TEAM ARES

Kaggle Username - DataStorm019

Data Storm 4.0

Storming Round

Highest F1 Score - 0.70833

GitHub link : https://github.com/sandani98/DataStorm-4.0.git

Chalani Ekanayake
Sachini Chandanayake
Sandani Jayawardena

Introduction

1.1. Business Problem

The goal of store profiling is to determine which stores are operating well and which are not, so that suitable steps can be made to improve their performance. (ex: distribution of resources, including as employees, marketing, and equipment, in order to boost the performance of low-rated stores and streamline the selection of an item range.) It's done based on the stores' sales and customer behaviour.

1.2. Use Case Definition

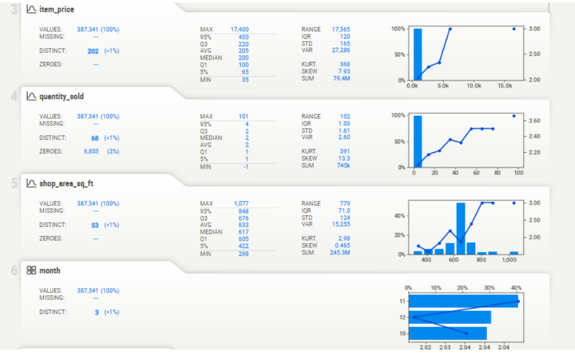
In the challenge, we are required to design and develop an advanced analytics solution that can be used by Beverages Company XYZ to perform store profiling (identify the profile of given outlets as High, Moderate or Low) and enhance their decision-making process.

1.3. Data sources

For training purposes, we are given a historical transaction data set consisting of 488,788 records which are collected from each customer purchase collected through transactions from 3 months (from 15th October 2021 to 15th December 2021). Additionally, a store info data set is given consisting of 124 stores with shop space and store profile.

Tools that were utilized to crack the case

- SQL workbench to make relations between the two datasets
- Sweetviz generates beautiful, high-density visualisations to kickstart EDA
- Excel for calculations
- Power BI to draw plots and analyse the datasets
- Matplotlib for creating static, animated, and interactive visualisations



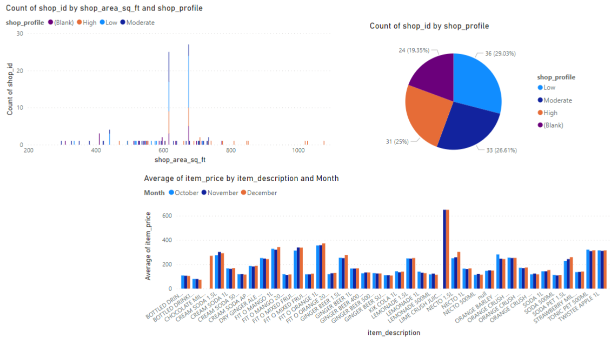
```
3 item_price
VALUES:   387,341 (100%)    MAX     17,400    RANGE  17,365
MISSING:  ---               95%        450    IQR       120
                            Q3         220    STD       165
DISTINCT: 202  (<1%)        AVG        205    VAR    27,286
                            MEDIAN     200
ZEROES:   ---               Q1         100    KURT.     368
                            5%          65    SKEW     7.93
                            MIN         35    SUM     79.4M

4 quantity_sold
VALUES:   387,341 (100%)    MAX        101    RANGE     102
MISSING:  ---               95%          4    IQR      1.00
                            Q3           2    STD      1.61
DISTINCT: 68   (<1%)        MEDIAN       2    VAR      2.60
                            AVG          2
ZEROES:   6,853  (2%)       Q1           1    KURT.     391
                            5%           1    SKEW     13.3
                            MIN         -1    SUM      745k

5 shop_area_sq_ft
VALUES:   387,341 (100%)    MAX      1,077    RANGE     779
MISSING:  ---               95%        848    IQR      71.0
                            Q3         676    STD       124
DISTINCT: 53   (<1%)        AVG        633    VAR    15,255
                            MEDIAN     617
ZEROES:   ---               Q1         605    KURT.    2.98
                            5%         422    SKEW    0.465
                            MIN        298    SUM    245.3M

6 month
VALUES:   387,341 (100%)
MISSING:  ---

DISTINCT: 3    (<1%)
```



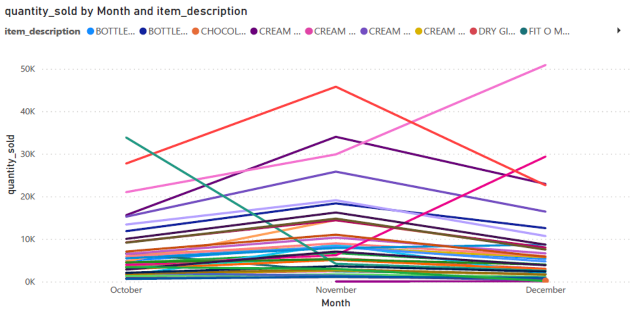
Approach

1. Exploratory Data Analysis (EDA)

We used Power Bi, matplotlib and Sweetviz to generate some plots and calculated value counts manually and using Excel to get a clearer idea of the data available.

