Степанова Е.Н.

Кейс № 3.

Курс: АБД-2022-1

**Оглавление**

[Описание кейса](#_gjdgxs) **2**

[Обогащение данных кейса ABC, XYZ и RFM - анализом.](#_30j0zll) **4**

[EDA, расширенный анализ данных, обучение моделей регрессии с использованием библиотек Python.](#_1fob9te) **12**

[Визуализация.](#_3znysh7) **19**

[Построение модели в Google BigQuery](#_j3nknlvqkcmh) **30**

[Список ссылок.](#_tyjcwt) **35**

# **Описание кейса**

Для итоговой работы взят кейс № 3 из предложенного списка наборов данных. Данные содержат информацию о работе магазина одежды в период с 2009 по 2013 год, которые хранятся в 5 таблицах.

Таблица **Orders**

OrderID - идентификатор заказа,

OrderDate - дата заказа,

Month - месяц,

Year - год,

CustomerID - идентификатор клиента,

EmployeeID - идентификатор сотрудника,

ShipperID - идентификатор грузоотправителя,

ProductID - идентификатор продукта,

Sales - продажи в $,

Costs - расходы в $,

Profit - прибыль в $,

Quantity - количество,

Discount - скидка в $,

Freight - стоимость перевозки в $.

Таблица **Categories**

CategoryID - идентификатор категории товаров,

Category - категория товаров,

Description - описание категории товаров.

Таблица **Products**

ProductsID - идентификатор продукта,

CategoryID - идентификатор категории,

Product - продукт,

SupplierID - идентификатор поставщика.

Таблица **Suppliers**

ID - идентификатор,

SupplierID - идентификатор поставщика,

Supplier - поставщик,

SupplierContact -контакт с поставщиком,

SupplierCountry - страна-поставщик.

Таблица **Employees**

EmployeeID - идентификатор сотрудника,

Extension -расширение,

EmployeeName - имя сотрудника,

Hire Date -дата найма,

Office - дата найма,

Reports To – перед кем отчитывается сотрудник,

Title - название должности,

Year Salary - годовая зарплата.

# **Обогащение данных кейса ABC, XYZ и RFM - анализом**.

Ссылка на файл Knime:

<https://drive.google.com/file/d/1hMoYpUv7_-qZMK1BEFf8WGQ_BjcX-N4T/view?usp=sharing>

1. Выполнение задания производилось на платформе KNIME. В результате загрузки данных в формате .csv были получены следующие ошибки:

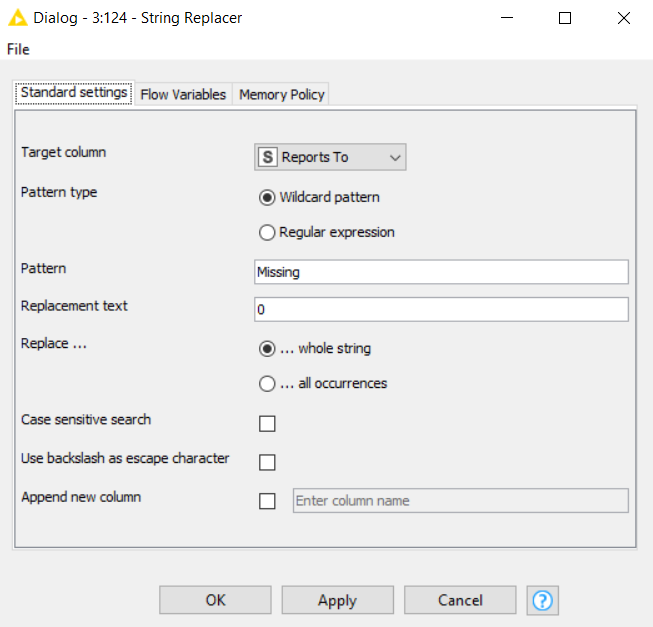
* ошибка при преобразовании формата данных в столбце OrderDate,
* данные столбцов Sales, Costs, Profit, Discount, Freight содержали вместо “8” - символ “В”, Category, Product - вместо апострофа символ “В”.

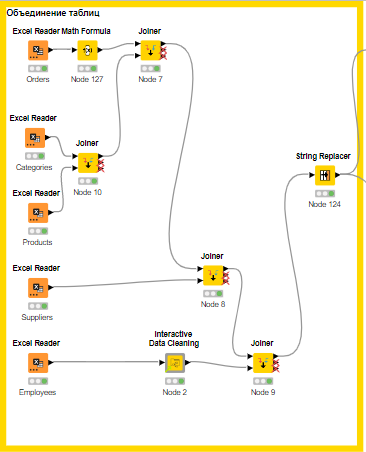
Решением этих проблем стало загрузка данных в формате Excel.

1. Объединение таблиц **Categories** и **Products** по идентификатору категории товаров CategoryID. Полученные данные были объединены таблицей **Orders** по идентификатору продукта ProductsID. Затем данные были объединены таблицей **Suppliers** по идентификатору поставщика SupplierID. Затем полученные данные были объединены с таблицей **Employees** по идентификатору сотрудника EmployeeID.
2. Параметр Discount указан в долларах, чтобы значение отображалось в процентах было добавлен новый столбец с помощью формулы

DiscountPercent=round($Discount$/$Sales$,2)

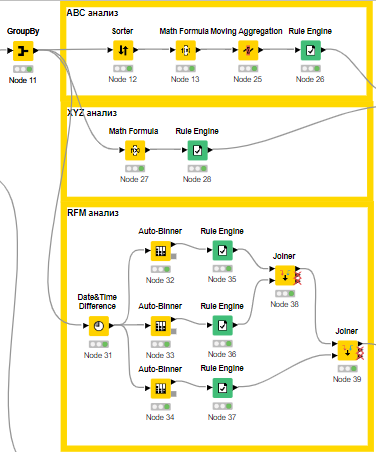
4) Пропущенные значения параметра "Reports To" соответствует значению Title=President, так как президент никому не подчиняется то можно заполнить пропущенные значения нулем (0).



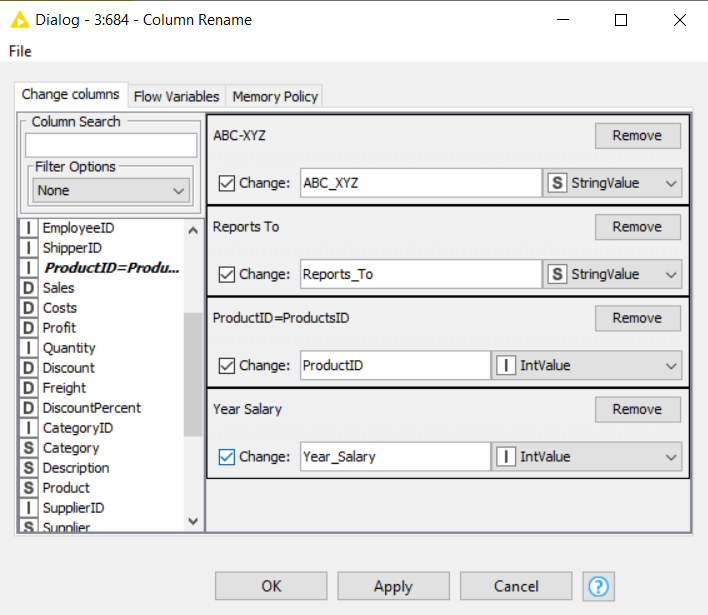


1. Далее был проведен ABC, затем XYZ и RFM анализ по клиентам.

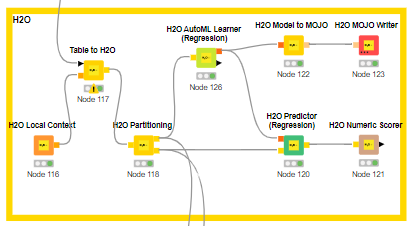
Более подробно рассмотрим ABC и XYZ анализ в разделе 3.

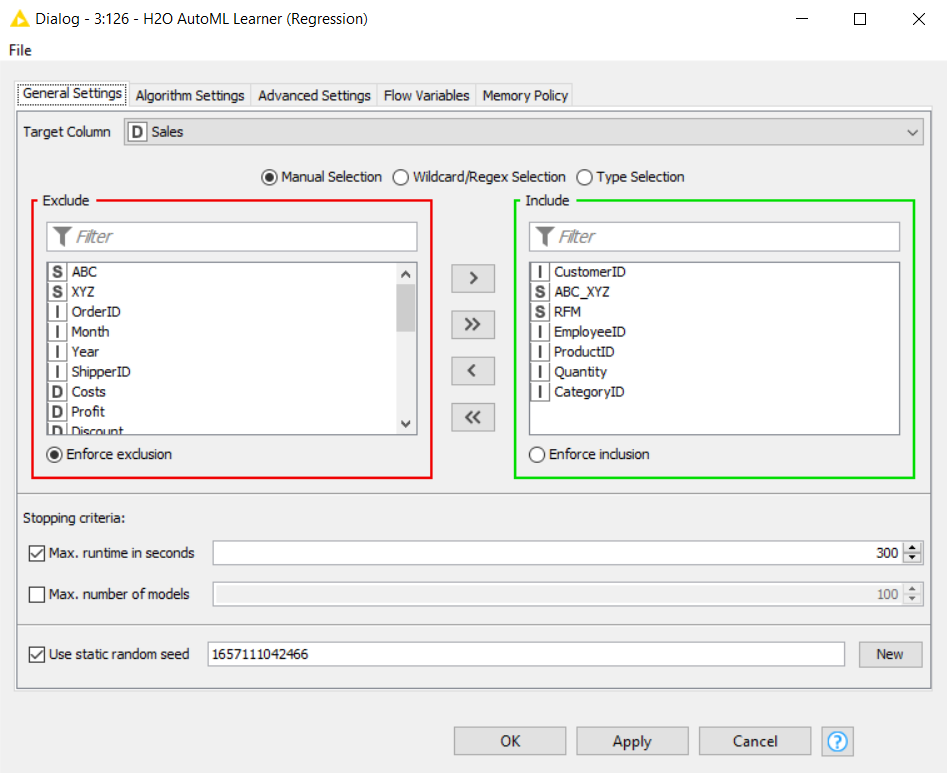


1. Далее было произведено объединение первоначальных данных с данными полученными в результате ABC-XYZ анализа и RFM анализа.
2. Название 4 столбцов были переименованы для выгрузки данных в Google BigQuery без ошибок.

