Data Storm 4.0

MageZaKi

F1 SCORE - 0.58333

Kaggle Username - DataStorm026
https://github.com/Buddhi19/Data-Storm-MageZaKI

Mr. D.M.U.P. Sumanasekara
Faculty of Engineering,
University of Peradeniya
e19391@eng.pdn.ac.lk

Mr. W.M.B.S.K. Wijenayake
Faculty of Engineering,
University of Peradeniya
e19445@eng.pdn.ac.lk

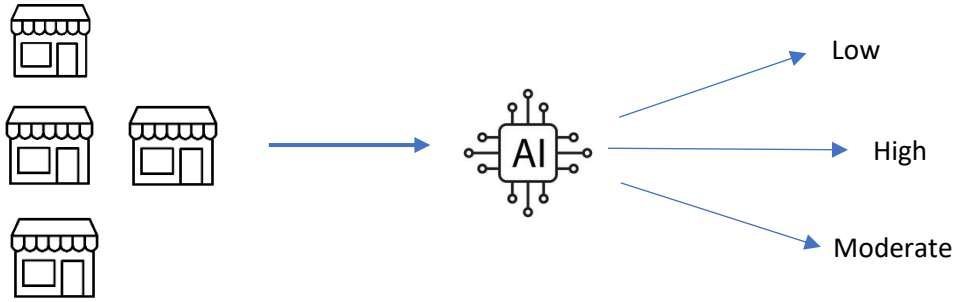
Mr. R.M.A.M.B. Ratnayake
Faculty of Engineering,
University of Peradeniya
e19328@eng.pdn.ac.lk

Introduction

Predicting the store profiling to enhance decision- making process, is more important for a business. In this project we look at a dataset with information about stores, their areas, items sold by each store and customer data. We use different models to predict the profile of each shop and using the results of the model with highest test F1 score, we will predict the profile as Low, Moderate and High.

We also visualize how each model performs about the predictions.

Approach



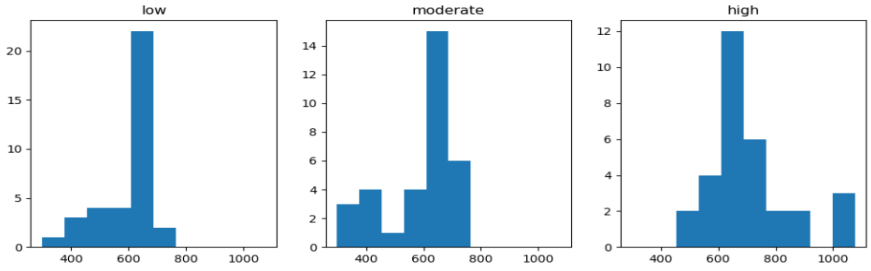
In order to group the given shops, we decided to use clustering algorithms with the Python library scikit-learn. Before applying clustering, we identified the relevant features/attributes for grouping the data.

Feature Engineering

Several features were inferred from the data obtained from the dataset about the transactions as well as from the data from the shop info dataset. A brief overview of the features obtained as such is given below. After evaluating each feature and their combinations with the different machine learning models a final feature set was chosen for the final model.

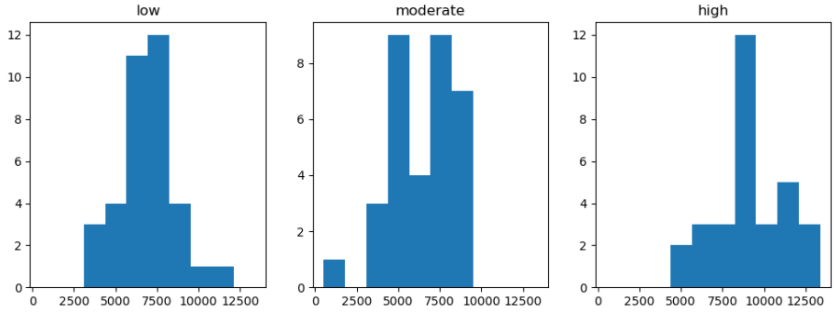
Shop Area

This feature is simply the area of the shop it was feasible to consider that there would be some correlation between the area and the shop profile. Considering the histograms for the three different profiles it can be seen that while there is some correlation it is low.