Datasets : We have been provided with two datasets;

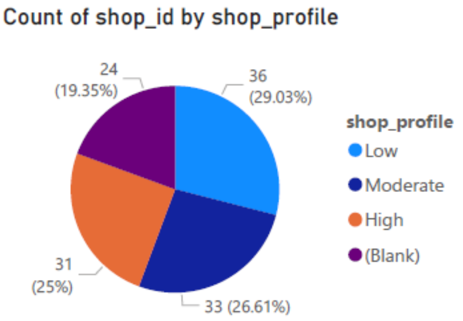
1. Historical transaction dataset
   The historical transaction data set consists of 488,788 records which are collected from each customer purchase. These purchases were expanded during three months from 15th October 2021 to 15th December 2021. We identified 37 unique items (considering different sizes of the same beverage as distinct items) sold by the XYZ Beverages company. Ex: Ginger Beer 1.5L, Tonic PET 500ML etc.

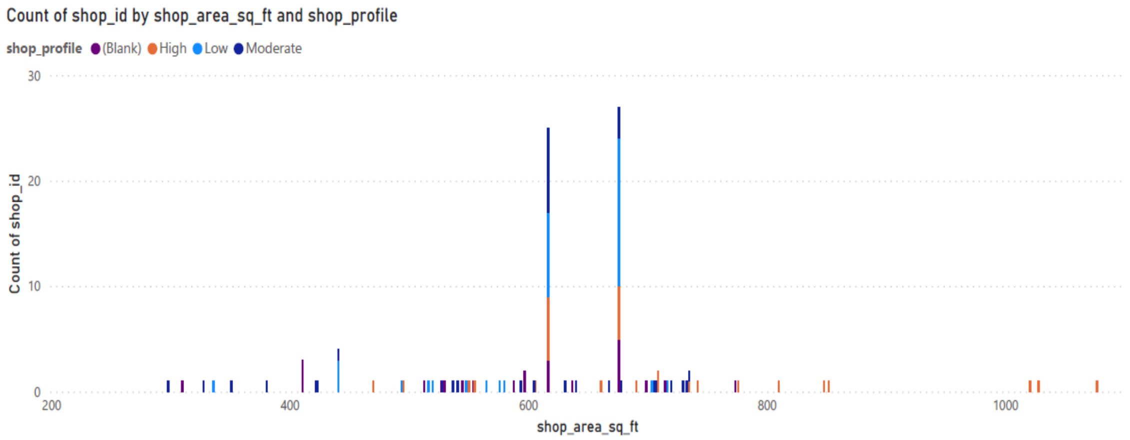


   This figure shows the total quantity sold from each item in all shops within a month.

2. Store info dataset
   The store info data set consists of 124 stores with shop space and store profile. Among these 124, 24 were in the testing dataset which is expected for us to classify as High, Moderate and Low. The remaining 100 were used for training.

   Summary of store profiles is illustrated in the figure. High, moderate and Low profile stores are approximately equal in the training set.





The figure above shows the count of shops from each profile vs the shop area. We observe that High profile stores have a comparatively larger shop area and the Low and moderate profile stores have a smaller shop area.

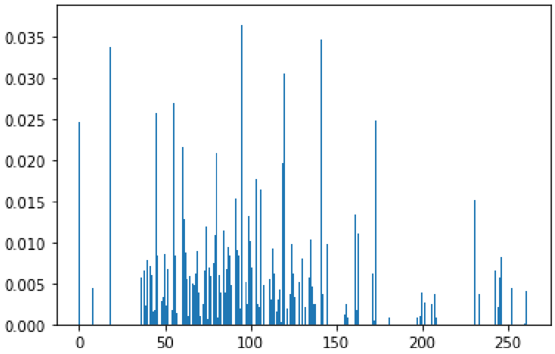
2. Dataset Preparation

2.1. Feature Engineering

We mapped information from historical transaction data into the store info dataset using python codes. There were 37 unique items sold by the XYZ Beverages company within the given duration. We came up with the following features for each store,

1. *shop_area* : already given
2. *Item_price per each item (37)* : We took one single value for each item based on the assumption that the price wouldn't vary drastically during 6 weeks' time. We took the most frequent value as the item_price.
3. *quantity_sold per each item in each month (October, November, December) (37 x 3)* : obtained by summing up the quantity_sold column in the historical transaction dataset for each item in each month
4. *earning per each item in each month (37 x 3)* : obtained by multiplying the item price by quntity_sold of each item in each month
5. *grand_quantity_sold* : total of *quantity_sold in that particular shop within the duration*
6. *grand_earning* : total of *earnings in that particular shop within that duration*

It gave altogether 262 features for each store.