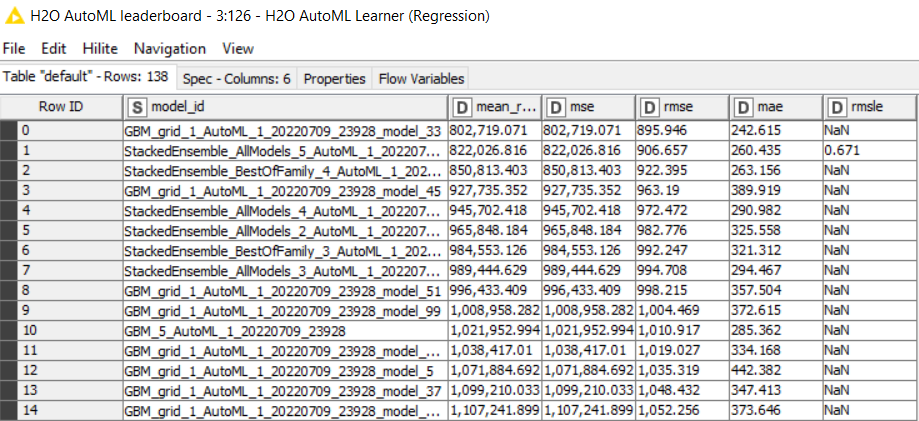


1. Далее была произведена выгрузка обогащенного датасета в формате Excel для дальнейшей работы.
2. Прогноз продаж с помощью AutoML H2O в зависимости от следующих параметров: ABC-XYZ, RFM, CustomerId, EmploeeId, ProductId, Quantity и CategoryId. Датасет был разделен на тренировочные и тестовые данные в пропорции 80% на 20%.

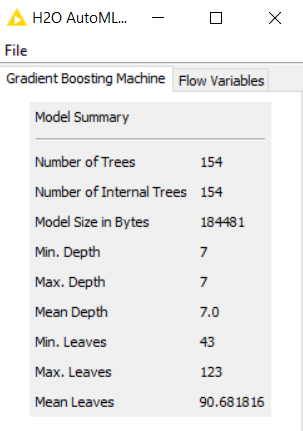




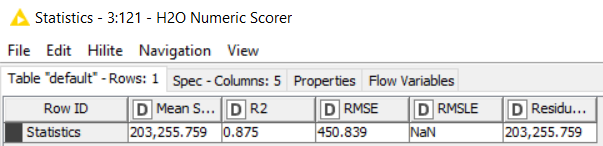
1. Было проверено 138 моделей. Лучшая модель - Gradient Boosting



Параметры модели:



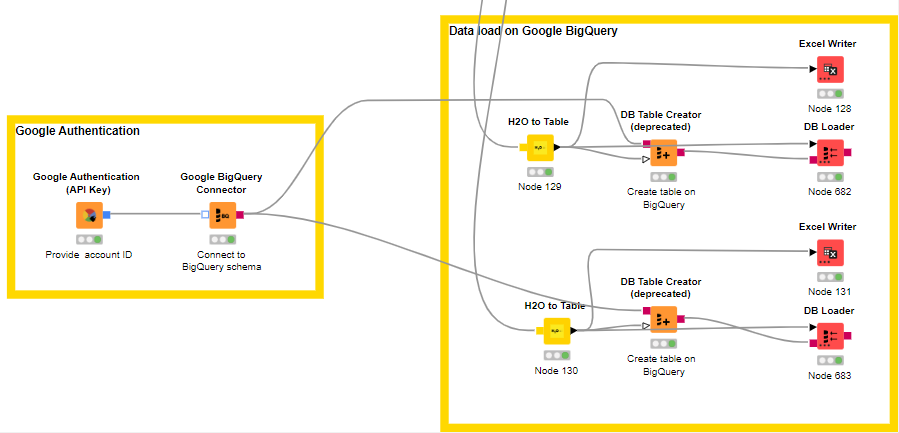
10) В результате сравнения продаж и предсказанного значения продаж получаем таблицу:



В результате получаем R2 = 0.875. Данная модель была сохранена на диске.

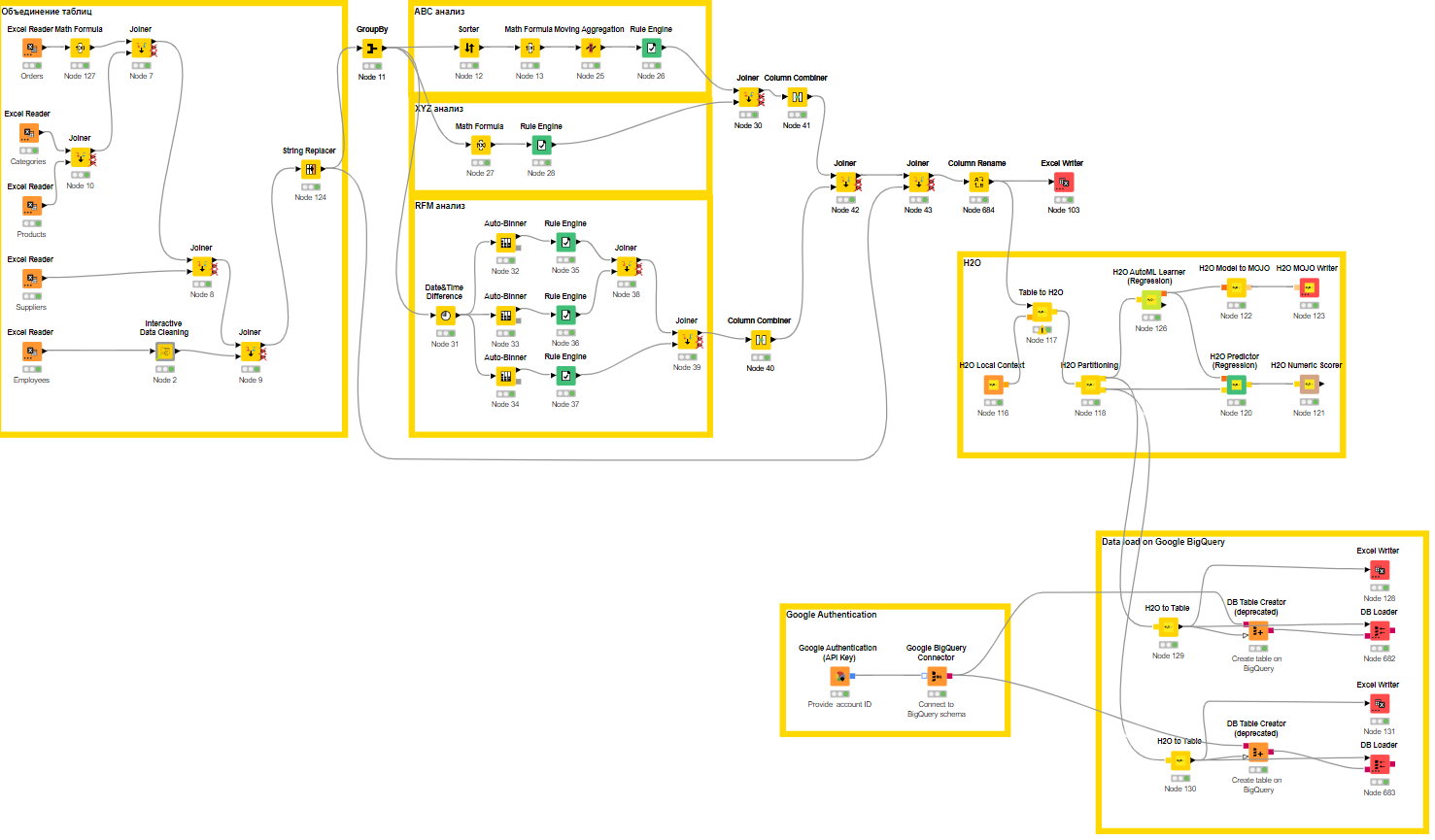
<https://drive.google.com/file/d/1s6amrtXLlupoDJGb-NsPr6BYeEpZP9S2/view?usp=sharing>

1. Далее выбо выполнено подключение к Google BigQuery, запись и преобразование в таблицу данных для моделирования и данных для предсказания, а также выгрузка в Google BigQuery.