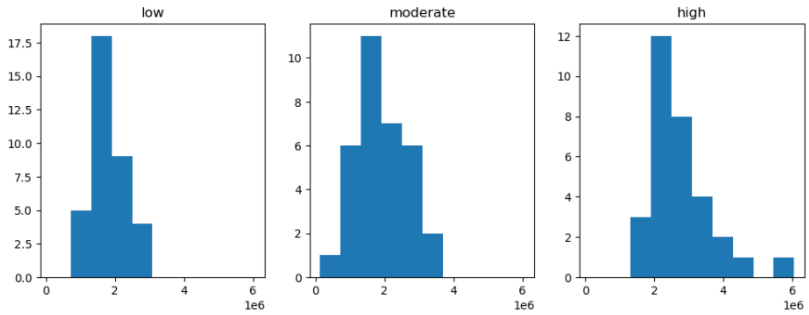
Quantity of Product Sold

This feature is simply the total number of products that were sold in all the transactions



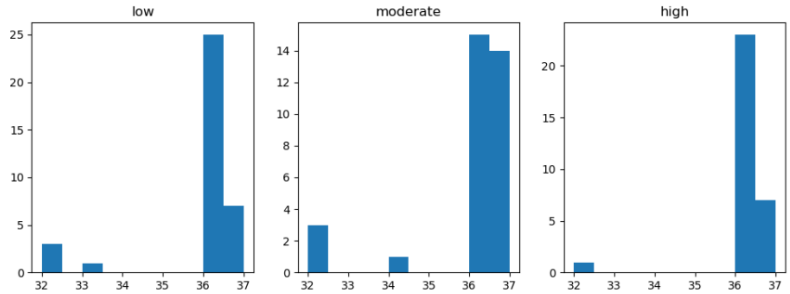
Income

This feature is the total income that was obtained from all the transactions. While distinguishing High stores is somewhat possible, distinguishing between the other two types is difficult.



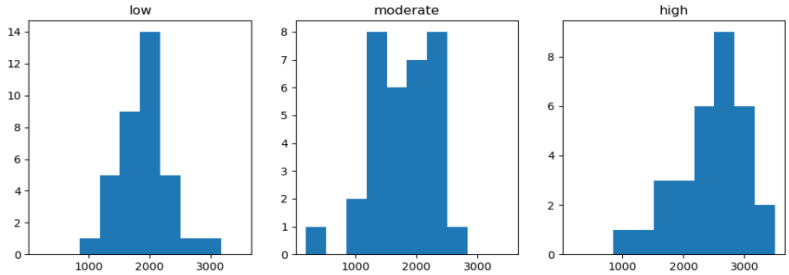
Product Diversity

Considers the number of unique types of products that is available in the store, it can be seen that this feature is not that useful as there is very little variation in this feature and all store carry almost all products, however a lot of moderate stores seem to carry all the products compared to the other two.



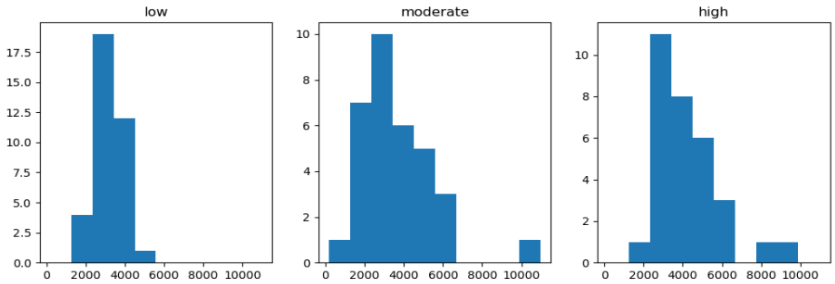
Customer Diversity

Similar to the product diversity this is the different number of customers that have visited the store in the time period



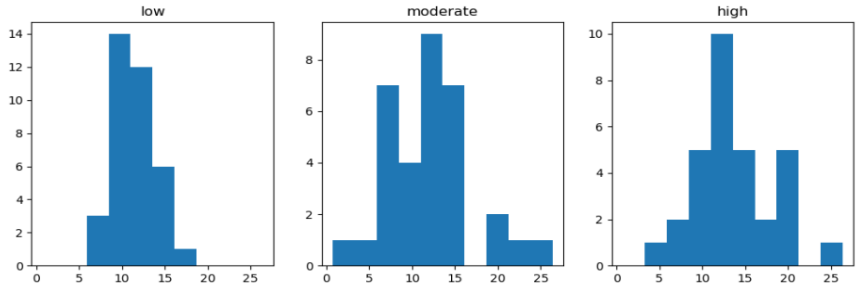
Income per Area

As it may be unfair to compare larger stores with smaller stores income wise, income per area was calculated. There is a slight potential to distinguish between low and other types of stores.