This figure shows the importance of each of those features towards the target variable.

2.2 Dataset Cleaning

After making suitable features, we replaced the zeros in item_price columns (The values are zero due to that item not being sold in that particular shop within that period of time) with the maximum value of that respective column.

The only categorical variable was the shop_profile which is the target column. We applied label encoding for shop_profile columns.

3. Model Training and Hyper-Parameter Tuning

We split the shop_profile given stores (100 stores) into training(80%) and validation(20%). We carried out experiments with several models including XGBoost classifier, decision tree and random forest as these models support multiclass classification problems.

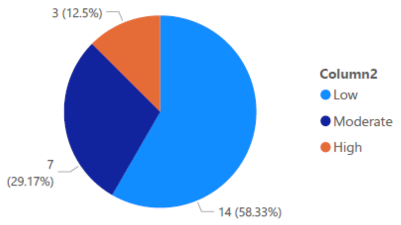
Model Name	F1 Score for the validation set
XGBClassifier	0.524
DecisionTreeClasifier	0.493
RadomForestClassifier	0.7

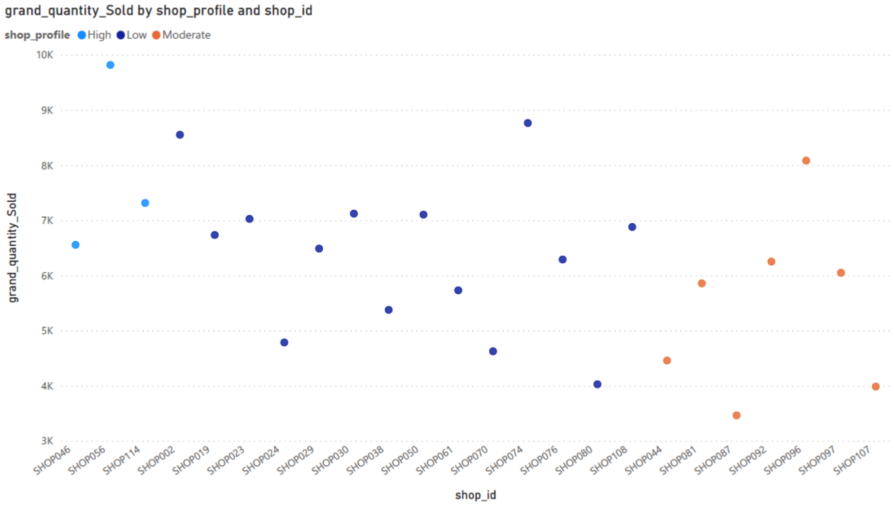
Table 01: Model Performance

We did parameter tuning for these three models using bayesian optimization and random search. The above table shows the best accuracy we obtained. Since the number of datapoints is small we couldn't obtain a better insight on the test set through the validation set. When submitted to the kaggle, results obtained with XGBClassifier gave the best result of 0.70833.

Business Insights

This pie chart illustrates the testing data stores' store profile after prediction. As seen most of the stores gave a LOW profile.





As the figure above shows the High profile stores always seem to have a higher amount of quantities sold and consequently have a higher earning.

Additional Attributes for Store Profiling

- Number of customers who visited the store (irrespective of whether they bought some item or not) during a day
- Festive Seasons
- Count of staff members in a store
- Stores’ location (Ex: distance to the nearby city)

Interventions that can be taken to Enhance the Decision-making Process

- Creating a dashboard that visualises data from the store profiling mechanism and gives a real-time overview of sales, inventory levels, consumer behaviour, and other vital indicators. Managers can use this to swiftly spot trends and make data-driven choices.
- Using predictive analytics models to forecast sales, demand, and inventory levels. This can help the company to optimise their operations, avoid stockouts, and reduce waste.
- Using the data collected from the store profiling mechanism to conduct A/B testing on various marketing campaigns, product placement, and pricing strategies.
- Using data from the store profiling mechanism to generate possible future scenarios, such as changes in consumer behaviour or economic conditions. As a result, the company can plan for many scenarios and make informed decisions.
- Using segmentation data to segment customers based on their behaviour, tastes, and purchase history can assist the company in tailoring their marketing tactics to specific client categories and creating more successful customised offers and promotions.