Ссылка на схему Knime в хорошем качестве:

<https://drive.google.com/file/d/1euozPUc0MIiUT1hqbaik_PLQERPdaCn3/view?usp=sharing>

****

# **EDA, расширенный анализ данных, обучение моделей регрессии с использованием библиотек Python.**

Блокнот Google Colab:

<https://colab.research.google.com/drive/1RX-hMqlXuSWr7DbIvHVcUyl0r2yKxTl6?usp=sharing>

В рамках работ в блокноте Google Colab был проведен исследовательский анализ данных с использованием библиотеки pandas-profiling, который был выгружен из Google Colab.

В результате было выявлено следующее:

* высокая корреляция параметров Sales и Profit, Sales и Cost, что довольно ожидаемо,
* параметр Discount имеет 61% нулевых значений, но это совершенно нормально,
* пропущенные значения отсутствуют.

Полный отчет pandas-profiling расположен по ссылке:

<https://docs.google.com/document/d/1e6lFV4r9POX_Q4lCY_wICg-7YeFABJdMRdq6z5t8Doo/edit?usp=sharing>

Параметры OrderID, Month, Year, EmployeeID, ShipperID, CategoryID, Description, SupplierContact, SupplierContact, SupplierContact, Extension, Office, Year Salary, Supplier, Reports To, Hire Date являются неинформативными для нашей модели. ABC и XYZ тоже удаляем, вместо этих параметров используем ABC-XYZ.

Можно было не удалять EmployeeID и CategoryID, но для большей сложности вместо них используем EmployeeName и Category.

Составляем второй датасет, который содержит только интересующие нас параметры. Разделяем и группируем.

С помощью sklearn строим модель:

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.fit\_transform(X\_test)

from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(n\_estimators = 400, random\_state = 50, max\_depth=15)

rf.fit(X\_train, y\_train.values.ravel())

y\_pred\_test = rf.predict(X\_test)

from sklearn.metrics import r2\_score

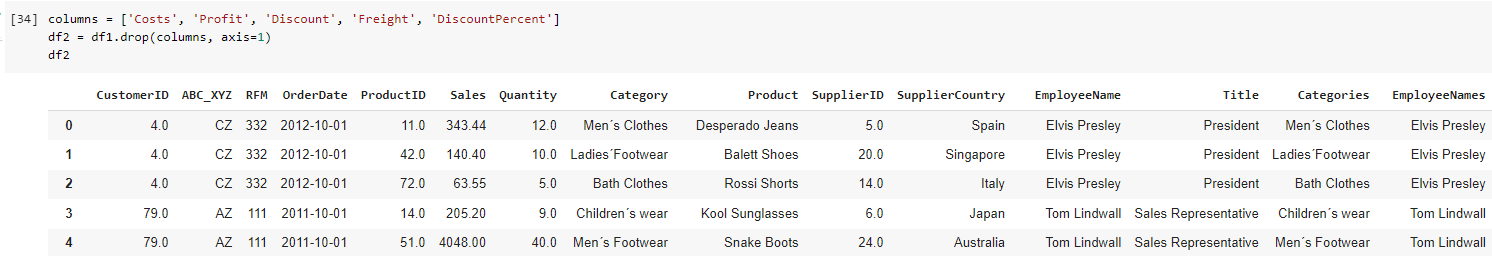
r2\_score(y\_test, y\_pred\_test)

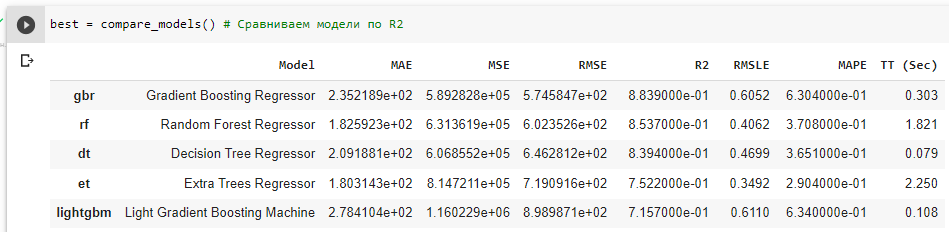
Получаем очень низкое значение r2\_score = 0.4116753361999924.

Для машинного обучения используем Pycaret.

Так как параметр Sales сильно коррелирован с Costs, Profit, Discount, Freight, DiscountPercent, то мы удаляем эти столбцы.

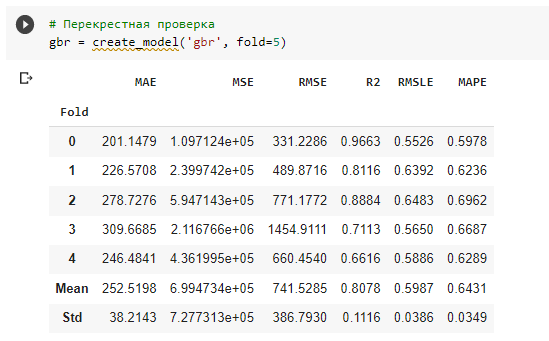
Далее разделяем датасет на данные для моделирования (95%) и данные для предсказаний (5%). Затем строим модель и сраниваем.

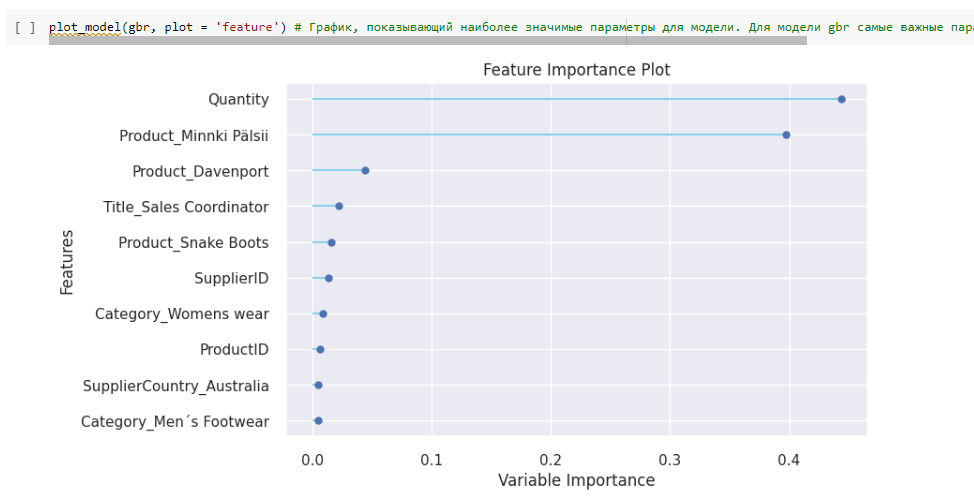


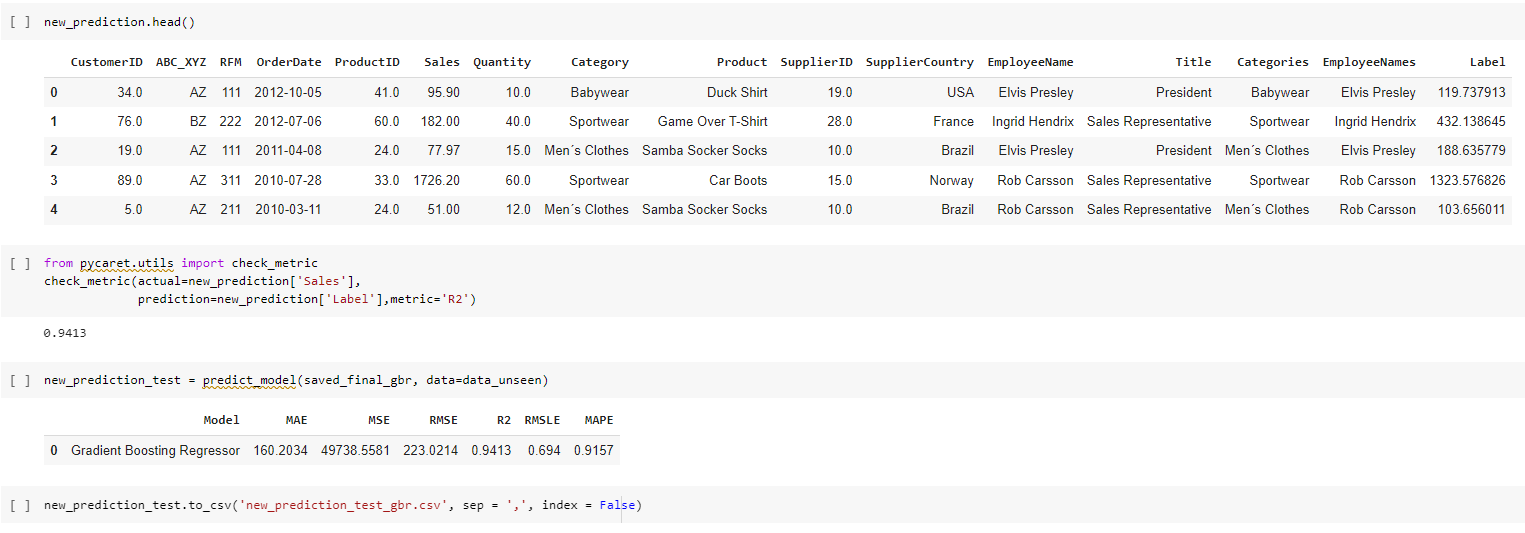


Согласно таблице, лучшие модели - Gradient Boosting Regressor (R2=0.88) и Random Forest Regressor (R2=0.85).

