. При этом частота покупок со стороны клиентов низкая и так же низкий средний чек. К этому стоит добавить, что большинство клиентов не использует промо-коды. Значит с новыми клиентами плохо ведётся последующая работа. Привлечь нового клиента всегда дороже, чем работать с постоянным.
3. 90% номенклатурных позиций имеют нестабильный спрос. Нестабильность всегда несет риски для бизнеса. За три года сбора данных ассортимент практически не менялся. Имеется только одна авиамодель, которая была введена в продажи в 2018 году и поднялась в категорию “A”.

EDA в KNIME.

Общая динамика за отчетный годы по Выручке и прибыли

рис. 1

Изображение выглядит как текст, График, линия, диаграмма

Автоматически созданное описание

Видно, что при общем росте Выручки, доходность падает начиная с 2019

Цены покупки и продажи товаров (2017, 2018, 2019 годы)

рис.2

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

рис.3

Изображение выглядит как текст, снимок экрана, График, число

Автоматически созданное описание

рис.4

Изображение выглядит как текст, снимок экрана, График, линия

Автоматически созданное описание

При анализе данной визуализации становиться ясно, что доходность падает по причине сильного снижения маржинальности. Часть заказов закрывалась с отрицательным результатом.

Необходимы новые поставщики для уменьшения закупочных цен и возможно для предоставления клиентам нового товара. Как видно из данных компания продаёт одни и те же авиамодели, но это тот тип вещей, что не подразумевают повторные покупки того же самого. Нужно либо стимулировать создание коллекций, либо предлагать дополнительные опции к существующим моделям.

ML в KNIME.

Обучались и тестировались три модели машинного обучения. Это Linear regression, XGBoost Linear regression, Time series analysis в двух видах Auto-SARIMA и ARIMA. Решалась задача предсказания получения прибыли.

Программный продукт позволяет логически расставлять необходимые узлы и обучаться на заранее разбитых на два датафрейма (тренировочный и тестовый) данных. После обучения тестовую часть данных прогоняем через созданную систему предсказания.

Результаты по одному из ключевых показателей R^2 следующие:

1. Linear regression - 0.87
2. XGBoost Linear regression - 0.65
3. Auto-SARIMA - 0.39
4. ARIMA – (-0.3)

**Colaboratory.**

Для EDA использовались следующие библиотеки Python pandas-profiling и

Python dataprep. Обе библиотеки предоставляют интерактивные визуализации и таблицы с отчетами. Основной инструмент, который используют эти библиотеки это группировка, подсчет суммы, среднего, минимум, максимум и прочие агрегирующие функции.

Например:

- Топ 5 дат заказов это три дня в сентябре и два в ноябре. Возможно, это рядом с какими-то национальными праздниками или прошла удачно рекламная кампания.

- Если посмотреть на даты отправления заказов, то видно в какие периоды происходили сбои и это не всегда периоды с большим количеством покупок.

- Количество штук номенклатуры в заказах от 1 до 3. Подавляющее большинство — это покупка одной авиамодели.

- В закупке лидируют дорогие товары.

- 84% покупок совершаются без скидки.

- Эти 3 наименования товара составляют более 50% оборота компании. Tailspin Heli - Max Pro Flight - 6ch - 26.3%, 6CCP-A Helicopter - 12.7%, Piper Cub 4 Channel - 12.4%

- С большим отрывом лидирую две категории товаров Collective pitch - 147083 шт. и Trainer - 115627 шт.

- Чуть больше покупают вертолеты чем самолеты

- Готовые модели (RTF) покупают 56.0% людей, собирать (KIT) любят 44.0%

- В основном предпочитают профессиональные (Professional - 147083 шт.) и продвинутые (Intermediate - 127671 шт.) модели с большим количеством каналов связи и подключения 6 и 4.

- Цены продаж сильно коррелируются с товарами из сегмента с дорогой закупкой.

- California, Florida и New York штаты лидеры покупок.

- По годам и количеству заказов 2019–203620 шт., 2018–129682 шт., 2017–44439 шт.

- Доход за эти годы. 117728 заказов, а это 31.2% закончились с прибылью 0.

- Моржинальность имеет низкий процент. 117492 продажи (31.1%) совершены с нулевой морженальностью.

- В 2017 году по классу «А» 233501 продажи, в 2018–220381, а в 2019 уже 194006.

- 91.0% заказов пришёлся на товары категории «Z»

- RFM классификация показывает, что на клиентов кластера 311 приходится 77.9% покупок, т. е. редкие и дешёвые, из группы 333 совершилось 11.6%, 322–10.1%. Есть более 20% клиентов, с которыми можно поработать и улучшить положение дел фирмы.

Как и в случае с предыдущим анализом данных (в KNIME), проблема видна очень хорошо, это низкая рентабельность совершаемых сделок. Необходимо искать новых партнеров и снижать закупку или менять ассортимент. Ещё вариант уйти в сегмент элитных продаж и торговать только дорогими моделями. Меньше, но с более высокой доходностью.

AutoGluon в Colaboratory.

Для работы с моделями машинного обучения был выбран фреймворк AutoML -AutoGluon.

Система была развернута в блокноте Colaboratory, разбит датафрейм на тренировочный и тестовый. При запуске обучения так же была выбрана целевая колонка Прибыль.

AutoGluon подбирает модели под предоставленные данные и производит вычисления и подбор гиперпараметров.

Как и в первом случае авто ML выбрал модели связанные с regression, это модели KNNModel, RFModel, LGBModel, CatBoostModel, WeightedEnsembleModel. В итоге лучшей моделью оказался CatBoost. Один из ключевых показателей R^2 у данного метода составил – 0.999.

Показатель высокий, но есть большая вероятность, что модель переобучилась.

Из плюсов это простота запуска процессов, буквально несколько строк кода. Но расчёт идёт очень продолжительное время. Модель в KNIME имеет хорошие результаты, представляет из себя low-code подход и тренирует модель гораздо быстрее.