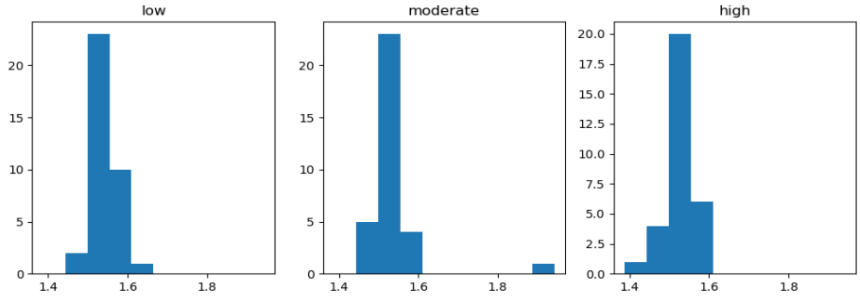
Quantity per Area

Similarly, quantity per area was also calculated and compared.



Income Growth

Income growth is a value calculated as follows.



$$Income\ growth = \frac{Income\ of\ a\ certain\ day \times Index\ of\ the\ given\ day}{Total\ income}$$

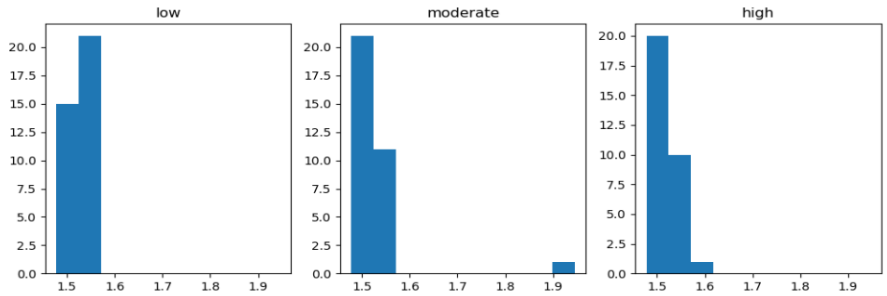
This was meant to represent any growth of the income as days that are more recent would have a higher weight however very little variation was observed.

Quantity Growth

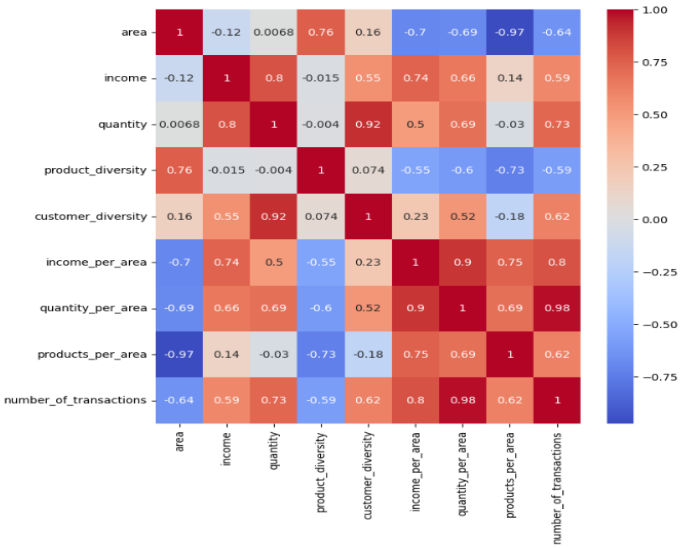
Similar to income growth, income growth was also calculated.

$$Income\ growth = \frac{Quantity\ of\ a\ certain\ day \times Index\ of\ the\ given\ day}{Total\ income}$$

This was meant to represent any growth of the income as days that are more recent would have a higher weight however very little variation was observed.