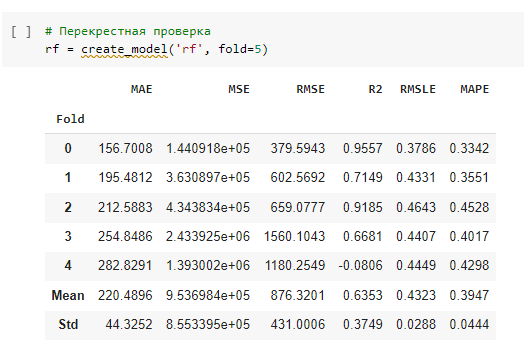
Рассмотрим модель Gradient Boosting Regressor.

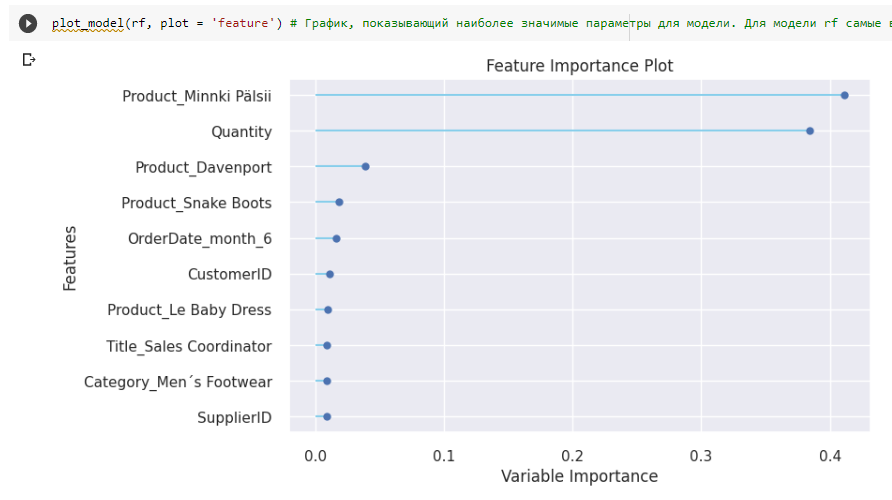






Рассмотрим модель Random Forest Regressor.







Модели **Gradient Boosting Regressor** и **Random Forest Regressor** имеют высокое значение R2 0,94 и 0,96 соответственно. Для этих моделей самыми важными параметрами являются **Quantity, Product, ABC\_XYZ**.

Модели близка к переобучению, предполагаю что это связано с тем, что наш датасет имеет маленький объем 2172 строки.

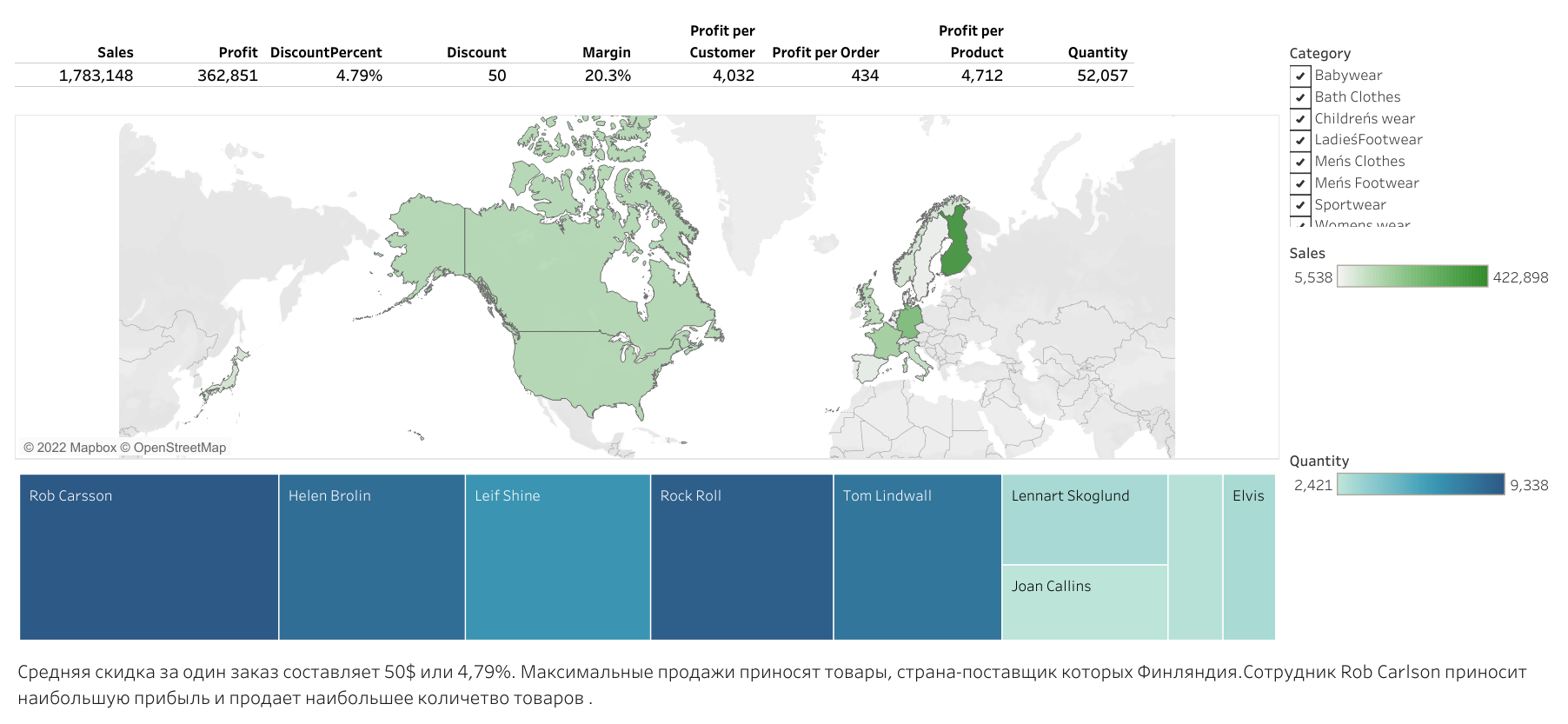
# **Визуализация**

В рамках работы был проведен анализ данных с использованием BI-платформ Tableau, созданы интерактивные отчеты, оформлены дашборды, опубликованные в аккаунте Tableau Public, а также данные были внедрены в блокнот Google Colab.

К сожалению, географические данные в датасете представлены только для параметра страна-поставщик SupplierCountry, но для примера визуализации мы построим такую карту совместно с древовидной картой по имени сотрудника EmployeeName и количеству продаж.

Ссылка на Tableau:

<https://public.tableau.com/app/profile/ekaterina2379/viz/StepanovaEN/QuantityOrderDate?publish=yes>



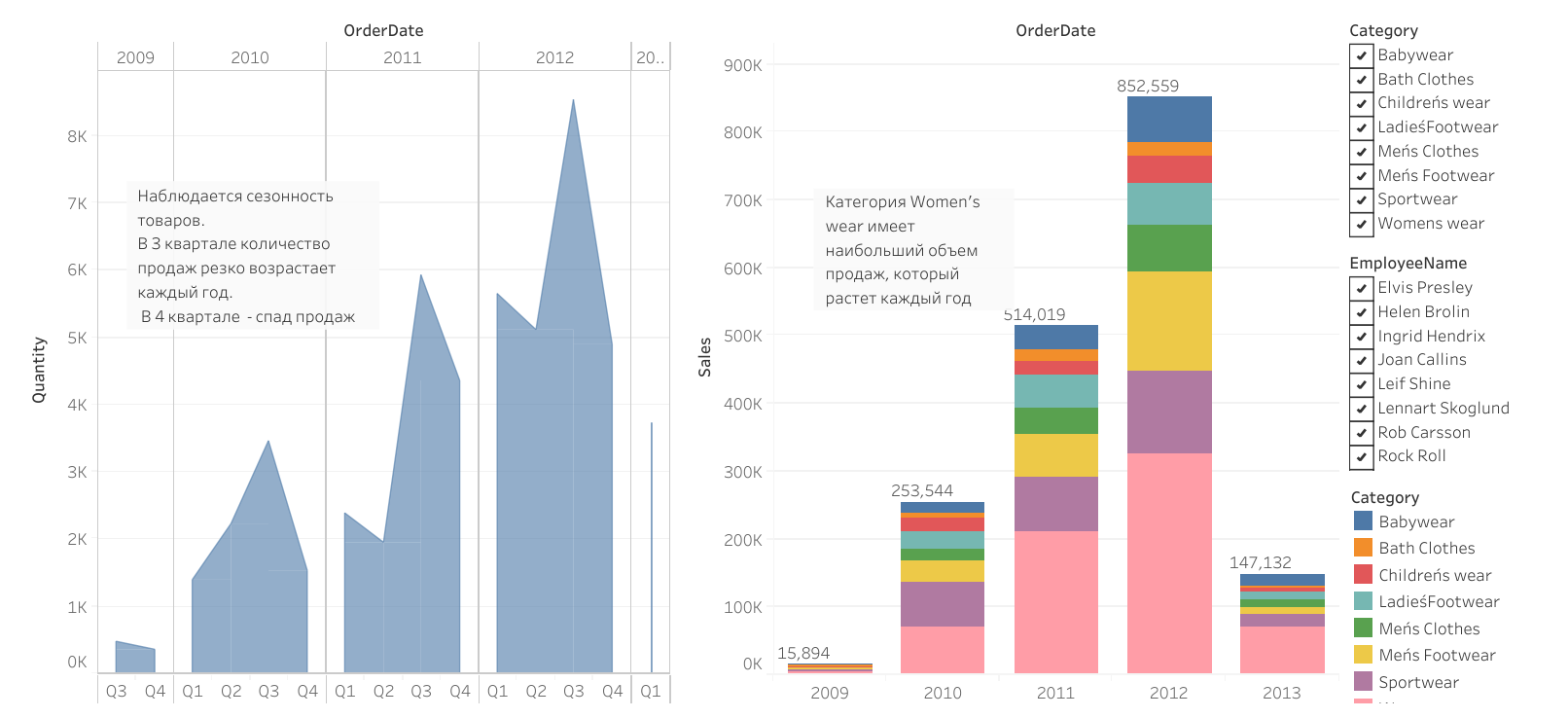
Визуализация географических данных и древовидная карта

Выводы:

1) Средний размер скидки составляет 50$ на один заказ, средняя скидка на заказ составляет 4,79%

2) Максимальные продажи приносят товары, страна-поставщик которых Финляндия.

3) Наибольшую прибыль компании приносит сотрудник Rob Carlson.

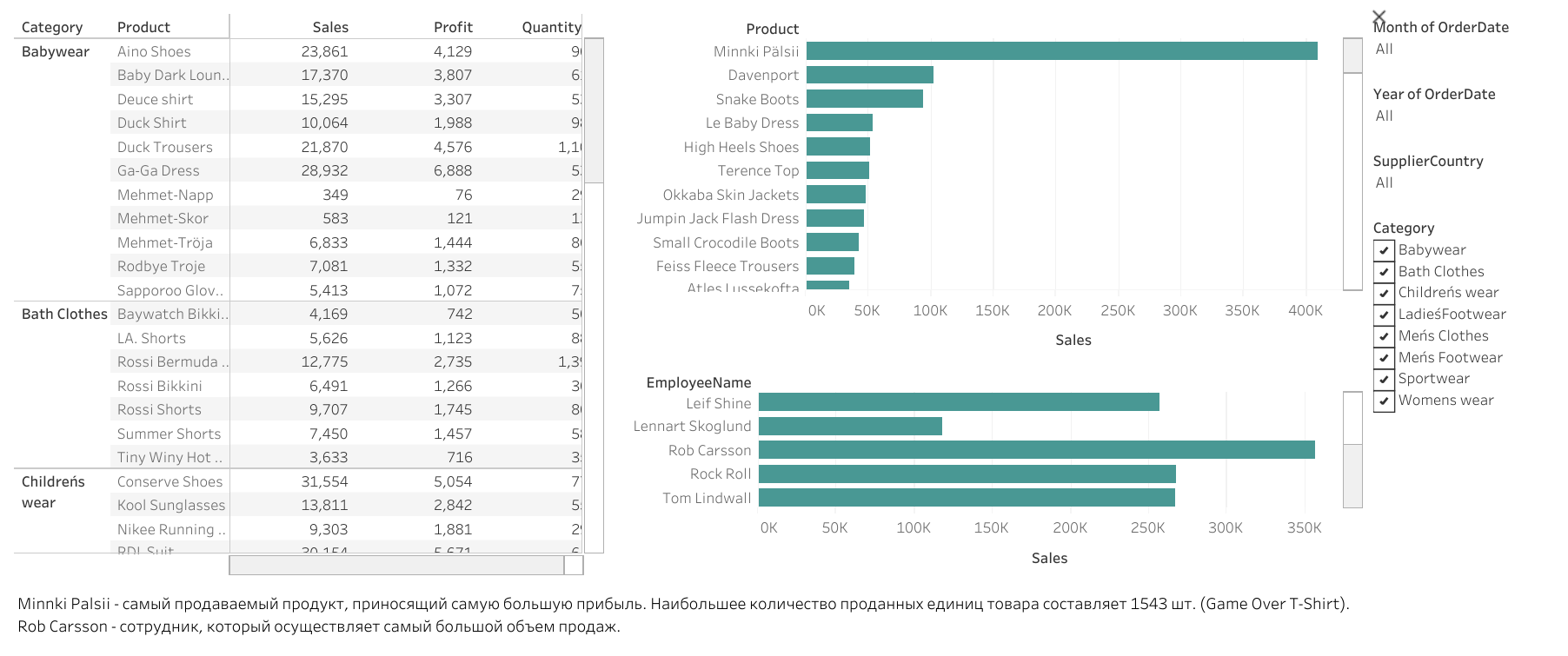


Количество и объем продаж в зависимости от даты заказа

Выводы:

1) На графике зависимости количества заказанных товаров от даты заказа видим, что товары являются сезонными. В 3 квартале количество продаж резко возрастает каждый год. В 4 квартале - спад продаж.

2) На столбчатой диаграмме видим, что категория товаров Women's wear имеет наибольший объем продаж, который растет каждый год. Оценить продажи за 2013 год не представляется возможным, так как есть данные только за 2 квартал.

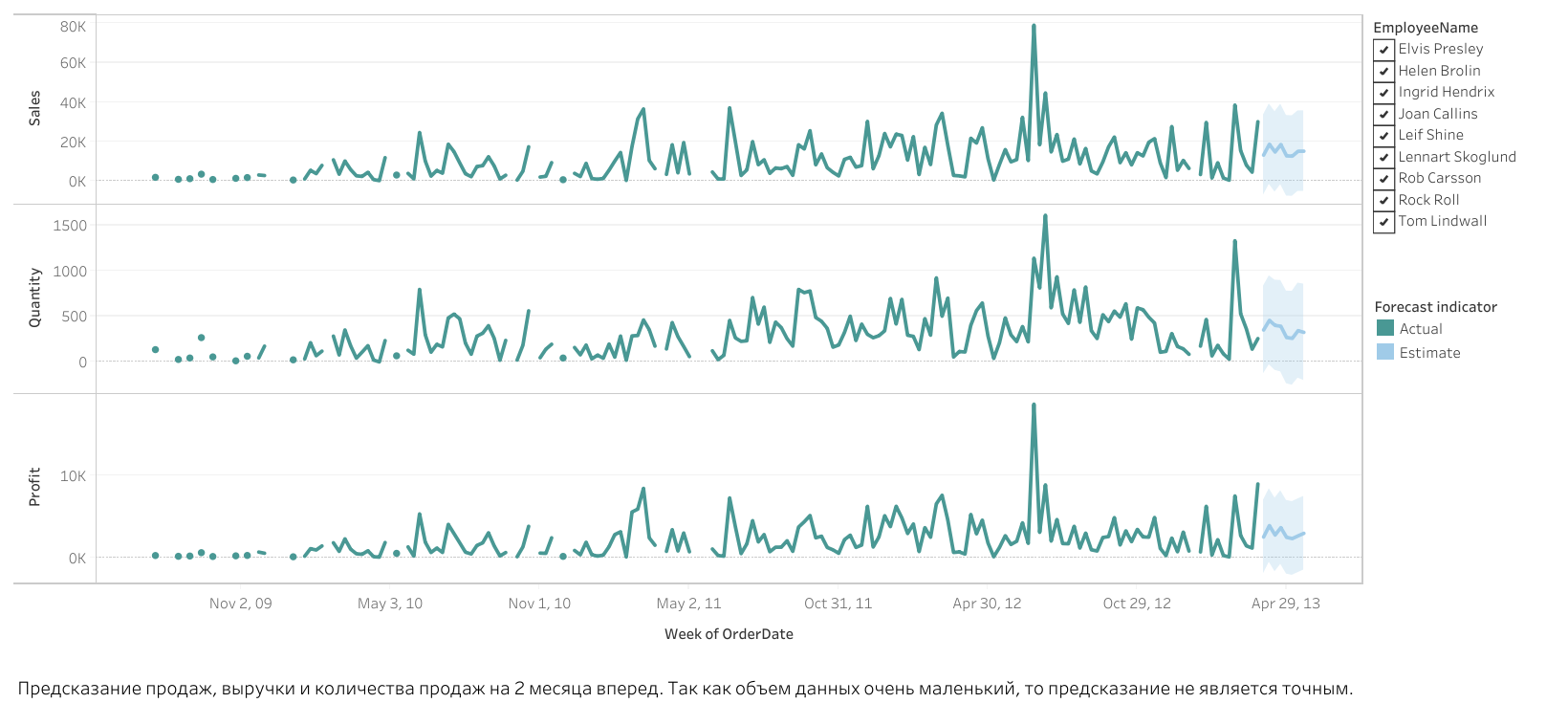


Объем, выручка и количество продаж в зависимости категории и сотрудника

Выводы:

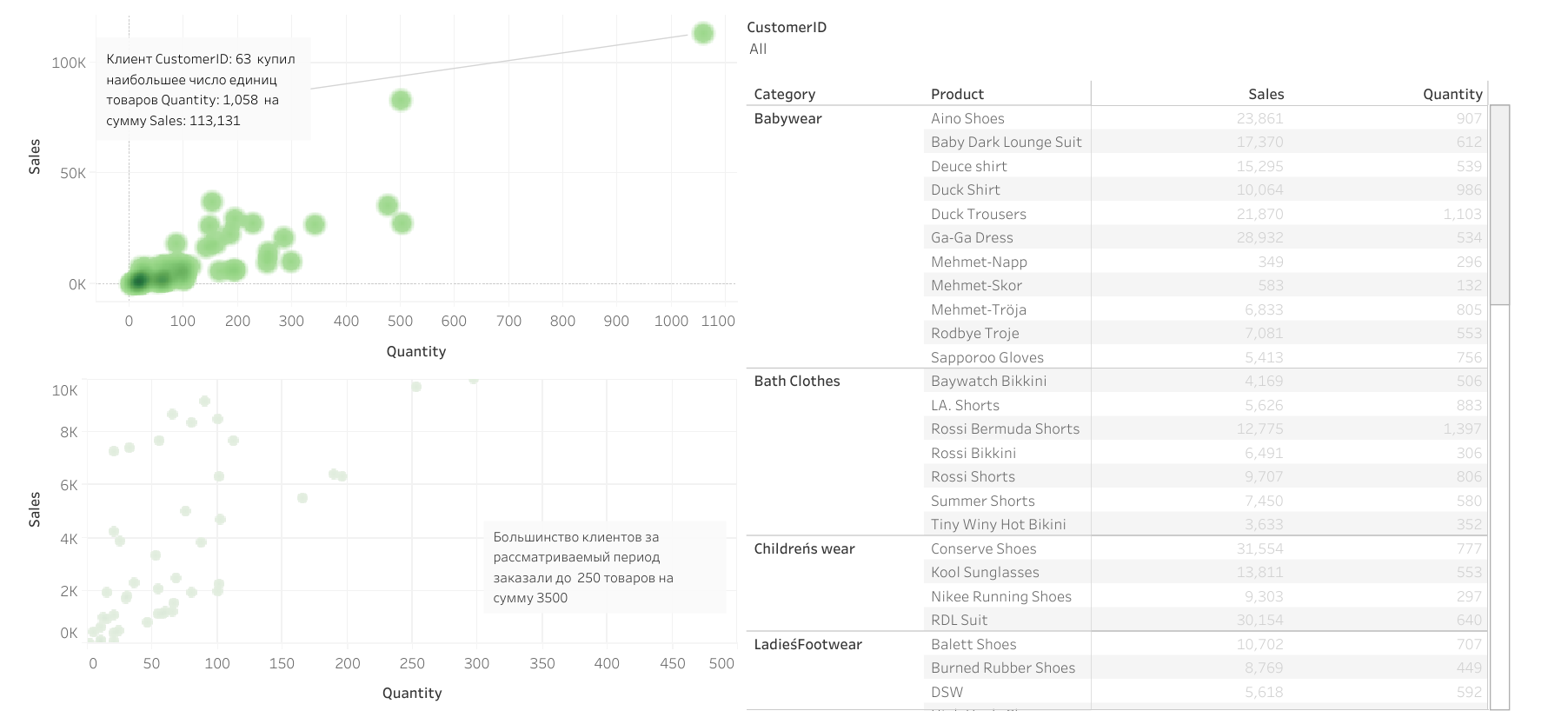
1) Minnki Palsii - самый продаваемый продукт, приносящий самую большую прибыль. Наибольшее количество проданных единиц товара составляет 1543 шт. (Game Over T-Shirt).

2) Rob Carsson - сотрудник, который осуществляет самый большой объем продаж.



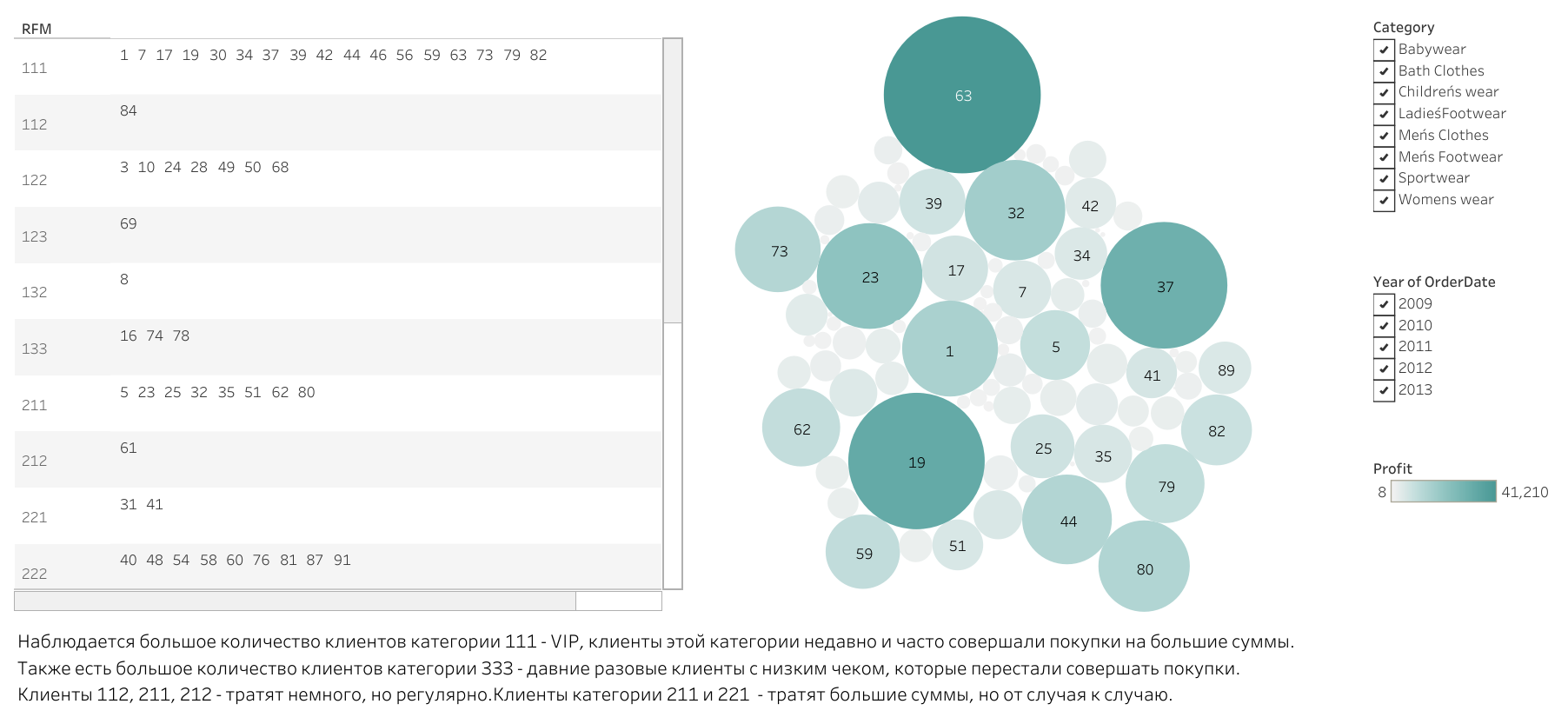
Выводы:

Предсказание продаж, выручки и количества на 2 месяца вперед не является точным, так как объем данных маленький

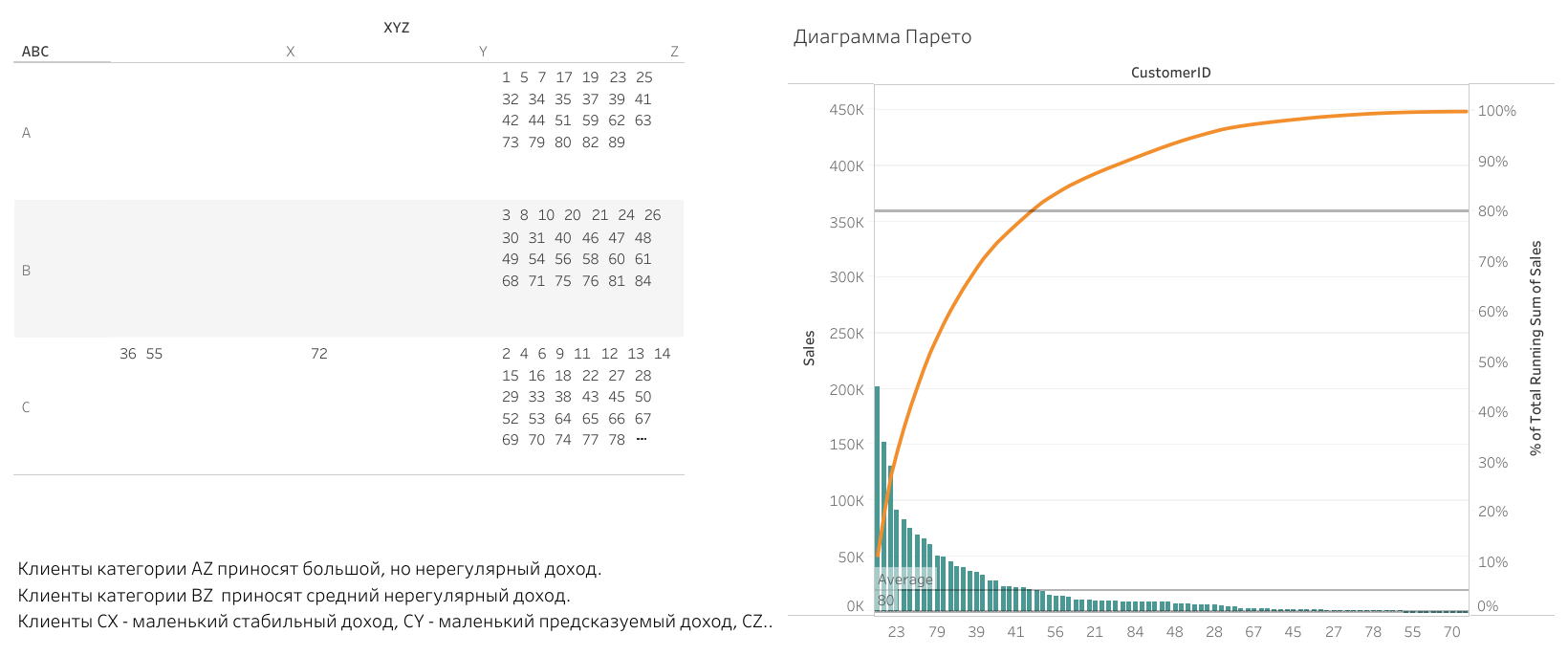


Объем продаж в зависимости от количества продаж для разных категорий и клиентов

Вывод:

Клиент CustomerID: 63 осуществил наибольшее число заказов Quantity:1058 на сумму Sales:113131$.Большинство клиентов за рассматриваемый период заказали до 250 товаров на сумму 3500.