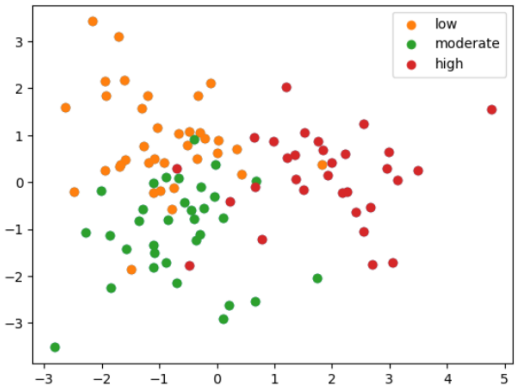


Correlation of features with Profile



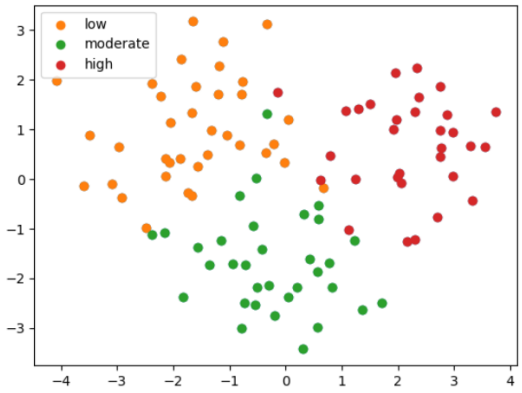
Juice Vectors

Juice vectors are vectors $\boldsymbol{v} \in Z_0^{+n}$ where n is the number of different juice available in the shop. Where the ith component of the vector represents the number of juices sold of that type. Then dimensionality of this space was reduced to two dimensions using linear discriminant analysis, and this reduced feature was used as a feature. The plot shown below shows good separation but this may be due to overfitting.



Time Vectors

Time vectors are vectors $\boldsymbol{v} \in Z_0^{+m}$ where m is the number of days in the given set of transactions. Where the ith component of the vector represents the number of juices sold of the ith dat. Then dimensionality of this space was reduced to two dimensions using linear discriminant analysis, and this reduced feature was used as a feature. The plot shown below shows good separation but this may be due to overfitting. However, the results when submitted using this feature was low, so this feature was abandoned.



Machine Learning Models and Training Process.

Several machine learning methods were used as candidate models. A complete list of all the models that were tried is as follows.

1. KNN
2. SVM
3. Gaussian Process
4. Decision Tree
5. Random Forest
6. Neural Net
7. AdaBoost
8. Naive Bayes

However, in this report, only the ones that were considered in submissions will be addressed.

Training Process.

In order to compare the models fairly and gain a true perspective as to the accuracy of the models without being misled by overfitting of the models, the training data was divided to two training and testing data sets for model and feature evaluation (70% and 30%).

However, for the final prediction the models were trained on the entire dataset.

Evaluation

To evaluate the performance two metrics were used.

1. F1 score
2. Confusion Matrix

SVM- Support Vector Machine was used with the RBF kernel as one of the models it gave good results and the f1 score and confusion matrix of the final submission were promising.

KNN- Nearest Neighbors in a supervised mode was also used as a model. While it gave descent results it was not as reliable as some of the other method.

Random Forest - Random Forest while managed to always model the training data perfectly was poor on the training data set thus indicating it had a tendency to overfit to the data.

Simple Neural Network- A simple Neural Network was used that had 2 hidden layers of 100 dimensions each. This had comparable results to the SVM.

Final Model

For the final model, we propose a model that has the following feature set.

1. Shop Area
2. Quantity per Area
3. Income per Area
4. Customer Diversity
5. 2 dimensions of the reduce juice vectors

This make the input a 6-dimensional vector (The first 4 are normalized using the z -score)

For the model SVM and the Neural Network gave comparable results but the SVM was chosen in the end for its reliability and simplicity.

Business Insights

Our analysis of XYZ company's data revealed significant key features in business insights that can help the company to improve its decision-making process optimistically. As we mentioned in previous topics, we observed that the behavior of the customers is diverse, and as given in the problem statement, it can be divided into 3 main segments namely, High, Moderate and Low. By making this division the company can tailor its marketing strategies to better meet the needs of each category of customers.

Apart from that, we identified that the patterns in sales and customer behavior can be used to make item range decisions. For instance, certain products may be more popular in certain shops, or some shops may have higher sales during specific periods. With that knowledge, we recommend to concentrate on popular products in certain areas while distributing the products among the outlets. In addition, during periods which a certain outlet has higher sales, more products and staff should be provided to the outlet.

Subsequently, we've observed that some products are constantly underperforming while some are growing with a significant potential. By identifying the products which has a significant growth and the products which are not, we can help the company to prioritize the products.

Based on our analysis, we have identified some outlets which have experienced a significant growth over time, while some have not exhibited a satisfactory progress in terms of income, sales volume, etc. In the light of this information, we can assist the company in directing more resources to those underperforming outlets to provide them with the necessary support to achieve growth and improve their outcome. By taking these measures beforehand helps the company to maximize its profitability and enhance sustainability.

In conclusion, our analysis of Beverage Company XYZ's data revealed some key business insights that can undoubtedly help the company to improve their decision-making process. By considering these insights, the company will certainly be able to have a significant improvement in marketing, staffing and product distribution. Overall, our analysis demonstrated the value of data-driven decision-making and the importance of using advanced analytics and machine learning techniques in business landscape.