RFM анализ



ABC-XYZ анализ

**АВС-XYZ анализ**

Первоначально АВС-XYZ анализ был составлен по товарам, но этот анализ вообще не одержал ни одного товара, который бы относился к категории X. Для большей наглядности было решено провести анализ по клиентам.

В результате ABC анализа были определены клиенты следующих категорий:

А - Клиенты, приносящие наибольшие продажи, их накопленная доля — до 80%,

В - Клиенты, для которых накопленная доля составляет от 80% до 95%,

C - Клиенты , для которых накопленная доля более 95%.

Клиенты категории А также относятся к категориям Z, эти клиенты приносят большой, но нерегулярный доход.

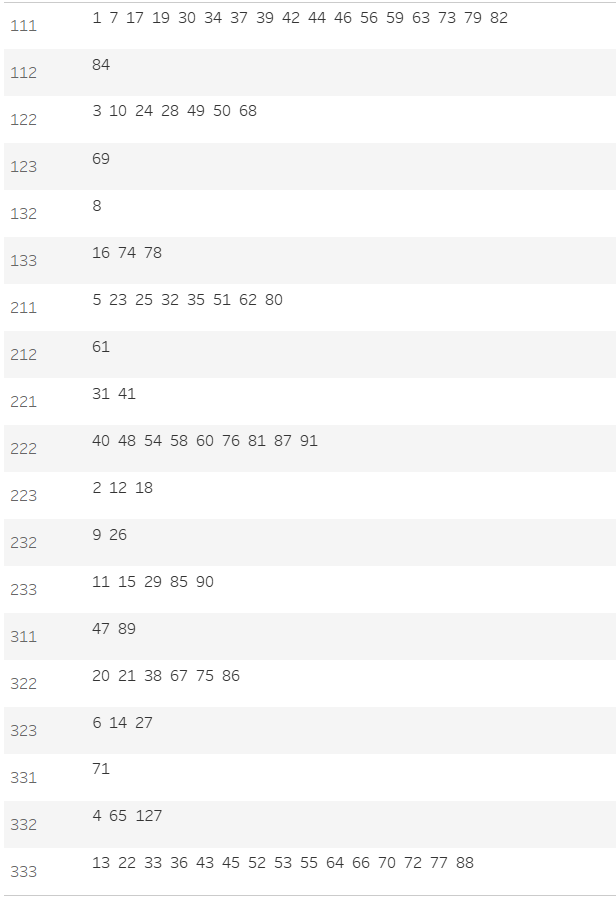
Клиенты категории В относятся к категории Y и Z. Клиенты BZ приносят средний нерегулярный доход.

Клиенты С приносят маленький доход. CX - маленький стабильный доход, CY - маленький предсказуемый доход, CZ - маленький нерегулярный доход.

Построим диаграмму Парето.

**RFM анализ**

В результате RFM анализа были определены клиенты следующих категорий:



111 - клиенты этой категории недавно и часто совершали покупки на большие суммы, так называемый VIP клиенты,

112 - недавние частные клиенты со средним чеком,

122 - недавние редкие клиенты со средним чеком,

123 - недавние редкие клиенты с низким чеком,

132 - недавние разовые клиенты со средним чеком,

133 - недавние разовые клиенты с низким чеком,

211 - спящие частые клиенты с высоким чеком,

212 - спящие частые клиенты со средним чеком,

221 - спящие редкие клиенты с высоким чеком,

222 - спящие редкие клиенты со средним чеком,

223 - спящие редкие клиенты с низким чеком,

232 - спящие разовые клиенты со средним чеком,

233 - спящие разовые клиенты с низким чеком,

311 - давние частые клиенты с высоким чеком,

322 - давние редкие клиенты со средним чеком,

323 - давние редкие клиенты с низким чеком,

331 - давние разовые клиенты с высоким чеком,

332 - давние разовые клиенты со средним чеком,

333 - давние разовые клиенты с низким чеком.

Наблюдается большое количество клиентов категории 111 - VIP, клиенты этой категории недавно и часто совершали покупки на большие суммы.

Также есть большое количество клиентов категории 333 - давние разовые клиенты с низким чеком, которые перестали совершать покупки.

Клиенты 112, 211, 212 - тратят немного, но регулярно. Это выгодные клиенты.

Клиенты категории 211 и 221 - тратят большие суммы, но от случая к случаю, спящие выгодные клиенты.

Клиенты 323, 331, 332, 333 - потерянные клиенты.

|  |
| --- |

# **Построение модели в Google BigQuery**

Для построения модели в Google BigQuery данные для моделирования и для предсказания были загружены из Knime.

Создаем модель градиентного бустинга GBM - BOOSTED\_TREE\_REGRESSOR с помощью библиотеки XGBoost

CREATE OR REPLACE MODEL`my-project-1-jule.mySchema.gbm`

OPTIONS(MODEL\_TYPE='BOOSTED\_TREE\_REGRESSOR',

BOOSTER\_TYPE = 'DART',

NUM\_PARALLEL\_TREE = 1,

MAX\_ITERATIONS = 300,

TREE\_METHOD = 'AUTO',

EARLY\_STOP = FALSE,

SUBSAMPLE = 0.85,

ENABLE\_GLOBAL\_EXPLAIN = TRUE,

INPUT\_LABEL\_COLS =['Sales']

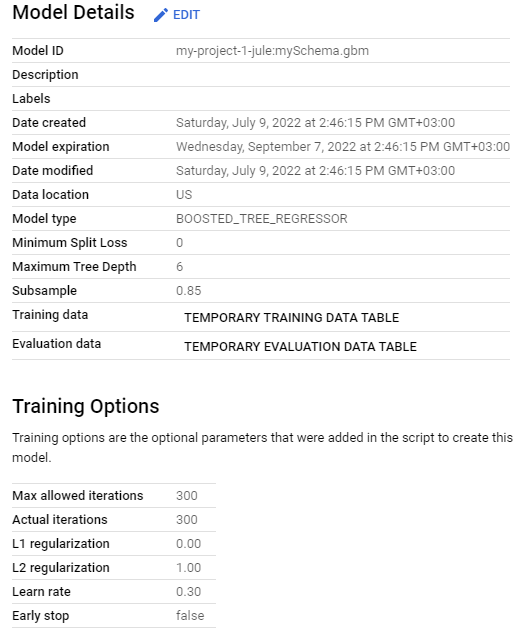
) AS

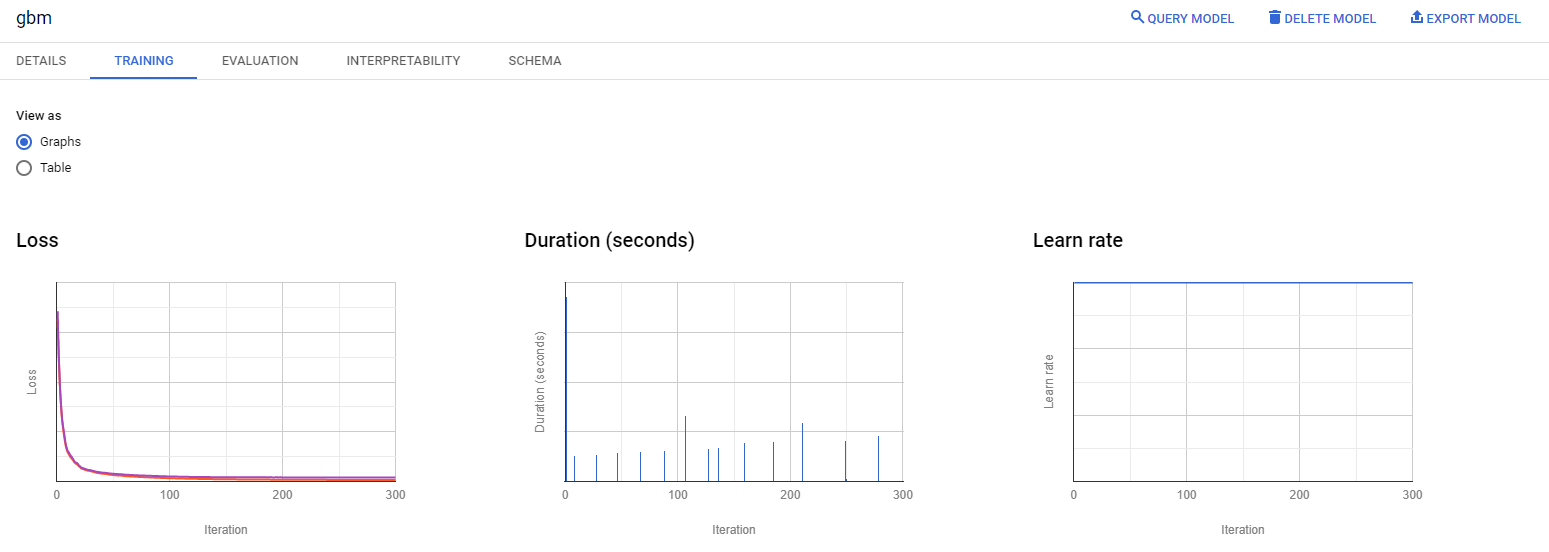
SELECT

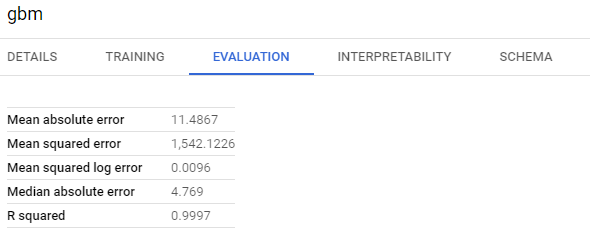
Sales AS Sales, ProductID, Quantity, SupplierID, CategoryID, EmployeeID, CustomerID, RFM

FROM

`my-project-1-jule.mySchema.DF\_train`







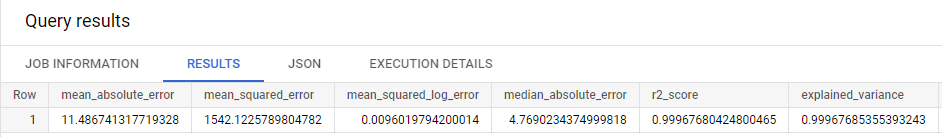
Оценка модели

Select

\*

from

ML.EVALUATE(MODEL `my-project-1-jule.mySchema.gbm`)



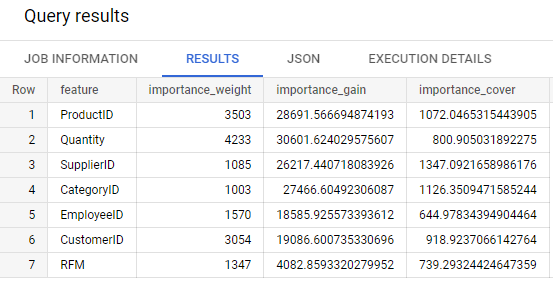
Определение наиболее важных параметров для модели

SELECT

\*

FROM

ML.FEATURE\_IMPORTANCE(MODEL `my-project-1-jule.mySchema.gbm`)



Модель сильно зависит от Quantity, ProductId и CustomerId.

Проверка модели на тестовых данных

SELECT

\*

FROM

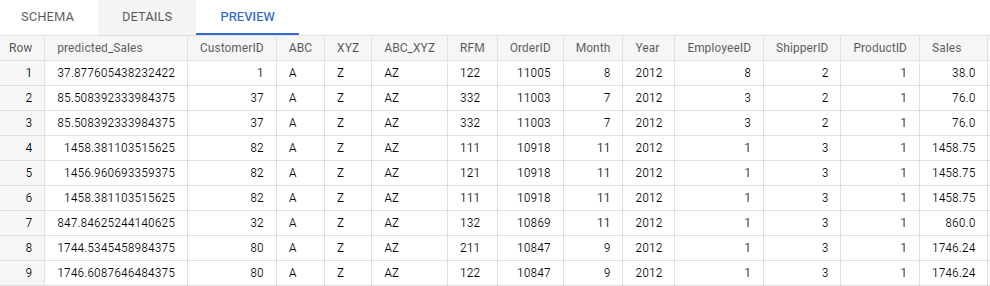
ML.PREDICT(MODEL `my-project-1-jule.mySchema.gbm`,

(

SELECT \*

FROM

`my-project-1-jule.mySchema.DF\_test`))

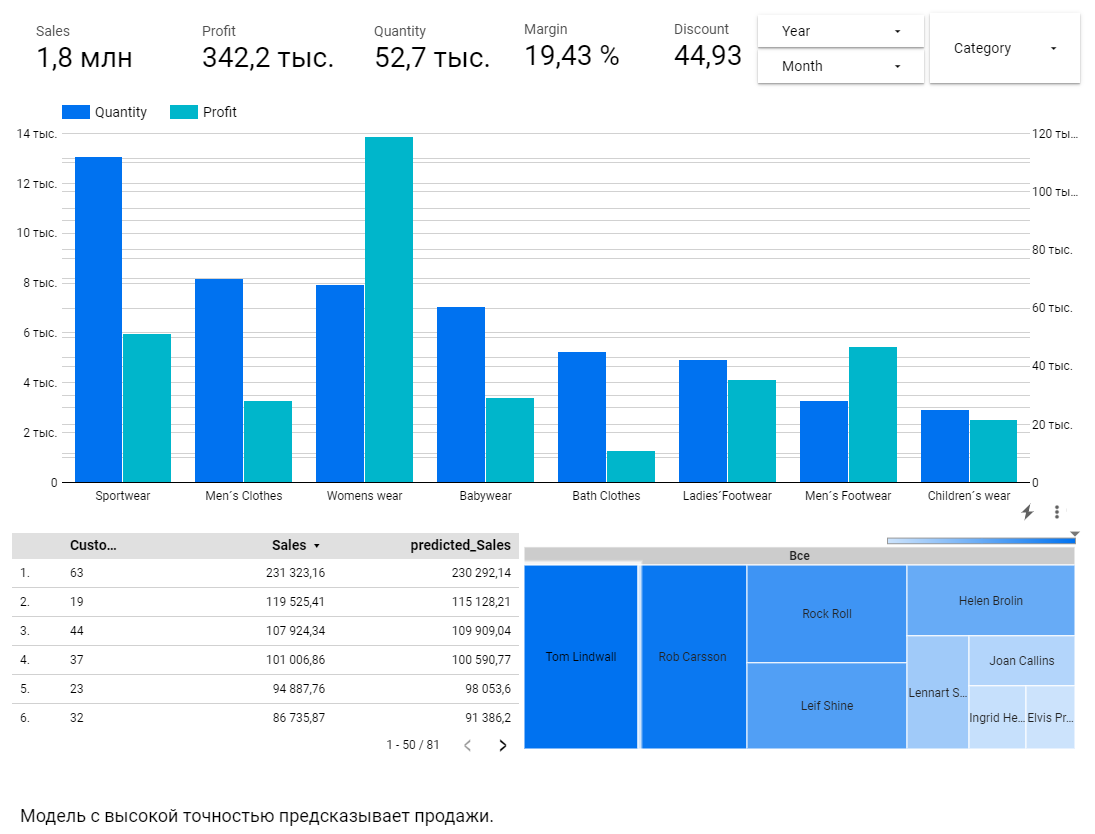


Ссылка на полученные данные:

<https://docs.google.com/spreadsheets/d/1__YxEuKXuTtiNzvAi7wJxLvian78iA1_hpb4SBeD9Lo/edit?usp=sharing>

Ссылка на DataStudio: <https://datastudio.google.com/reporting/d379f933-03ea-4d5b-b556-39620290a706>

Отчет DataStudio был интегрирован в блокнот Colab.



# **Список ссылок**

1. Ссылка на папку с файлами:

<https://drive.google.com/drive/folders/1xCUW04K9qFULosEfBRgahnRVjVjzjF0J?usp=sharing>

1. Исходные данные:

<https://drive.google.com/drive/folders/1R1GiUITeYnz4IklHL_f14Bg1XgP1IZf_?usp=sharing>

1. Knime Workflow:

<https://drive.google.com/file/d/1hMoYpUv7_-qZMK1BEFf8WGQ_BjcX-N4T/view?usp=sharing>

1. Модель H2O, построенная в Knime:

<https://drive.google.com/file/d/1s6amrtXLlupoDJGb-NsPr6BYeEpZP9S2/view?usp=sharing>

1. Обогащенные данные, выгруженные из Knime:

<https://docs.google.com/spreadsheets/d/1TA94J9DteaznuhPqyHGKSYKUuen0fsG_/edit?usp=sharing&ouid=110111397391109158013&rtpof=true&sd=true>

1. Общая схема работы в Knime:

<https://drive.google.com/file/d/11dJwtLT5_9VntGtxQ7FBOgrDNd3MGaJM/view?usp=sharing>

1. Блокнот Google Colab

https://colab.research.google.com/drive/1RX-hMqlXuSWr7DbIvHVcUyl0r2yKxTl6?usp=sharing

1. Полный отчет pandas-profiling:

<https://drive.google.com/file/d/1OxWZUyVTIjAThuP33Ua4hUVQYr0mipkb/view?usp=sharing>

1. Визуализация Tableau:

<https://public.tableau.com/app/profile/ekaterina2379/viz/StepanovaEN/QuantityOrderDate?publish=yes>

1. Ссылка на предсказанные данные BigQuery:

<https://docs.google.com/spreadsheets/d/1__YxEuKXuTtiNzvAi7wJxLvian78iA1_hpb4SBeD9Lo/edit?usp=sharing>

1. Ссылка на отчет DataStudio: <https://datastudio.google.com/reporting/d379f933-03ea-4d5b-b556-39620290a706>